

August 12, 2025
DRAFT

Thesis Proposal
**On the Developmental Science of
Language Models**

Emmy Liu

August 2025

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Graham Neubig (Carnegie Mellon University)
Chenyan Xiong (Carnegie Mellon University)
Aditi Raghunathan (Carnegie Mellon University)
Jacob Andreas (Massachusetts Institute of Technology)
Andrew Lampinen (Google Deepmind)

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

Copyright © 2025 Emmy Liu

August 12, 2025
DRAFT

Keywords: natural language processing, machine learning, cognitive science

Abstract

The remarkable progress of language models in recent years has primarily been driven by scaling to larger models as well as larger quantities of data. Despite this, this approach is starting to face increasing friction as models reach the limits of available high-quality web data, especially for highly specialized use cases. A fundamental issue is that of inconsistent development – not all capabilities improve equally well when scaling the number of tokens of generic web data. Furthermore, there are many seemingly contradictory results where state-of-the-art language models perform at a superhuman level at tasks that we think of as highly specialized, while failing at others we consider trivial such as arithmetic or counting the “r”s in “strawberry”. We argue that in order to resolve these tensions, it is important to understand the process of capability development within pretraining as well as posttraining at a more granular level. We would also like to argue that this set of problems is not unique – it has been studied by developmental psychologists in the case of human development. Although we cannot draw a direct analogy between humans and language models, in this thesis we propose to take inspiration from this field to better understand and improve language models by investigating how specific capabilities emerge, interact, and can be targeted for improvement throughout training.

We organize our investigation into three complementary approaches. First, we **evaluate** core cognitive capabilities that are foundational to human generalization – analogical reasoning, compositionality, and forms of inference – revealing that models partially succeed at these tasks but exhibit surprising deficits from over-reliance on surface frequency and poor representation of long-tail phenomena. Second, we seek to **understand** capability emergence by investigating factors beyond scale that predict performance, examining the critical “midtraining” phase of model training, and proposing to decompose complex capabilities into elemental skills that can be easily measured and study their training trajectories. Finally, we seek to **intervene** through targeted interventions to data or training methods in order to improve long-tail, compositional, and data-scarce skills such as translating idioms, repo-level code generation, and creative reasoning in scientific idea generation. In summary, we believe the ability to systematically target specific abilities and reason about their development in the course of training offers a complement to existing scaling approaches that could contribute to more principled training of language models.

August 12, 2025
DRAFT

Acknowledgments

Note: This section will be expanded in the final dissertation document.

Thank you to the friends, family, mentors and collaborators that have supported me and guided me on this path so far.

August 12, 2025
DRAFT

CONTENTS

1	Introduction	1
1.1	Thesis Overview	3
1.2	Proposed Timeline	6
2	Background	7
2.1	The Lifecycle of a Language Model	7
2.2	Lessons from Human Cognitive Development	8
2.2.1	Capabilities that enable transfer learning	8
2.2.2	Learning and Development	9
3	Evaluating Complex Semantic Reasoning	11
3.1	Background	11
3.2	Testing the Ability to Interpret Figurative Language	11
3.2.1	Introduction	11
3.2.2	Dataset Creation and Validation	12
3.2.3	Figurative Language Typologies	13
3.2.4	Baseline Models and Evaluation	15
3.2.5	Results	17
3.2.6	Performance and Error Analysis	20
3.2.7	Related work	24
3.3	The (Non*)compositionality of Representations	26
3.3.1	Introduction	26
3.3.2	Methods and Experimental Details	27
3.3.3	Data and Compositionality Judgments	28
3.3.4	Approximating a Composition Function	29
3.3.5	Examining Compositionality across Phrase Types	31
3.3.6	Comparing Compositionality Judgments of Humans and Models	36
3.4	Abductive Reasoning	39
3.4.1	Introduction	39

3.4.2	Three Types of Reasoning in Language Models	40
3.4.3	Instruction Following	41
3.4.4	Few-Shot Prompting	41
3.4.5	Instruction Inference	41
3.4.6	Methods	42
3.4.7	Domains and Evaluation	42
3.4.8	Linear Functions	43
3.4.9	Simple Artificial Languages	44
3.4.10	Kalamang Translation	44
3.4.11	Results	45
3.4.12	When Does Instruction Inference Improve Over In-Context Learning?	46
3.4.13	How Does the Ability to Induce Instructions Relate to In-Context Learning?	46
3.4.14	Conclusion	49
3.5	Conclusion	51
4	Studying the Development of Complex Skills in Training	53
4.1	Background	53
4.2	Linking Downstream Capabilities to Architecture and Data	54
4.2.1	Introduction	54
4.2.2	Scaling Laws	54
4.2.3	Building a Database of Publicly-Available Language Models	55
4.2.4	Predictive Modeling	60
4.2.5	Results	61
4.2.6	Related Work	64
4.2.7	Conclusion and Future Work	65
4.3	Midtraining Bridges Pretraining and Posttraining	66
4.3.1	Introduction	66
4.3.2	Defining Midtraining	66
4.3.3	Experimental Setting	68
4.3.4	Which downstream tasks benefit most from midtraining?	70
4.3.5	What data is most effective for midtraining?	71
4.3.6	When and how much midtraining data should be introduced?	73
4.3.7	How does midtraining change model representations?	75
4.3.8	Related Work	76
4.3.9	Conclusion	78
4.4	Elemental Skills and their Emergence (proposed)	79
4.4.1	Motivation	79
4.4.2	Research Questions	79
4.4.3	Proposed Experiments	80
4.4.4	Expected Outcomes	80
4.5	Conclusion	81

5	Scaffolding Development for Real-World Tasks	83
5.1	Background	83
5.2	Learning to Translate Idiomatically	84
5.2.1	Introduction	84
5.2.2	Non-Compositional Translation	84
5.2.3	Idioms and Data Collection	85
5.2.4	Evaluating Non-Compositional Translation	87
5.2.5	Methods to Improve Non-Compositional Translation	89
5.2.6	Experimental Settings	92
5.2.7	Results	93
5.2.8	Related Work	94
5.2.9	Conclusion	96
5.3	Skill-driven Model Training: A Study in Repo-level Coding (proposed)	97
5.3.1	Motivation	97
5.3.2	Research Questions	97
5.3.3	Expected Outcomes	98
5.4	Active Exploration in Language Model Training (proposed)	99
5.4.1	Motivation	99
5.4.2	Research Questions	99
5.4.3	Proposed Approach	100
5.5	Conclusion	101
6	Conclusion	103
A	Appendix for Chapter 3.2	105
A.1	Crowdsourcing Details	105
A.2	Invalid Examples	106
A.3	Backward accuracies	107
A.4	Paired accuracies	107
A.5	Accuracy breakdown by Part-of-Speech	108
A.5.1	Subject	108
A.5.2	Relation	108
A.5.3	Object	109
A.6	Accuracy breakdown by hypernyms	109
A.6.1	Subject	109
A.6.2	Object	110
A.7	Generation examples	111
B	Appendix for Chapter 3.3	113
B.1	Treebank dataset tree types	113
B.2	Treebank dataset phrase lengths	114
B.3	Probe learning curves	114
B.4	Length Correlation	114
B.5	Idiom and matched pair examples	114

B.6	Error ratio of probes	116
B.7	Annotation setup and instructions	119
B.8	Compositionality scores without anisotropy correction	123
B.9	AUC of approximative probes	126
B.10	Mean deviation of phrase types by tree type	127
B.11	Further named entity results	127
B.12	Frequency and length correlations	131
C	Appendix for Chapter 3.4	133
C.1	Model Inference Settings	133
C.2	Prompts for Linear Functions Domain	135
C.3	Prediction Fit of Other Models On Linear Functions	139
C.4	Predicted Coefficients compared to Real Coefficients in Linear Functions	139
C.5	Grammar and Details of Colours Domain	139
C.6	Prompts for Colours Domain	140
C.7	chrF For Colours Domain	144
C.8	Vocabulary Induction Accuracy for Colours Domain	144
C.9	MTOB Experimental Settings	146
C.10	MTOB Prompts	146
C.11	Kalamang Grammar Sketch	152
C.12	Predicted Kalamang Grammar Sketches and Accuracy	152
C.13	Vocabulary Induction Accuracy for MTOB	154
C.14	Example Kalamang Translations	156
D	Appendix for Chapter 4.2	157
D.1	List of all models	157
D.2	List of all architectural and data features	160
D.2.1	Architectural Features	160
D.2.2	Data Features	161
D.2.3	Freegen-derived Features	161
D.3	List of all evaluations and settings	163
D.4	Task Deviations from Kaplan-style Scaling Laws	163
D.5	Free-generation Domain Classification	164
D.6	Domain Classifier Validation	164
D.7	Free-generation Validation	164
D.8	Free-generation Examples	165
D.9	XGBoost Settings	167
D.10	Selected Features by Task	167
D.11	LightGBM Results	167
D.12	SHAP Plots for remaining benchmarks	167
D.13	Details on confirmatory pretraining runs	167
D.13.1	Training	167
D.13.2	Evaluation	168
D.13.3	Conversion to Loss-Based Metrics	168

D.13.4	Full Results	169
E	Appendix for Chapter 4.3	179
E.1	Pretraining Settings	179
E.2	Posttraining Settings	179
E.3	Dataset Similarity Matrix	180
E.4	SFT in-domain loss and C4 Losses after Finetuning for 70m and 160m models	181
E.5	Midtraining Leg Length vs. Benefit	183
E.6	Representative training loss curves for midtrained vs. base models	183
E.7	Additional CKA results on APPS	183
E.8	CKA results on C4	183
F	Appendix for Chapter 5.2	191
F.1	Synthetic Dataset Training Details	191
F.2	Idiom Sources	191
F.3	OpenSubtitles Training Details	192
F.4	Native Annotator Recruiting	192
F.5	Standards for Human Evaluation	193
F.6	Statistical Significance Testing	194
F.7	Statistical Significance Results	194
F.8	Effect of Frequency on Commercial Translations	194
F.9	Effect of Frequency on DeltaLM	200
F.10	Automatic Metrics in Detail	200
	Bibliography	203

August 12, 2025
DRAFT

LIST OF FIGURES

3.1	Bar chart of the 25 most frequent subjects, relations, and objects in the FigQA dataset.	14
3.2	Bar chart comparing performance of few-shot prompts on Fig-QA.	19
3.3	Scatterplot relating predicted probability of an answer by itself to a model’s final prediction; models over-rely on predicted probability of the answer alone.	22
3.4	An illustration of the local composition prediction problem with [CLS] representations.	26
3.5	Bar chart of mean compositionality scores of each approximative probe in approximating model representations.	31
3.6	Density plot for compositionality scores of named entities and non-named-entities with RoBERTa _{AVG} representations. Higher means more compositional.	33
3.7	Density plots for compositionality scores of different named entity types with RoBERTa _{AVG} representations. Higher means more compositional. . .	34
3.8	Tree reconstruction error (cosine distance) for each parent phrase type present in the Treebank, ordered from highest mean error to lowest. Based on the affine approximation for each model and representation type. Expanded version with all tree types is presented in § B.10.	35
3.9	Diagram of abductive reasoning for an LM. Red arrows show data flow in inductive reasoning (few-shot prompting), while blue arrows show data flow in deductive reasoning (instruction following). Black arrows indicate data flow unique to abductive reasoning (instruction induction). Instruction inference generally improves on few-shot prompting and zero-shot chain of thought. However, success at inductive reasoning and success at instruction inference are not related.	40

3.10	Real coefficients of linear functions and relationship to hypothesized coefficients for GPT-3.5-turbo and GPT-4-turbo. Remaining models can be found in § C.4. The x-axis has been truncated for visualization purposes (as there are some large outlier hypotheses). GPT-4-turbo is able to induce a reasonable function in-context, but other models struggle.	47
3.11	Accuracy of models in synthetic domains with and without hypothesis generation. Error bars indicate standard error. The top row shows results for the functions domain, while the bottom row shows results for the colours domain. Results are aggregated across 6 runs, and zero values are marked with '0'.	48
3.12	chrF scores for Kalamang under different methods, in English to Kalamang direction (top row) and Kalamang to English direction (bottom row)	49
4.1	Data taxonomy for pretraining sources.	56
4.2	Scaling contour plots (expected) vs. actual model performance in diverse tasks.	59
4.3	SHAP visualization depicting impact of design decisions on predicted performance across tasks.	62
4.4	Line plot showing the SHAP impact of code percentage on natural language vs. code.	63
4.5	Scatterplot showing proximity advantage vs. midtraining performance . .	71
4.6	Validation loss and C4 loss for the starcoder-midtrained model and base pretrained model <i>after</i> supervised fine-tuning on the Pycode dataset, with each point on the x-axis representing the number of steps the pretrained checkpoint was trained on.	72
4.7	Example similarity matrix between pre/midtrain and posttraining datasets. For the complete matrix, see Appendix E.3.	73
4.8	Timing ablation: validation losses (SFT and C4) by model size and starting step for the Pycode downstream dataset (in-domain). Top: SFT losses after Pycode finetuning; bottom: C4 losses showing forgetting after Pycode finetuning. In this setup, Starcoder is mixed at a mixed 20% ratio. . .	75
4.9	Mixture ablation: validation losses (SFT and C4) by model size and mix for the Pycode downstream dataset (in-domain). Top: SFT losses after Pycode finetuning; bottom: C4 losses showing forgetting after Pycode finetuning. The x-axis depicts the mixture weight of Starcoder mixed in at 40k steps.	76
4.10	CKA analysis of model activations in the 70m model, probed with the APPS code dataset.	77
5.1	Accuracy of a transformer in translating a non-compositional phrase after training on datasets of different sizes, with different numbers of non-compositional patterns (only non-compositional translation accuracy is depicted). Results are averaged across 5 seeds, and standard deviation is shown.	88

5.2	Automatic metrics – Quality of DeepL French translations on idiomatic test set bucketed by idiom frequency. The bottom 20% of least common idioms are excluded, as they may occur fewer than 3 times and not be in our test set.	90
5.3	Results of automatic metrics. In most cases, combining loss weighting with KNN-MT improves automatic metrics the most on all three test sets, including the out-of-distribution (Random) test set.	92
5.4	Automatic metrics for fr-idiom sentences, plotted by frequency, for base and upweight+knn.	95
B.1	Length distribution of phrases mined from the treebank, in number of words. The modal length was 3 words, followed closely by 2 words. Few phrases contained more than 50 words.	117
B.2	Learning curves of approximative probes trained on differing percentages of train data.	118
B.3	Examples of compositionality judgments shown to annotators	120
B.4	First page of annotation interface for a practice phrase	121
B.5	Second page of annotation interface for a practice phrase	122
B.6	Mean deviation from predicted representation across full tree types. . . .	128
B.7	Distributions of compositionality score for named entities and non-named entities across model types and representation types. The AVG representation matches the intuition that named entities are usually less semantically compositional, as they point to an entity in the real world that may not relate to their name.	129
B.8	Visualization of distribution of compositionality scores across different types of named entities.	130
D.1	SHAP values for GSM8k	175
D.2	SHAP values for Lambada	176
D.3	SHAP values for Hellaswag	176
D.4	SHAP values for MMLU 0-shot	177
D.5	SHAP values for MMLU 5-shot	177
D.6	SHAP values for ANLI	177
D.7	SHAP values for XNLI	178
D.8	SHAP values for MathQA	178
D.9	SHAP values for LogiQA2	178
E.1	Token-based similarity matrix for pre/midtrain and SFT datasets. Note that these midtrain datasets are corrected for mix weight in this matrix. . .	181
E.2	Relationship between the two “legs” $\text{dist}(\text{C4}, \text{midtrain})$ and $\text{dist}(\text{midtrain}, \text{SFT})$ and the benefit of midtraining.	184
E.3	Representative training loss curve for a midtrained model and base model on Pycode, for Pythia-410m. The midtrained model starts with a lower training loss, and maintains a slight gap throughout training.	185

E.4	CKA layer analysis for Pythia-160M with APPS as a probe.	186
E.5	CKA layer analysis for Pythia-410M with APPS as a probe.	187
E.6	CKA layer analysis for Pythia-70M with C4 as a probe.	188
E.7	CKA layer analysis for Pythia-160M with C4 as a probe.	189
E.8	CKA layer analysis for Pythia-410M with C4 as a probe.	190
F.1	Effect of idiom frequency on translations by DeepL in Finnish	195
F.2	Effect of idiom frequency on translations by Google in Finnish	196
F.3	Effect of idiom frequency on translations by Google in French	197
F.4	Effect of idiom frequency on translations by DeepL in Japanese	198
F.5	Effect of idiom frequency on translations by Google in Japanese	199
F.6	Translation metrics for Finnish idiomatic sentences, for base and upweight+knn models.	200
F.7	Translation metrics for Japanese idiomatic sentences, for base and upweight+knn models.	201

LIST OF TABLES

1.1	Status of projects in the thesis. Proposed projects are marked in blue and have a rough (estimated) time of completion, whereas past projects are indicated by their conference.	5
3.1	Example sentences from the dataset	12
3.2	Examples in each data split	13
3.3	Metaphor types based on types of knowledge required (not mutually exclusive)	14
3.4	Zero-shot and finetuned test accuracies (%), finetuned is averaged across 5 seeds. Dev set accuracies can be found on the leaderboard under the "validation" split.	18
3.5	Spearman r -values and p-values between $P(y_i x_i)$ and $P(y_i)$	21
3.6	The performance of models across different commonsense categories, in terms of accuracy on examples annotated with that category (%). The strongest category of each model is highlighted.	23
3.7	Examples of idioms with their matched phrases, selected based on having the same syntactic pattern and most similar log frequency in the Syntactic Ngrams dataset. Examples depicted here have the same log frequency. Note that the frequency is based on the most common dependency and constituency pattern found in Syntactic NGrams. Humans were asked to rate each phrase for its compositionality.	30
3.8	Phrases rated most and least compositional using RoBERTa _{CLS} representations, from several syntactic phrase types. "... indicates that a phrase continues but is too long to display. Long phrases and abbreviated names tend to have a higher compositionality score.	32
3.9	Example judgments of one annotator on the pilot set. Annotators were asked to rate each phrase from 1 to 3, where 1 meant not compositional and 3 meant fully compositional. They were also asked how much each subphrase contributed to the meaning.	36

3.10	Spearman correlation between human judgments of compositionality and compositionality score generated by different model and representation combinations. P-values are corrected for multiple comparisons with the Holm-Bonferroni correction.	37
3.11	Accuracy of model representations on the subphrase test and idiomaticity test.	38
3.12	Most and least compositional phrases in CHIP by human judgments and RoBERTa compositionality scores. Human scores are the average of 3 annotators.	39
3.13	Summary of reasoning types and analogues in language models. Citations to instruction following, in-context learning, and chain-of-thought are limited to the original paper due to the high number of papers on these topics, while we have tried to list all currently available papers on instruction inference (for abductive reasoning) with LMs.	40
3.14	Examples of in-context examples, queries, and hypotheses in each domain.	43
3.15	Summary of results. A checkmark indicates that the property held for all or almost all language models, a half-checkmark indicates a partial success for all or almost all language models, while an X-mark represents lack of success for most language models.	45
3.16	Spearman correlation and p-values of true vs predicted coefficients for each model in the functions domain.	47
4.1	Downstream evaluation datasets that we predict performance on.	58
4.2	MAE comparison across predictors for each downstream dataset.	61
4.3	Midtraining mixes used in our experiments and dataset(s) from which they were derived.	69
4.4	SFT and C4 validation losses for 410M model	70
4.5	SFT and C4 validation losses for 70m and 160m models comparing default midtraining mixes to mixes of only the midtraining data (marked with 100%), averaged across 5 seeds for each SFT dataset. Bold values indicate best performance within each dataset and model size combination. Results reflect two continued pretraining trials in the code and math domains.	74
5.1	Size of test sets for each language. The idiomatic and literal sentences contain strings matching known idioms (after lemmatization), and the in-domain random set contains unrelated sentences from OpenSubtitles, but the out-of-domain random set contains unrelated sentences from the Ted Talks corpus.	86
5.2	Performance of commercial systems on idiomatic, literal, and random test sets. There is a clear degradation in performance on idiomatic sentences.	89
5.3	Human-judged accuracy on sentence-level semantics.	94
5.4	Rate of major and severe errors in translations.	96

A.1	Examples that were rejected due to being nonsensical.	106
A.2	Examples that were rejected due to having significant spelling or grammar errors.	106
A.3	Examples that were rejected due to not following the specified format. . .	107
A.4	Zero-shot and finetuned backward auto-regressive model accuracies on the test set	107
A.5	Accuracy for models on the test set, counted in terms of pairs of sentences in which both are correct (%). Results are from one run.	108
A.6	Accuracy breakdown and frequency of parts of speech in metaphor subjects. Only part-of-speech patterns with greater than 10 occurrences are shown.	108
A.7	Accuracy breakdown and frequency of parts of speech in metaphor relations. Only part-of-speech patterns with greater than 10 occurrences are shown.	109
A.8	Accuracy breakdown and frequency of parts of speech in metaphor objects. Only part-of-speech patterns with greater than 10 occurrences are shown.	109
A.9	Accuracy breakdown and frequency of WordNet hypernyms in metaphor subjects. Only hypernyms with 10 or greater occurrences are shown. . . .	110
A.10	Accuracy breakdown and frequency of WordNet hypernyms in metaphor objects. Only hypernyms with 10 or greater occurrences are shown. . . .	111
A.11	Examples of completions generated by GPT-3 Davinci.	111
B.1	Counts of the top 20 grammatical tree types found in the WSJ and Brown sections of the Penn Treebank, with some examples given.	113
B.2	Spearman ρ correlation between phrase length (in words) and compositionality score in the treebank.	114
B.3	Examples of idioms with their matched phrases, selected based on having the same syntactic pattern and most similar log frequency in the Syntactic Ngrams dataset. Examples depicted here have the same log frequency. Note that the frequency is based on the most common dependency and constituency pattern found in Syntactic NGrams.	114
B.4	Error ratio ($\frac{dist_{probe}}{dist_{control}}$) for probes trained to predict representations from different model types. Mean across 10 folds.	116
B.5	Mean reconstruction score (cosine similarity) and standard deviation of each approximative probe across 10 folds. Not corrected for anisotropy in each representation/model type.	124
B.6	AUC scores for probes trained on various percentages of the training set. .	126
B.7	Correlations of frequency and length with human and model compositionality scores. Corrected with Holm-Bonferroni correction.	131

C.1	Prompts for the functions domain. Newlines are depicted visually for ease of reading. Variables substituted per question are marked like {this}. The wording for the “prompt with self-induced hypothesis” and “zero-shot chain-of-thought prompt” was slightly changed from the few-shot examples prompt to reduce formatting errors.	135
C.1	Prompts for the functions domain. Newlines are depicted visually for ease of reading. Variables substituted per question are marked like {this}. The wording for the “prompt with self-induced hypothesis” and “zero-shot chain-of-thought prompt” was slightly changed from the few-shot examples prompt to reduce formatting errors.	136
C.1	Prompts for the functions domain. Newlines are depicted visually for ease of reading. Variables substituted per question are marked like {this}. The wording for the “prompt with self-induced hypothesis” and “zero-shot chain-of-thought prompt” was slightly changed from the few-shot examples prompt to reduce formatting errors.	137
C.1	Prompts for the functions domain. Newlines are depicted visually for ease of reading. Variables substituted per question are marked like {this}. The wording for the “prompt with self-induced hypothesis” and “zero-shot chain-of-thought prompt” was slightly changed from the few-shot examples prompt to reduce formatting errors.	138
C.2	Prompts for the colours domain. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The “prompt with a self-induced hypothesis” was slightly modified from the base prompt in order to encourage models to follow the correct formatting, while the “prompt for zero-shot chain-of-thought” was slightly modified to encourage models to generate a concrete chain of thought, and also to follow the correct formatting.	140
C.2	Prompts for the colours domain. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The “prompt with a self-induced hypothesis” was slightly modified from the base prompt in order to encourage models to follow the correct formatting, while the “prompt for zero-shot chain-of-thought” was slightly modified to encourage models to generate a concrete chain of thought, and also to follow the correct formatting.	141
C.2	Prompts for the colours domain. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The “prompt with a self-induced hypothesis” was slightly modified from the base prompt in order to encourage models to follow the correct formatting, while the “prompt for zero-shot chain-of-thought” was slightly modified to encourage models to generate a concrete chain of thought, and also to follow the correct formatting.	142

C.2	Prompts for the colours domain. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The “prompt with a self-induced hypothesis” was slightly modified from the base prompt in order to encourage models to follow the correct formatting, while the “prompt for zero-shot chain-of-thought” was slightly modified to encourage models to generate a concrete chain of thought, and also to follow the correct formatting.	143
C.3	Vocabulary induction correctness and relation to few-shot translation correctness for gpt models. Significant correlations after correction for multiple comparisons are bolded.	145
C.4	Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. “” indicates that the prompt is the same as in § C.6.	146
C.4	Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. “” indicates that the prompt is the same as in § C.6.	147
C.4	Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. “” indicates that the prompt is the same as in § C.6.	148
C.4	Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. “” indicates that the prompt is the same as in § C.6.	149

C.4	Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. New-lines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. "" indicates that the prompt is the same as in § C.6.	150
C.4	Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. New-lines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. "" indicates that the prompt is the same as in § C.6.	151
C.4	Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. New-lines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. "" indicates that the prompt is the same as in § C.6.	152
C.5	Overall grammar sketch accuracy of models. Unsure answers were marked as incorrect.	153
C.6	Accuracy and segment-level chrF for vocabulary hypotheses in Kalamang translation. If morphology was excluded, all words with a * or - symbol in the translation (corresponding to prefixes/suffixes) were excluded, whereas otherwise we matched whether or not the model's hypothesis began/ended with the correct characters.	154
C.7	Point-biserial correlation between segment-level chrF of each translation with the correctness of vocabulary hypotheses induced for that question (in the instruction-inference:verbalized confidence setting). No correction for multiple comparisons was done because we examined hypotheses at the model level (see § C.8 for how this was done in the colours domain)	155
C.8	Example translations by the GPT-3.5-turbo model in different settings. . .	156
D.1	Model Parameter Counts by Organization (sorted by size)	158
D.2	Number of models evaluated for each benchmark task. Note that some models encountered technical errors when being loaded or in lm-eval-harness. The number of models will continue to be updated.	170
D.3	Overview of R^2 values by benchmark.	171

D.4	Domain correlations between pretraining data and free-generations. Web content shows the strongest Pearson correlation, suggesting models most faithfully reproduce web distribution patterns, while books content shows the weakest relationship.	171
D.5	Linguistic feature correlations between pretraining data and free-generations. Connective elements like conjunctions show stronger correlations, while structural features like character length are less preserved in model generations.	171
D.6	Greedy-selected features per benchmark.	174
D.7	MAE comparison of Scaling Laws and All Features predictors versus Baseline. Note that a significance test was not carried out for LGBM, so this reflects results for one run, though a hyperparameter search is still carried out over the same values as in § D.9 for both predictors. Brier scores are scaled $\times 100$ for comparability. Both predictors here use LGBM.	175
D.8	Loss-based evaluation metrics across different data mix ratios	175
D.9	TruthfulQA evaluation metrics across different web/other ratios	176
E.1	Core pretraining hyperparameters for Pythia-70M, 160M, and 410M. . . .	179
E.2	Finetuning datasets.	180
E.3	Grid of candidate peak learning rates swept during tuning.	180
E.4	Selected peak learning rates for fine-tuning (cosine schedule with 10% warmup).	180
E.5	SFT and C4 validation losses for 70m and 160m models across downstream datasets and midtraining mixtures, averaged across 5 seeds for each SFT dataset. Bold values indicate best performance within each dataset and model size combination.	182
F.1	Examples of categorized errors made by Δ LM.	194
F.2	p-values obtained using approximate randomization on translations produced by the base model and the upweight+knn model.	195
F.3	Results of automatic metrics. In most cases, combining loss weighting with KNN-MT improves automatic metrics the most on all three test sets, including the out-of-distribution (Random) test set.	202

CHAPTER 1

INTRODUCTION

The remarkable progress of language models, driven largely by the scaling paradigm, has polarized the research community in recent years. On one side are those that see benchmark-shattering performance as a sign that these systems will soon be able to automate all jobs that humans perform, while on the other side, skeptics point out brittleness, weakness to distribution shifts, and alignment failures as fundamental limitations. This debate persists because we conceptualize models as monolithic entities with a given level of performance, rather than systems with different cognitive capabilities which develop at different rates according to their characteristics and training. When a model outperforms humans on professional exams, yet fails on simple logical reasoning problems, or excels at creative writing while struggling with simple semantics, what are we to make of this? These observations are not necessarily contradictory, but we lack systemic frameworks below the granularity of task performance to explain why they occur. This leaves researchers to draw conflicting conclusions from the same evidence based on their prior beliefs.

A more productive approach may be to focus on the growth and interaction of specific capabilities over the course of training rather than focusing on final benchmark scores. To break this impasse, we propose to draw inspiration from developmental psychology, which asks similar questions in the case of humans: How do complex cognitive abilities emerge from simpler foundations, and when and why do they emerge [135]? Although humans are very different from language models as learning systems, it is worth thinking through what it would mean to have an equivalent developmental science for language models. Although this may seem like a stretch, I will argue that it is not only practical to think in such a way, but that it could accelerate development towards adapting language models to increasingly difficult and open-ended tasks.

As a starting point, we need to establish what capabilities we are looking for and how to measure them reliably. Cognitive science has identified a handful of abilities that are foundational to human generalization. Analogical reasoning enables us to transfer knowledge from one domain to another by mapping relational structures; composition-

ality underlies our ability to build complex meanings from simpler parts while handling exceptions like idioms; and the three forms of inference – deduction (deriving guaranteed conclusions), induction (generalizing explanations from examples), and abduction (inferring the most plausible explanations) – drive our capacity to learn and apply rules that we transfer to new situations. In **Part I** (Chapter 3), we evaluate precisely these capabilities in language models. Our findings show unexpected patterns: models partially succeed at tasks such as interpreting figurative language, representing compositional and non-compositional expressions, as well as inducing the best explanation for few-shot examples and instructing themselves. However, we also find surprising deficits arising from over-reliance on surface frequency, poor representation of long-tail phenomena such as idioms, and dissociations between capabilities that are often linked in humans, such as answering a question and explaining how to answer a question.

These uneven capability profiles point to a deeper issue with how models develop. Currently, language models undergo an inflexible pipeline from pretraining to posttraining in which they consume vastly more textual data than humans see in their lifetimes, yet this approach yields inconsistent results. While some capabilities scale predictably with model size, others have more unpredictable patterns, scale inversely or suddenly “emerge”. This suggests that certain capabilities may benefit from specific developmental conditions, which we can study empirically.

In **Part II** (Chapter 4), we therefore shift from measuring capabilities to understanding their emergence during training. We first investigate what factors beyond scale predict performance on downstream benchmarks, discovering that although model size and data size explain most variation, in all examined benchmarks data composition and other features can improve predictions of a model’s performance. We then examine the critical “midtraining” phase between pretraining and supervised fine-tuning, finding that it effectively serves to “bridge” pretraining and posttraining token distributions, leading to better in-domain performance as well as reduced forgetting. Finally, we propose decomposing complex capabilities into elemental skills whose emergence can be tracked and accelerated through targeted interventions. Together, these studies reveal that capability development follows predictable patterns that we can measure and influence.

Finally, **Part III** (Chapter 5) translates these insights into actionable interventions. Rather than hoping capabilities emerge from scale or fixing problems post-hoc, we demonstrate targeted developmental interventions. In machine translation, we show how identifying and addressing a specific developmental gap (recognizing idiomatic expressions) can improve performance without collateral damage. Our proposed work extends this principle to more complex domains: the first proposed work in this section is on decomposing repository-level code generation into simpler skills which we can create large-scale training data for, while the second is on making language model training more flexible through extending the framework of reinforcement pretraining [65].

1.1 Thesis Overview

In this section, we detail the contributions in each chapter, and sketch out a timeline for completing the thesis. We plan to complete the proposed work in the next two years, with the approximate timeline given in [Table 1.1](#).

- In Chapter 3, we focus on evaluating several capabilities key to reasoning in humans.
 - In Section 3.2, we create resources for evaluating the understanding of novel figurative language in English. We test whether models can correctly interpret metaphorical phrases through a minimal inference task – choosing between opposing interpretations of novel metaphors. We found that state-of-the-art LMs of the time performed markedly worse than human performance even after fine-tuning, suggesting that understanding figurative language represented a bottleneck in semantic understanding at the time. This may remain true for low-resource languages now [130], though this has improved considerably in English since the time of this paper [283].
 - In Section 3.3, we study the compositionality of phrase representations in language models through approximating the composition operation with simpler functions. We find that LM representations are effectively approximated through affine transformations of the inputs, but that conversely non-compositional expressions such as idioms and multi-word expressions are not represented in a non-compositional manner as expected.
 - In Section 3.4, we study the relationship between deductive, inductive, and abductive reasoning in LMs. We present a simple way to improve performance on structured tasks by having LMs propose and rerank rules underlying in-context examples, inspired by abductive reasoning. However, we show that counterintuitively, these general modes of reasoning are not related, and that despite success on simple tasks, it is difficult to scale abductive reasoning to difficult tasks such as documenting and translating unknown languages.
- In Chapter 4, we instead focus on the acquisition of individual abilities during pretraining.
 - In Section 4.2, we ask what architectural or data-distributional patterns predict performance across diverse tasks over and above traditional scaling factors (number of trainable parameters and data size). We rediscover phenomena like the optimal code percentage in pretraining to balance natural language reasoning and coding (15-25% and task specific effects such as the negative impact of web data on truthfulness. This shows that different tasks may benefit from different types of data and training, rather than uniformly benefitting from scale, and that some counterintuitive scaling like inverse or U-shaped scaling may be due to data composition or other factors.
 - In Section 4.3, we investigate the “midtraining” stage of training between the end of pretraining and the beginning of posttraining. Through controlled ex-

periments, we find that midtraining is most effective for domains that are distant from general web text (such as math and code), when the midtraining domain aligns with downstream posttraining tasks. Additionally, midtraining reduces catastrophic forgetting compared to both direct fine-tuning and continued pretraining, with the timing of specialized data introduction being more critical than mixture composition. This establishes midtraining as an effective domain adaptation technique that bridges the distribution gap between pretraining and specialized posttraining, demonstrating that LMs benefit from smooth transitions rather than abrupt distributional shifts.

- In Section 4.4, we propose breaking down complex tasks such as factual QA into a composable set of atomic skills, which are easily evaluable and for which we can easily generate synthetic data. We propose to study the emergence of these elemental skills and their compositions in training, and show that we can accelerate the learning of elemental tasks using our synthetic data and this transfers to real data.
- In Chapter 5, we pivot from diagnosis and understanding to targeted model development.
 - In Section 5.2, we address a persistent gap in translating non-compositional expressions such as idioms and multi-word expressions. Through controlled experiments with the “elemental task” of non-compositional translation, we find that this skill only emerges with large data volumes. By combining retrieval augmentation with loss reweighting for sentences with potentially idiomatic expressions, we improve translation of these phrases across three languages without degrading overall translation.
 - In Section 5.3, we propose to break down the task of repo-level coding in order to improve open-source coding agents. This represents a practical use of the approach in Section 4.4. Repo-level coding is a long-horizon task that is compositionally structured, yet has sparse data. This is because naturally-occurring training data in the form of pull requests with tests is relatively rare, limiting the training data available. However, the individual steps that go into implementing a feature or bug fix are not necessarily difficult to define, consisting of skills such as code localization, documentation generation, and refactoring based on small changes. By defining these elemental skills related to coding, we can potentially generate a large amount of synthetic data from repositories on github without natural PRs, and compose these skills to make more complex trajectories. We aim to improve the performance of open coding models on downstream benchmarks such as SWE-bench [127] and Commit0 [314] with targeted interventions at the posttraining and midtraining stages.
 - In Section 5.4, we propose to make the training pipeline more flexible in the context of open-ended reasoning domains where verification is subjective or partial. While recent RL approaches for reasoning succeed in closed domains with clear correctness signals such as math and coding [99], many real-world

capabilities like scientific reasoning and creative problem-solving lack binary feedback. We propose to investigate how models can develop autonomous reasoning improvement through: (1) reinforcement pre-training that encourages exploratory hypothesis formation beyond next-token prediction, (2) graded reward systems that capture partial correctness and logical coherence rather than binary truth, and (3) autonomous discovery of evaluative principles from contrastive examples, akin to in Section 3.4. This approach mirrors how humans learn in open-ended domains – through self-assessment, hypothesis testing, and self-selecting learning materials without constant external validation.

Chapter	Referenced Papers	Status
§3	Testing the Ability of Language Models to Interpret Figurative Language	NAACL 2022
	Are Representations Built from the Ground Up? An Empirical Examination of Local Composition in Language Models	EMNLP 2022
	An Incomplete Loop: Instruction Inference, Instruction Following and In-Context Learning in Language Models	COLM 2024
§4	Not-Just-Scaling Laws: Towards a Better Understanding of the Downstream Impact of Language Model Design Decisions	Submitted to EMNLP 2025
	Midtraining Bridges Pretraining and Posttraining	To be submitted to ICLR 2026
	Elemental Tasks	November 2025
§5	Crossing the Threshold: Idiomatic MT through Retrieval Augmentation and Loss Weighting	EMNLP 2023
	Case study: Developmental training of repo-level code agents	February 2026
	Adaptive Reasoning Development	January 2027

Table 1.1: Status of projects in the thesis. Proposed projects are marked in blue and have a rough (estimated) time of completion, whereas past projects are indicated by their conference.

1.2 Proposed Timeline



Aug 2025	•	Thesis Proposal
Aug–Nov 2025	•	Elemental Tasks (ongoing)
Oct 2025–Feb 2026	•	Case Study: Repo-level Code Agents (ongoing)
May–Aug 2026	•	Summer Internship
Sep 2026–Jan 2027	•	Adaptive Reasoning Development
Oct 2026–Apr 2027	•	Job Hunting
Feb–Apr 2027	•	Thesis Finalization
May 2027	•	Thesis Defense

CHAPTER 2

BACKGROUND

Note: This section will be expanded in the final dissertation document.

2.1 The Lifecycle of a Language Model

The dominant approach to training language models typically follows a progression through increasingly specialized stages. **Pretraining** involves training autoregressive language models – which estimate the probability $p(y_t|y_1, \dots, y_{t-1})$ of each token given previous context – on massive, diverse web-scale corpora using enormous computational resources [133]. **Supervised fine-tuning (SFT)** then narrows the focus to smaller, curated datasets of instruction-response pairs, teaching models to follow directions and behave appropriately [202]. Finally, **reinforcement learning methods** such as RLHF and RLVR use human preference feedback or AI-generated rewards to fine-tune alignment and safety [15, 99], operating at the smallest scale with the most constrained data. After a model is deemed sufficiently trained – usually after achieving acceptable scores on standard benchmarks compared to competing models—its weights are frozen and users query it through **inference**.

Consider a user asking: *"I'm planning a trip to Japan in cherry blossom season. What are the best places to visit, and when should I book my flights for the best prices?"* A successful response requires the model to integrate diverse knowledge: geographical facts about Japan, seasonal timing of cherry blossoms, travel logistics, and economic patterns of airline pricing. The model must also understand the conversational format, provide a helpful structure, and avoid potentially harmful advice. But how exactly does each training stage contribute to producing such a response?

The general intuitive understanding is that pretraining establishes broad world knowledge (knowledge about Japan and the best places to visit), SFT creates a reliable instruction-following foundation (ability to respond reliably to the user query rather than simply writing an article about Japan), and RL alignment refines behavior according to some

human values (knowledge of general preferences on travel and good ways to format the response to the query). Each stage provides a stable foundation for the next while progressively constraining the model’s behaviour. However, this ordering appears to be driven by both technical stability considerations and computational convenience, rather than principled understanding of optimal capability development sequences. Moreover, once deployed, models become frozen at inference time, unable to learn from user interactions. Reasoning models do recover some flexibility through search and other forms of scaffolding at test time, but the success of test-time techniques is also dependent on the performance of the base model [252, 288]. This rigid pipeline leaves us with limited understanding of how specific capabilities emerge, interact, or can be targeted for improvement. This motivates the need for a more systematic, capability-centered approach to model development.

2.2 Lessons from Human Cognitive Development

In order to better optimize model development and understand current gaps in capabilities, we need a framework for understanding how complex and generalizable behaviours develop from simpler patterns in response to the environment. Human cognitive development provides such a framework: research has identified core skills such as analogical reasoning, compositionality, and learning from instructions that enable efficient generalization. More importantly, developmental psychology offers methodologies for tracking how these capabilities emerge, interact, and build upon each other over time. We do not claim that language models or humans share similarities in how they learn, but it is worth considering if we can also develop a similar framework for language models that would motivate targeted improvements.

2.2.1 Capabilities that enable transfer learning

We focus on a few key capabilities that enable cognitive flexibility in humans, although this is not an exhaustive list. We focus on language specifically, and highlight how each capability would be useful for language models.

Analogical reasoning Analogical reasoning is the ability to recognize common structures across situations,¹ and to transfer things learned in one situation to another based on this shared structure [89, 152]. In language, this is expressed through verbal metaphors. Although it may not seem essential for language models to possess the ability to reason analogically beyond just being able to interpret figurative language generated by humans, it is clear that better analogical reasoning could greatly decrease data needs of models.

¹More precisely, isomorphisms between situations

Compositionality and idiomaticity Many have argued that cognition in humans depends on composition of mental representations, which is to say that the representations of more complex phrases, actions, or routines can be formed systematically from simpler representations [265]. In terms of evaluation, we generally refer to compositional generalization: if a model is able to do task A and is able to do task B separately, it should be able to do a task that combines A and B [139, 150]. There are many benchmarks that measure compositional generalization [139, 150], and arguably it remains a major challenge today even in very strong models [187]. However, language is not always compositional, and many phrases such as “break a leg” (good luck) cannot be decomposed like this. Therefore, this poses a dual challenge for language models: they must learn when to compose meanings systematically and when to treat expressions as non-decomposable units. Failure to handle non-compositionality leads to literal mistranslations of idioms, misunderstanding of cultural expressions, and inability to recognize when conventional meanings override compositional interpretation [58].

Abductive Reasoning and Hypothesis Testing Abductive reasoning involves inferring the rule that best explains observed experience [78, 209]. This type of reasoning often guides information-seeking behaviour, even in young children [153]. By making hypotheses and actively testing them, humans can learn efficiently from limited data. This capability could be valuable for language models, as they could potentially infer implicit rules from unstructured data and use these inferred rules to guide generation or further exploration.

2.2.2 Learning and Development

Developmental psychology offers a framework for understanding how complex cognitive abilities emerge from simpler foundations in humans. Systematic investigation into how children develop cognitively started with Piaget, who proposed that children progress through discrete stages: sensorimotor, preoperational, concrete operational and formal operational [211, 212]. Each stage is characterized by distinct ways of thinking and characteristic mistakes – for instance in the preoperational stage, children mistakenly think that pouring water from a fat, wide glass into a tall, thin glass increases the volume of water in the glass because they have not internalized the concept of conservation of volume.

While Piaget’s discrete stages provided a foundation for the field, modern theories have moved in a more nuanced direction. Siegler’s “overlapping waves” theory [249] proposes that children simultaneously use multiple strategies for problem-solving, with more advanced approaches gradually becoming dominant rather than suddenly replacing earlier ones. This parallels findings in language models, where multiple solution strategies can coexist during training, with models initially relying on shortcuts before eventually developing more general solutions, as seen in phenomena like grokking [217] and mechanistic studies of emergent generalization [193]. Similarly, Karmiloff-Smith’s theory of representational redescription [134] challenges both strong nativist views (that

modules are prespecified) and pure constructivist accounts (that everything is learned from scratch). Instead, she argues that development involves making implicit procedural knowledge progressively more explicit and flexible.

Beyond these general frameworks, we also focus on two general concepts from developmental science in the context of language models:

Critical periods in development While some neural pathways are highly plastic throughout development, while others (mostly those controlling sensory and motor function) are essentially frozen, with no input from experience [141]. However, most circuits are somewhere in between – and for these circuits, there is often a *critical period* in which the circuit is particularly sensitive to experience, and inputs received can greatly alter the pattern of connectivity. There are many examples of critical periods: for instance, in the case of vision, the ability to perceive images in 3D by combining inputs from both eyes requires early experience with properly aligned binocular vision [56, 121]. For language development, the ability to process a language proficiently also requires early exposure to that language [128, 146, 194, 203, 284].

Dynamic interaction with environment Human learners actively shape their learning through curiosity, hypothesis testing, and self-directed exploration [94, 243]. Current language model training, by contrast, typically involves static data presentation regardless of the model’s emerging capabilities or knowledge gaps. Our proposed work (Chapter 5.4) will investigate how models can move towards more autonomous reasoning in semi-verifiable domains like scientific hypothesis formation and creative writing.

CHAPTER 3

EVALUATING COMPLEX SEMANTIC REASONING

3.1 Background

In this chapter, we evaluate several capabilities that we have previously established are important from a cognitive science perspective: analogical reasoning through figurative language (§3.2), dealing with compositionality and non-compositionality in language (§3.3), and inducing instructions from in-context examples (§3.4).

Overall, we find that while language models partially implement these mechanisms, they often exhibit surprising deficiencies and counterintuitive dissociations that differ markedly from human cognitive patterns.

3.2 Testing the Ability to Interpret Figurative Language

3.2.1 Introduction

When humans encounter metaphorical phrases like “All our words are but crumbs that fall down from the feast of the mind”[91] they readily interpret the intended meaning despite the non-literal language. This ability to understand figurative language, which occurs fairly often in human speech and writing [186], has been viewed as a bottleneck in natural language understanding [248, 268]. Figurative language often relies on shared cultural knowledge that may be difficult to acquire from language statistics alone, presenting challenges for language models trained primarily on text [19].

Most previous work at the time of this work focused on metaphor detection or paraphrase generation rather than testing genuine interpretive understanding [247, 272]. In this work, we ask whether language models can correctly make inferences regarding creative, relatively novel metaphors generated by humans. This task is harder because (1) inference requires understanding underlying semantics rather than surface patterns, and (2) our metaphors are novel creations that may not appear in training data. We propose a minimal task where models choose between opposing inferences from metaphor-

ical premises, such as determining whether “Her commitment is as sturdy as oak” implies she was “committed” or “uncommitted.”

Paired sentences	Possible answers
The pilot flew like a <u>ballet dancer</u> The pilot flew like a <u>modern dancer</u>	The pilot flew in a (restrained way creative way) The pilot flew in a (restrained way creative way)
The meteor was as bright as New York City The meteor was as bright as <u>coal</u>	The meteor was (very bright not bright at all) The meteor was (very bright not bright at all)
The atom is like a solar system The atom is like a <u>cloud</u>	Electrons (orbit the nucleus are in probability densities) Electrons (orbit the nucleus are in probability densities)
He hustles like rent was due three days ago He hustles like <u>he’s a billionaire’s son.</u>	He (hustles hardcore. doesn’t hustle at all.) He (hustles hardcore doesn’t hustle at all)
Life is as easy as <u>kindergarten for a high school senior</u> Life is as easy as <u>kindergarten for a newborn</u>	Life is (basic beyond comprehension) Life is (basic beyond comprehension)

Table 3.1: Example sentences from the dataset

3.2.2 Dataset Creation and Validation

Crowdsourcing Task

We crowdsourced data from workers on Amazon Mechanical Turk (details in § A.1). Workers were asked to generate paired metaphors with different meanings, as well as literal implications of the two metaphors in context. We instructed workers to try to generate rare or creative metaphors, namely “metaphors that would not appear often in text on the internet, books, social media, or news sites, but that can still be easily understood by people.” Workers were given examples of valid pairs that fit the format, and examples of invalid ones to discourage errors. Some examples of generated pairs are shown in Table 3.1.

To help workers, we employ the *randomness as genesis* and *narrow limits of change* principles of Cognitive Load Theory [264]. To add soft constraints, we generate 3 different random words to be shown to each batch of workers. However, workers were not required to use these words, as we wanted to encourage maximum diversity. To ensure that the random words were metaphorically rich, we selected them based on the metaphorical frames in Lakoff and Johnson [152].

Data Validation

The dataset was manually validated by three authors of this article. Each author covered roughly one-third, evenly split between training, validation, and test. Examples were excluded if they (a) did not make sense given the figurative expression, (b) had grammar or spelling errors that rendered them unintelligible, or (c) did not follow the format of the task. Examples of excluded samples are included in § A.2. We collected 13,324 sentences and interpretations from the crowdsourcing task, and 10,256 sentences remained after filtering.

Final Dataset

The release version of our dataset contains the named data splits in [Table 3.2](#). The medium train, dev, and test splits were generated by partitioning the first stage of the data collected. The large train split additionally contains all the new examples from the second collection stage, and the small train split is a small random sample.

S	Train M	L	Dev	Test
200	1,458	8,016	1,094	1,146

Table 3.2: Examples in each data split

3.2.3 Figurative Language Typologies

In this sample, we perform an analysis of the collected data to demonstrate its trends and categorize examples for further error analysis. Specifically, we examine (a) subjects, objects, and relations, and (b) types of common-sense knowledge needed to interpret the metaphor.

Figurative Language Structure

We note that most metaphors and similes can be characterized by three components, (S, R, O) , where S is a subject, R is a relation, and O is an object. For instance, "Her commitment is as sturdy as plywood" can be written (Her commitment, sturdy, plywood). Interpretation involves inferring an attribute of the subject by extracting a relational attribute from the object [72]. In a simile, R is explicit, whereas it is usually implicit in a metaphor. The most common subjects, relations, and objects in the medium train dataset are shown in [Figure 3.1](#). These were obtained by first segmenting the phrases with syntactic patterns constructed from observation, followed by lemmatization and removal of punctuation and determiners "the", "an", "a" and "that". There are 441 unique subjects, 646 unique relations, and 1,198 unique objects in the medium training set.

Type of knowledge required	Paired sentences
Common-sense (objects)	The new mattress is just as comfortable as sleeping on a (cloud/rocks outside)
Visual	The professor’s argument had the clarity of a (crystal glass/marine fog)
Common-sense (social)	She is as embarrassed as a kid that (forgot homework/got an A)
Cultural	The construction was as disastrous as the (1981 musical Cats/The 2019 film based on the musical Cats)

Table 3.3: Metaphor types based on types of knowledge required (not mutually exclusive)

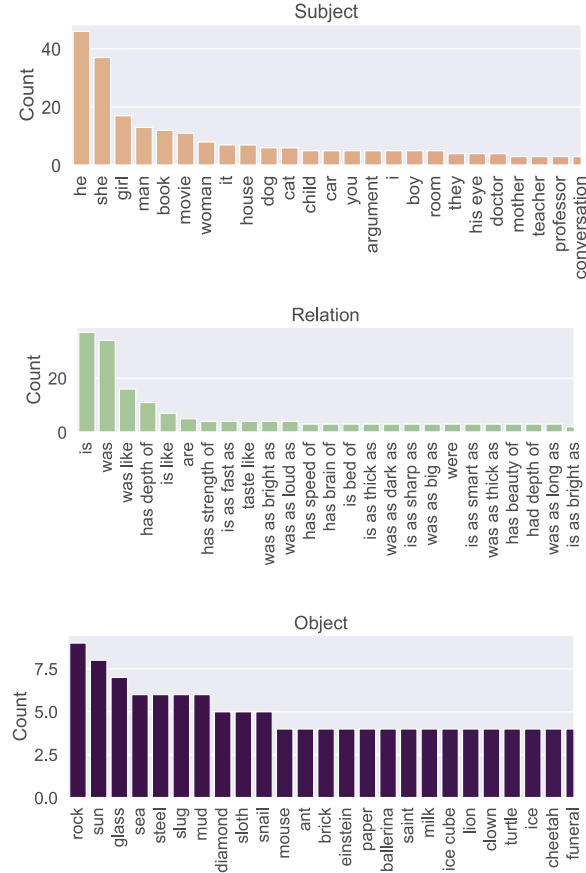


Figure 3.1: Visualization of 25 most frequent subjects, relations, and objects in the medium train set.

Common-sense Knowledge Types

Next, we examined the test set to determine the types of commonsense knowledge needed to interpret metaphors. Through thematic analysis, we devised 4 categories based on common-sense knowledge, which are not mutually exclusive: common-sense object knowledge, visual metaphors, common-sense social understanding, and cultural knowledge. The same 3 authors of the paper annotated the test set for these categories, with annotators responsible for separate categories.

Common-sense object knowledge consisted of metaphors that made reference to properties of common objects and animals, such as volume, height or mass of objects, or properties of materials. 68.35% of the test set was found to require common-sense object knowledge.

Visual metaphors were a subset of common-sense object metaphors, based primarily on the visual modality, including attributes such as brightness or colour. Some visual metaphors also sketched a vivid visual scene. These examples comprised 14.73% of the test set.

Common-sense social understanding examples required knowing how humans would react in different circumstances or required knowing about human emotions. These examples comprised 27.55% of the test set.

Cultural metaphors required knowing cultural traditions, works of art/artefacts, or religion. Due to crowdworkers being recruited from the US, these are centered around US culture. These examples comprised 16.56% of the test set.

3.2.4 Baseline Models and Evaluation

Auto-regressive Language Models

Auto-regressive LMs generate a probability distribution of the next token given all preceding tokens. As such, we can directly compute the probability of a sentence by multiplying the conditional probability of each token at each time step.

$$\tilde{P}(w_1 \dots w_N) = p(w_1) \prod_{i=2}^N p(w_i | w_1 \dots w_{i-1})$$

The ability to directly extract probabilities enables the *zero-shot* reasoning of these LMs. For a pair of metaphorical expressions x_1 and x_2 with two corresponding interpretations y_1 and y_2 , we feed in the concatenation of the metaphor and the interpretation to the pretrained model without finetuning. We define “forward” and “backward” probabilities assigned to interpretations and figurative language expressions, respectively. For the **forward probability**, for figurative phrase x_i and correct answer y_i , we take

$$P(y_i | x_i) = \frac{P(x_i, y_i)}{P(x_i, y_i) + P(x_i, y_j)}$$

since there are only two answer options. From this, we can calculate accuracy when we take the indicator of $P(y_i | x_i) > 0.5$. Similarly for the **backward probability** (predicting phrase based on answer), we take

$$P(x_i | y_i) = \frac{P(x_i, y_i)}{P(x_i, y_i) + P(x_j, y_i)}$$

with analogous backward accuracy.¹

We examine three state-of-the-art large transformer-based LMs of this category: **GPT-2** (with 117M parameters, trained on 40GB of text), **GPT-neo** (with 1.3B parameters, trained on 800GB of text) and **GPT-3** (4 variants between 350M and 175B parameters, trained on 45TB on text) [29, 34, 226]. We also examine the performance of these models after finetuning on the training data. GPT-2 and GPT-neo were trained with a batch size of 8, with early stopping with patience of 1 epoch, and a minimal hyperparameter search was performed with learning rates 1e-5 to 5e-5. GPT-3 was trained with the default parameters of the GPT-3 finetuning API.

Masked Language Models

We also evaluate the performance of masked LMs on this task. Unlike auto-regressive LMs, masked LMs cannot directly output the probability of a sentence, so it is not possible to directly test the zero-shot performance of these models. Instead, we test the transfer performance by first finetuning them in two ways: on WinoGrande, which is also a binary choice task based on common-sense reasoning, and on several NLI datasets, including SNLI, MNLI, FEVER-NLI and ANLI [196, 237]. The input to the model trained on WINOGRANDE is formatted as [CLS] [metaphor] [SEP] [answer1] [SEP] [answer2], and we use an extra linear layer on the [CLS] token embedding to perform the classification. In addition to the transfer performance, we also use contrastive fine-tuning by feeding in each metaphor along with both answer choices, and training the model with our dataset to classify which answer is correct. For the NLI model, we examine accuracy using all three labels the model was originally trained with (entailment, neutral, and contradiction), as well as using a forced binary choice paradigm in which the logits for the contradiction label are subtracted from the logits for the entailment label, and the higher “entailment score” is the ending the model predicts. We treat these two conditions as the analog of “zero-shot” for these models.

We examine two masked LMs that are commonly used as baselines on many NLP tasks: **BERT** [62], a transformer-based LM jointly trained on the masked LM and next-sentence prediction objectives, and **RoBERTa** [169], an improved variant of BERT which consistently outperforms BERT across most tasks. We use the large variant of both models (350M parameters). BERT and RoBERTa were finetuned on the medium dataset for 8 epochs with batch size 8, following the setting in [237]. A hyperparameter search was done with learning rates 5e-6 to 2e-5. Both BERT and RoBERTa were used for the WinoGrande experiments, while only RoBERTa was used for the NLI experiment.

¹In actuality, we use the length-normalized probability that a model assigns to a sentence as a heuristic for the total probability, to minimize the effect that the length of a sentence has on the decision (though this is not the probability of the sequence in a strict sense): $P(w_1...w_N) = \exp(-\frac{1}{N} \log \tilde{P}(w_1...w_N))$. Initial experimentation showed marginal differences in accuracy when using these two methods, so we used normalized probabilities by default.

Forced-choice Paradigm

Due to the inherent creativity of metaphors, there may be different interpretations of the same metaphor. For instance, in Table 3.1, the example "he hustles like he's a billionaire's son" could also be interpreted in other ways, for instance "he uses his father's contacts and social privileges to make money". In a structural-mapping context, the forced choice between two answers constrains the possible meaning of the metaphor to be along one axis [88]. In this case, it would be whether or not he is required to work hard.

Of course, many of these metaphors have other valid interpretations. In the "billionaire's son" example, another valid axis of interpretation could be the manner in which he works. For instance, the alternative pair could be "he hustles like he's a (billionaire's son | single mother working three jobs)" with answers "he (uses his contacts and social privileges to make money | works extremely long hours with multiple ventures to make money)". It is possible that LMs could come up with other valid interpretations that are not the ones originally intended, motivating us to also look at generation performance in section § 3.2.5.

Human Performance

To estimate the expected human performance on this task, we ran a benchmark on the test set with 10 human volunteers who were not the authors of the article. The human annotators were not shown any training examples, so this would be equivalent to the zero-shot setting for models. Participants ranged from 20 to 29 years old, and there were 5 male and 5 female participants. 5 each were native- and non-native English speakers respectively. Participants were mainly graduate student volunteers.

We shuffled the test set and split it into 10 partitions of ≈ 115 examples for each annotator. The examples were presented with pairs shuffled and separated, in order to create a better comparison with model performance.

Due to differences in vocabulary or cultural background, we instructed participants to mark examples where they weren't confident, such as those that contained words or cultural references they didn't understand.

3.2.5 Results

Inference Results

The first question is **whether strong LMs can interpret metaphors at all when presented with two opposing meanings, in zero-shot or supervised settings**. These results are presented in Table 3.4. The results for masked language models are higher than those for autoregressive language models, and fine-tuning significantly improves performance for all models.

²This is the accuracy score when transferring from Winogrande. Pretrained NLI results were 50.47 when using original labels (entailment/contradiction/neutral), and 66.32 when forcing a binary decision.

Model	Zero-shot	Tuned (L)	Tuned (XL)
GPT-2	53.93	54.80	62.65
GPT-neo 1.3B	56.89	69.98	72.00
GPT-3 Ada	59.08	69.17	73.56
GPT-3 Babbage	62.91	73.97	77.31
GPT-3 Curie	65.35	79.04	81.94
GPT-3 Davinci	68.41	-	-
BERT	58.14	83.16	85.69
RoBERTa	66.18 ²	89.22	90.32
Human	94.42	-	-
Human (confident)	95.39	-	-

Table 3.4: Zero-shot and finetuned test accuracies (%), finetuned is averaged across 5 seeds. Dev set accuracies can be found on the [leaderboard](#) under the “validation” split.

Zero-shot Performance For the zero-shot setting, we examine the test accuracy based on zero-shot forward probabilities for the GPT models, and the pseudo “zero-shot” transfer performance for BERT and RoBERTa using models pretrained on the WinoGrande task [237]. As shown, the GPT-3 models outperform the GPT-2 and GPT-neo models. Among the GPT-3 models, there is a clear correlation between model size and performance, with the largest model (GPT-3 Davinci) achieving the highest zero-shot test accuracy. BERT and RoBERTa achieve accuracy within the range of GPT-3 models. While our models mostly perform much better than chance in the zero-shot setting, there is still a large gap of 26 percentage points between our best model and human level performance.

Fine-tuned Performance For the fine-tuned setting, all models listed are fine-tuned on the small data set split. GPT models were trained with language modeling loss, whereas BERT and RoBERTa are trained with contrastive loss. We did not evaluate fine-tuning of GPT-3 Davinci due to budget. Overall, fine-tuning significantly improved accuracy for all models, with GPT-3 models uniformly improving by about 13 percentage points, and BERT/RoBERTa improving by about 25 points. Our best model after fine-tuning is RoBERTa, which is within 5% of our human performance.

Prompting We also experiment with prompting methods. First, we use a simple *suffix* prompting method, where we simply append the phrase “that is to say” between the metaphor and the interpretation, which we hypothesized may “explain” to the LM that the previous statement is figurative. We also evaluate the effectiveness of the *examples* method, by appending k random correct metaphor/interpretation pairs before the actual pair we are testing. The results of these experiments can be seen in Figure 3.2. We found that the suffix method provided a small (1-2%) improvement over the baseline, while the example method was generally ineffective.

Backward accuracies Note that the accuracies reported in this section are for the forward direction, and the backward direction is reported in § A.3. Backward accuracy is lower, with GPT-3 Curie for example having a 7% reduction in accuracy in the zero-shot case. This suggests that selecting a metaphorical expression to match a literal phrase is more challenging than the reverse for LMs.

Paired Evaluation Because our dataset is formatted as a Winograd schema, we can take advantage of *group scoring* to evaluate models more stringently [68]. We found that performance for autoregressive models plummeted under this evaluation scheme, while masked language models also suffered in accuracy. Human scores were the least affected. Details are in § A.4. This is likely related to the phenomenon found in § 3.2.6.

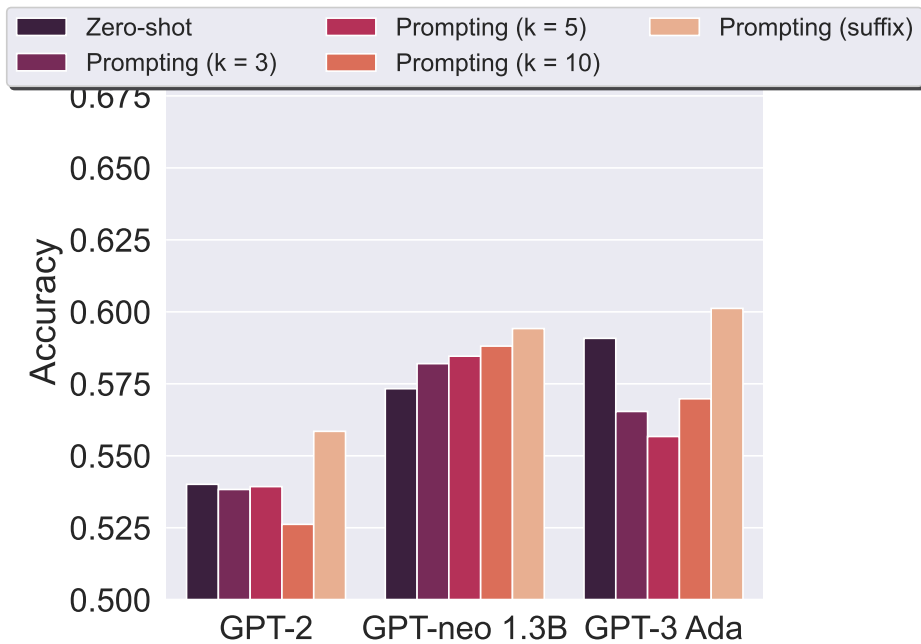


Figure 3.2: Comparison of prompting methods with autoregressive models

Generation Results

Next, we examine **if models can generate sensible interpretations for metaphors**. Given the difficulty of evaluating text generation, compounded by the difficulty of figurative language, we opted for manual evaluation of one tenth of the test dataset using generations of the strongest auto-regressive model: GPT-3 Davinci (≈ 175 B parameters).

The metaphor was given as input to the model, and 4 completions were generated for each metaphor, with a maximum length of 100 tokens. Completions were also truncated to the first sentence, as initial experiments showed contradictory statements (e.g. “he was talented. But he was not very talented”) were often generated across subsequent

sentences. Suffix prompting was also used because of the lack of context, with “That is to say,” appended to each metaphor. Only the first sentence of the output was evaluated. The temperature parameter was determined through grid search through values [0.2, 0.4, 0.6, 0.8, 1] on a small separate set of metaphors. A human annotator inspected the generated completions and found that a temperature of 0.4 produced the most correct results.

Three authors of the article labeled the completions generated by GPT-3 Davinci as correct, incorrect, or literal. In some cases, there were valid interpretations that were not the same as the answer given by crowdworkers, which were also marked correct. If the model simply restated the metaphor with no interpretation, the completion was marked as literal. Because some metaphors are ambiguous when presented without context, those examples were not counted. Inter-rater reliability was moderate due to differing standards for correctness (Krippendorff’s $\alpha = 0.5567$). The majority vote was taken between annotators’ judgments.

GPT-3 Davinci’s accuracy, counting literalized metaphors as incorrect, was 50.8%. Not counting literalized metaphors, the accuracy was 63.9%. In 37.7% of cases, GPT-3 generated contradictory completions among the 4 completions. There was at least one correct completion for 78.1% of the phrases, but only 19.3% of phrases had all completions correct. Examples of annotated generations can be found in § A.7.

3.2.6 Performance and Error Analysis

With these results in mind, we examine **what kinds of errors models make, and what factors make the task difficult.** This is covered in Section 3.2.6. We find that autoregressive models rely on the predicted probability of each answer by itself to predict the answer, and that this holds true for all models, before and after training. We find that models have difficulty in interpreting “sarcastic” metaphors and sometimes inexplicably interpret very simple metaphors wrong. We also examine error typology according to the commonsense typology of § 3.2.3 and find that models improve significantly on object, visual, and social commonsense when trained, but not on cultural commonsense.

Reliance on Probability of Answers

We find that models often rely solely on the predicted probability of answers y_1 and y_2 to make their final predictions, regardless of the context. This led the models to make the same prediction for the paired sentences in many cases. Figure 3.3 and Table 3.5 show that this trend improves with fine-tuning, and that GPT-3 is best able to disentangle the probability of y_i and the probability of $P(y_i|x_i)$, but all three models show a heavy tendency to predict based on the relative probability of an answer alone.

We hypothesize that this may be one reason why BERT and RoBERTa achieve the best finetuned performance; they use a contrastive finetuning strategy providing both the correct and incorrect options as input to the model. On the other hand, the GPT models were finetuned with only the correct option, making the comparison unfair. One way to

Model	r	p
Untrained		
GPT-2	0.8128	6.700×10^{-136}
GPT-neo	0.7891	6.075×10^{-123}
GPT-3	0.7392	4.329×10^{-100}
Trained		
GPT-2	0.6765	6.700×10^{-78}
GPT-neo	0.6689	1.456×10^{-75}
GPT-3	0.4157	2.598×10^{-25}

Table 3.5: Spearman r -values and p-values between $P(y_i|x_i)$ and $P(y_i)$

finetune GPT models contrastively is to include both options into a cleverly engineered prompt, but we leave this as a direction for future work.

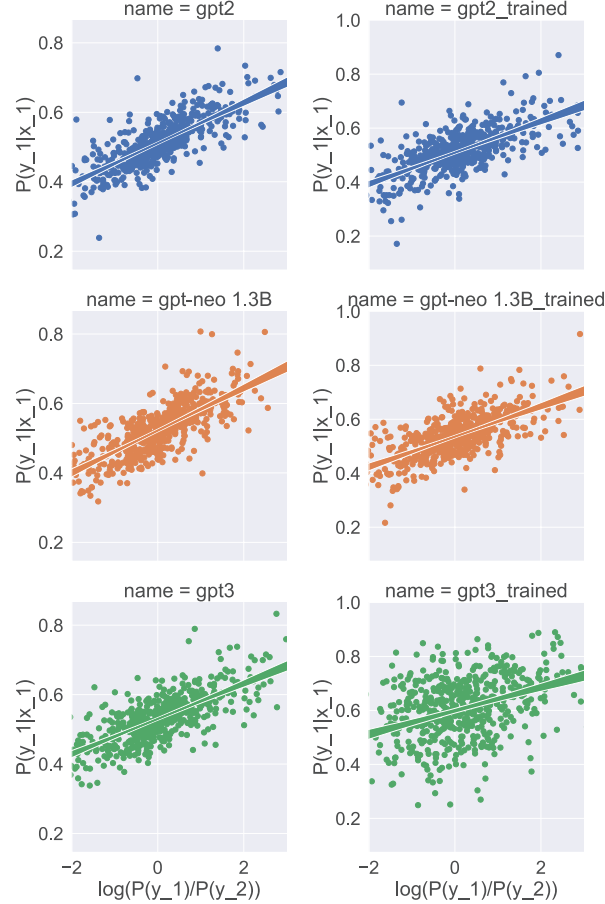


Figure 3.3: Models over-rely on predicted probability of the answer by itself to do their predictions. y -axis is predicted probability of the first interpretation (answer) given metaphor while x -axis is log odds of the first interpretation.

Other Factors Influencing Correctness

We also examined the influence of several other factors on correctness. The point-biserial correlation between length of the context phrase and the binary correctness value was -0.1544 with a p -value of 1.50×10^{-7} , indicating that longer phrases are harder to interpret correctly. The point-biserial correlation between answer probability and binary correctness was 0.1840 , with a p -value of 3.50×10^{-10} , indicating that examples where the answer was already more probable were more likely to be answered correctly, in line with our findings that models tended to predict the answer that was already more plausible alone.

Furthermore, we conducted an analysis on subjects, objects, and relations as defined in § 3.2.3. We examined accuracy by part of speech patterns in each part of the metaphor, as well as by wordnet hypernyms present in each part of the metaphor. This is detailed in § A.5 and § A.6 [73]. We used NLTK for POS tagging [174].

Qualitative Analysis of Error Trends

Common Sense Knowledge We first examine the error tendencies by the type of common sense knowledge described in § 3.2.3. Table 3.6 summarizes accuracies for these types of commonsense questions compared to humans.

Model	Obj	Vis	Soc	Cul
Untrained				
GPT-2	52.17	52.07	55.38	58.42
GPT-neo	56.38	55.62	56.01	62.10
GPT-3 Curie	75.00	71.00	72.47	78.42
Trained				
GPT-2	53.57	51.48	57.91	57.37
GPT-neo	70.15	72.78	68.67	70.00
GPT-3 Curie	87.50	84.62	83.86	83.16
BERT	87.37	92.31	84.18	77.37
RoBERTa	91.20	94.08	89.56	83.68
Human	95.41	96.45	93.99	90.00

Table 3.6: The performance of models across different commonsense categories, in terms of accuracy on examples annotated with that category (%). The strongest category of each model is highlighted.

We find that both humans and trained models tend to find object commonsense and visual commonsense metaphors easier to interpret. We find that as models improve, most of the performance gain comes from the object, visual, and social commonsense categories. Interestingly, the untrained models do quite well on cultural examples, but do not improve much on the culture category when trained. This makes sense, as cultural examples tend to be quite disparate, so training would not help as much with other examples.

Sarcastic Metaphors For both humans and LMs, many of the errors are “sarcastic” metaphors, such as saying “the girl was as bubbly as still water” to mean “the girl was bland”, rather than “the girl was vivacious”. These sentences can be difficult if the model or the human focuses on simple word association (bubbly with vivacious) without reading the entire sentence to understand the sarcasm.

Inexplicable Errors We examined the errors made by GPT-3 Curie (trained) and found that there was little overlap with the errors made by humans. Of the 64 human errors, 13 were also errors made by GPT-3. GPT-3 made many more “obvious” errors, such as predicting “The ball is a big red sun” to mean “the ball is small” rather than “the ball is big and red”. This is in contrast to sentences in which humans made errors, which often contained rare vocabulary or unfamiliar cultural references.

3.2.7 Related work

Figurative Language Identification

Most existing work focuses on identifying figurative language at the word level. The VU Amsterdam Metaphor Corpus (VUA) is the largest available corpus of metaphorical language, annotated by humans [259]. Two shared tasks on metaphor identification have been run [154, 155]. Both have utilized the VUA corpus, and the latter also introduced the TOEFL corpus, sampled from essays written by non-native English speakers [17, 155]. Most participants in the shared tasks used neural models, notably BERT, RoBERTa, and Bi-LSTMs [27, 81, 155, 219]. These models are generally improved when augmented with semantic data, such as concreteness and multimodal information.

Another line of work focuses on probing models to determine the extent of metaphor recognition. For example, BERT assigns higher pseudo-log-likelihood scores to metaphors than nonsense expressions, and its contextualized representations show some signs of contextualizing the object domain [208]. Another study uses linear probes trained on BERT layers to predict whether a word usage is literal or nonliteral, and finds that this can be done effectively, especially using middle layers as a representation [3],

Despite the utility of these tasks and datasets, they have drawbacks. Most of the metaphor use is conventional, so this task does not capture novel metaphors well. The word-level annotation also does not lend itself well to capturing extended conceptual metaphors. Finally, metaphor interpretation may be a more difficult, although less studied, task.

Figurative Language Interpretation

Recent studies mostly focus on metaphor paraphrases, either through identification [28] or generation [179, 247, 260]. However, there has not been as much work done on interpretation as on detection, and framing metaphor interpretation as a paraphrase task may not capture the emergent meaning of metaphors, such as the intended emotion, or the interaction of subject, relation and object in the metaphor [188, 268].

Other work has focused on interpreting figurative language in narratives in context, based on plausible continuations of figurative language such as idioms and similes from stories [41] or dialogues [122]. This represents a promising direction, and our work focuses on expanding our understanding of LMs’ ability to interpret non-conventionalized metaphors.

Other Figurative Language Datasets

We note that there are several other challenging NLI datasets available that contain figurative language, including the DNC corpus and the RTE dataset [42, 214]. Other datasets, such as RiddleSense, explicitly test models through difficult commonsense inference, involving figurative language [162].

Our work is distinguished by the Winograd schema format, as this format provides a better guard against the possibility that models have simply memorized common word

associations that occur in figurative language. Additionally, we specifically instructed crowdworkers to be creative, and this resulted in longer figurative phrases which require more detailed commonsense knowledge. It is likely that a fair number of these figurative phrases have never appeared in any training corpus. However, our figurative phrases also differ from riddles, as they are not supposed to be difficult to reason about, given that the source, relation, and object are properly contextualized.

Human Language Processing

Humans typically do not have any more difficulty processing metaphorical statements in context compared to literal statements [80, 92]. This may be because certain words serve as a *dual reference*, that is to say, they simultaneously refer to a physical referent and an abstract superordinate category [92]. For instance, "shark" may refer to literal sharks as well as anything that is considered vicious, leading to utterances such as "that lawyer is a shark".

Metaphorical language processing has also been studied in second-language learners, in the case of idioms. In most cases, the meaning of an unfamiliar idiom is inferred from the context or from word association [38, 54, 293].

As LMs excel at word-association based tasks, this is an encouraging finding. However, there is still a gap between LM and human performance even in our task, in which one answer is obviously wrong when the input is correctly understood.

We take into account that these results are for conventionalized figurative language and that some of the more creative phrases in this dataset may take a longer time to process for humans as well. This is especially true for non-native English speakers. However, the high human accuracy on this task with half the participants being non-native English speakers suggests that this was not a major barrier.

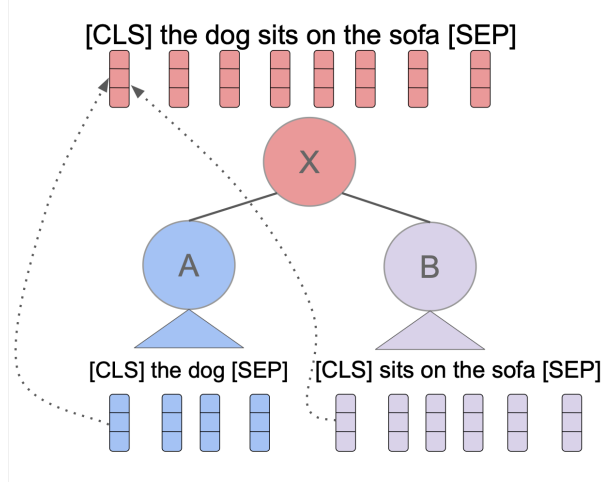


Figure 3.4: An illustration of the local composition prediction problem with [CLS] representations.

3.3 The (Non*)compositionality of Representations

3.3.1 Introduction

Compositionality—the principle that “the meaning of an expression is a function of the meanings of its parts and how they are syntactically combined” [210]—is argued to be a hallmark of linguistic generalization [266]. However, some phrases are non-compositional and cannot be reconstructed from individual constituents. A phrase like “I own cats and dogs” is compositional, whereas “It’s raining cats and dogs” is not. Language representations must handle both compositional construction and non-compositional exceptions.

Both lack [117, 149] and excess [58] of compositionality have been cited as sources of errors in NLP models. We focus on **localism**: whether the representation of a complex phrase is derivable from its local structure and immediate constituents [11, 117], rather than substitutivity-based approaches that require contrastive pairs.

We ask whether compositional probes can predict language model phrase representations from their syntactic children, which phrase types are more compositional, and whether this aligns with human judgments. Using datasets of 823K Penn Treebank phrases and human-annotated idiomatic expressions (CHIP), we find that phrase embeddings exhibit predictable affine compositional structure, with significant differences across phrase types. However, human compositionality judgments do not align well with model representation compositionality, suggesting gaps in how models capture appropriate compositional structure.

3.3.2 Methods and Experimental Details

Tree Reconstruction Error

We follow [11] in defining deviance from compositionality as *tree reconstruction error*. Consider a phrase $x = [a][b]$, where a and b can be any length > 0 . Assume we always have some way of knowing how x should be divided into a and b . Assume we also have some way of producing representations for x , a , and b , which we represent as a function r . Given representations $r(x)$, $r(a)$ and $r(b)$, we wish to find the function which most closely approximates how $r(x)$ is constructed from $r(a)$ and $r(b)$.

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \delta_{x,ab} \quad (3.1)$$

$$\delta_{x,ab} = d(r(x), f(r(a), r(b))) \quad (3.2)$$

Where $\mathcal{X}$ is the set of possible phrases in the language that can be decomposed into two parts, $\mathcal{F}$ is the set of functions under consideration, and d is a distance function. An example scenario is depicted in Figure 3.4.

For d , we use cosine distance as this is the most common function used to compare semantic vectors. The division of x into a and b is specified by syntactic structure [49]. Namely, we use a phrase’s annotated constituency structure and convert its constituency tree to a binary tree with the right-factored Chomsky Normal Form conversion included in NLTK [174].

Language Models

We study representations produced by a variety of widely used language models, specifically the base- (uncased) variants of Transformer-based models: **BERT**, **RoBERTa**, **DeBERTa**, and **GPT-2** [61, 105, 170, 225].

Representation extraction Let $[x_0, \dots, x_N]$ be a sequence of $N + 1$ input tokens, where x_0 is the [CLS] token if applicable, and x_N is the end token if applicable. Let $[h_0^{(i)}, \dots, h_N^{(i)}]$ be the embeddings of the input tokens after the i -th layer.

For models with the [CLS] beginning of sequence token (BERT, RoBERTa, and DeBERTa), we extracted the embedding of the [CLS] token from the last layer, which we refer to as the **CLS** representation. For GPT-2, we extracted the last token, which serves a similar purpose. This corresponds to $h_0^{(12)}$ and $h_N^{(12)}$ respectively.

Alternately, we also averaged all embeddings from the last layer, including special tokens. We refer to this as the **AVG** representation.

$$\frac{1}{N + 1} \sum_{i=0}^{N+1} h_i^{(12)} \quad (3.3)$$

Approximating a Composition Function

To use this definition, we need a composition function $\hat{f}$. We examine choices detailed in this section.

For parameterized probes, we follow the probing literature in training several probes to predict a property of the phrase given a representation of the phrase. However, in this case, we are not predicting a categorical attribute such as part of speech. Instead, the probes that we use aim to predict the parent representation $r(x)$ based on the child representations $r(a)$ and $r(b)$. We call this an *approximative probe* to distinguish it from the usual use of the word probe.

Arithmetic Probes In the simplest probes, the phrase representation $r(x)$ is computed by a single arithmetic operation on $r(a)$ and $r(b)$. We consider three arithmetic probes:³

$$\mathbf{ADD}(r(a), r(b)) = r(a) + r(b) \quad (3.4)$$

$$\mathbf{W1}(r(a), r(b)) = r(a) \quad (3.5)$$

$$\mathbf{W2}(r(a), r(b)) = r(b) \quad (3.6)$$

Learned Probes We consider three types of learned probes. The linear probe expresses $r(x)$ as a linear combination of $r(a)$ and $r(b)$. The affine probe adds a bias term. The MLP probe is a simple feedforward neural network with 3 layers, using the ReLU activation.

$$\mathbf{LIN}(r(a), r(b)) = \alpha_1 r(a) + \alpha_2 r(b) \quad (3.7)$$

$$\mathbf{AFF}(r(a), r(b)) = \alpha_1 r(a) + \alpha_2 r(b) + \beta \quad (3.8)$$

$$\mathbf{MLP}(r(a), r(b)) = W_3 h_2 \quad (3.9)$$

Where

$$h_1 = \sigma(W_1[r(a); r(b)])$$

$$h_2 = \sigma(W_2 h_1),$$

W_1 is (300×2) , W_2 is (768×300) , and W_3 is (1×768) . We do not claim that this is the best MLP possible, but use it as a simple architecture to contrast with the linear models.

3.3.3 Data and Compositionality Judgments

Treebank

To collect a large set of phrases with syntactic structure annotations, we collected all unique subphrases (≥ 2 words) from WSJ and Brown sections of the Penn Treebank (v3) [180].⁴

³Initially, we considered the elementwise product $\mathbf{PROD}(r(a), r(b)) = r(a) \odot r(b)$, but found that it was an extremely poor approximation.

⁴We converted the trees to Chomsky Normal Form with right-branching using NLTK [174]. We note that not all subtrees are syntactically meaningful. However, we used this conversion to standardize the number of children and formatting. We exclude phrases with a null value for the left or right branch [25].

The final dataset consists of **823K** phrases after excluding null values and duplicates. We collected the length of the left child in words, the length of the right child in words, and the tree’s production rule, which we refer to as *tree type*. There were 50260 tree types in total, but many of these are unique. Examples and phrase length distribution can be found in § B.1, and § B.2.

English Idioms and Matched Phrase Set

Previous datasets center around notable bigrams, some of which are compositional and some of which are non-compositional [230, 231]. However, there is a positive correlation between bigram frequency and human compositionality scores in these datasets, which means that it is unclear whether models are capturing compositionality or merely frequency effects if they correlate well with the human scores.

Because models are likely more sensitive to surface features of language than humans, we gathered a more controlled set of phrases to compare with human judgments.

Since non-compositional phrases are somewhat rare, we began with a set of seed idioms and bigrams from previous studies [122, 230, 231]. We used idioms because they are a common source of non-compositional phrases. Duplicates after lemmatization were removed.

For each idiom, we used Google Syntactic NGrams to find three phrases with an identical part of speech and dependency structure to that idiom, and frequency that was as close as possible relative to others in Syntactic Ngrams [93].⁵ For example, the idiom “sail under false colors” was matched with “distribute among poor parishioners”. More examples can be found in Table B.3. An author of this paper inspected the idioms and removed those that were syntactically analyzed incorrectly or offensive.

3.3.4 Approximating a Composition Function

Methods

To approximate the composition functions of models, we extract the **CLS** and **AVG** representations from each model on the Treebank dataset. We used 10-fold cross-validation and trained the learned probes on the 90% training set in each fold. The remaining 10% were divided into a test set (5%) and dev set (5%).⁶

To fairly compare probes, we used minimum description length probing [277]. This approximates the length of the online code needed to transmit both the model and data, which is related to the area under the learning curve. Specifically, we recorded average cosine similarity of the predicted vector and actual vector on the test set while varying

⁵The part of speech/dependency pattern for each idiom was taken to be the most common pattern for that phrase in the dataset

⁶The learned probes were trained with early stopping on the dev set with a patience of 2 epochs, up to a maximum of 20 epochs. The Adam optimizer was used, with a batch size of 512 and learning rate of 0.512.

Idiom	Matched phrase	Syntactic pattern	Log frequency
Devil’s advocate	Baker’s town	JJ/dep/2 NN/pobj/0	2.398
Act of darkness	Abandonment of institution	NN/dobj/0 IN/prep/1 NN/pobj/2	4.304
School of hard knocks	Field of social studies	NN/pobj/0 IN/prep/1 JJ/amod/4 NNS/pobj/2	6.690

Table 3.7: Examples of idioms with their matched phrases, selected based on having the same syntactic pattern and most similar log frequency in the Syntactic Ngrams dataset. Examples depicted here have the same log frequency. Note that the frequency is based on the most common dependency and constituency pattern found in Syntactic NGrams. Humans were asked to rate each phrase for its compositionality.

the size of the training set from 0.005% to 100% of the original.⁷ We compare the AUC of each probe under these conditions to select the most parsimonious approximation for each model.

Results

We find that **affine probes** are best able to capture the composition of phrase embeddings from their left and right subphrases. A depiction of probe performance at approximating representations across models and representation types is in [Figure 3.5](#). However, we note that scores for most models are very high, due to the anisotropy phenomenon. This describes the tendency for most embeddings from pretrained language models to be clustered in a narrow cone, rather than distributed evenly in all directions [70, 157]. We note that it is true for both word and phrase embeddings.

Since we are comparing the probes to each other relative to the same anisotropic vectors, this is not necessarily a problem. However, in order to compare each probe’s performance compared to chance, we correct for anisotropy using a control task. This task is using the trained probe to predict a random phrase embedding from the set of treebank phrase embeddings for that model, and recording the distance between the compositional probe’s prediction and the random embedding. This allows us to calculate an error ratio $\frac{\text{dist}_{\text{probe}}}{\text{dist}_{\text{control}}}$, where $\text{dist}_{\text{probe}}$ represents the original average distance from the true representation, and $\text{dist}_{\text{control}}$ is the average distance on the control task. This quantifies how much the probe improves over a random baseline that takes anisotropy

⁷We look at milestones of 0.005%, 0.01%, 0.1%, 0.5%, 1%, 10% and 100% specifically. This was because initial experimentation showed that probes tended to converge at or before 10% of the training data. Models were trained separately (with the same seed and initialization) for each percentage of the training data, and trained until convergence for each data percentage condition.

into account, where a smaller value is better. These results can be found in § B.6. The results without anisotropy correction can be found in § B.8. In most cases, the affine probe still performs the best, so we continue to use it for consistency on all the model and representation types.

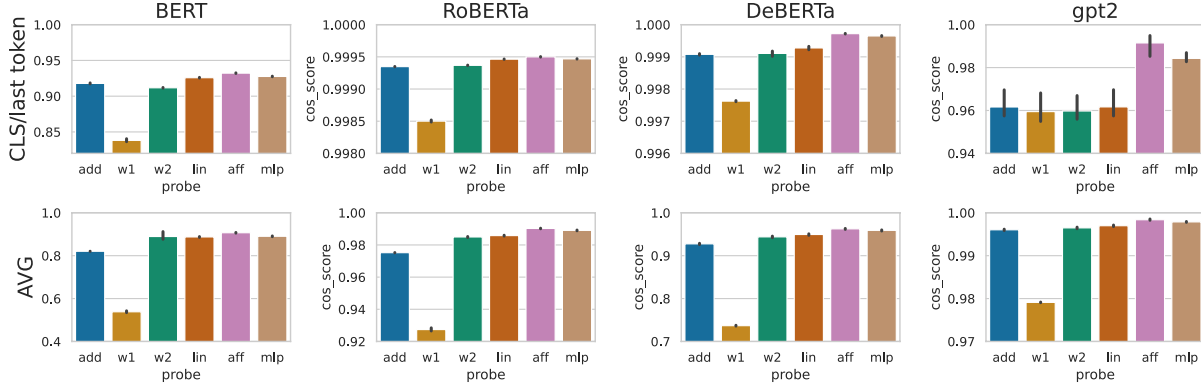


Figure 3.5: Mean compositionality score (cosine similarity) and standard deviation of each approximative probe across 10 folds. Error bar indicates 95% CI.

We also compare the AUC of training curves for each probe and find that the affine probe remains the best in most cases, except $\text{RoBERTa}_{\text{CLS}}$ and $\text{DeBERTa}_{\text{CLS}}$. Training curves are depicted in § B.3. AUC values are listed in § B.9.

Interestingly, there was a trend of the right child being weighted more heavily than the left child, and each model/representation type combination had its own characteristic ratio of the left child to the right child. For instance, in BERT, the weight on the left child was 12, whereas it was 20 for the right child.

For example, the approximation for the phrase "green eggs and ham" with BERT [CLS] embeddings would be: $r_{\text{CLS}}(\text{"green eggs and ham"}) = 12r_{\text{CLS}}(\text{"green eggs"}) + 20r_{\text{CLS}}(\text{"and ham"}) + \beta$.

3.3.5 Examining Compositionality across Phrase Types

Methods

Intuitively, we expect the phrases whose representations are close to their predicted representation to be more compositional. We call similarity to the expected representation, $\text{sim}(r(x), \hat{f}(r(a), r(b)))$, the *compositionality score* of a phrase.

We record the mean reconstruction error for each tree type and report the results. In addition to comparing tree types to each other, we also examine the treatment of named entities in § 3.3.5. We examine the relationship between length of a phrase in words and its compositionality score in § 3.3.5.

Results

There is a significant difference between the mean compositionality score of phrase types. Particularly, the **AVG** representation assigns a lower compositionality score to NP → NNP NNP phrases, which is expected since this phrase type often corresponds to named entities. By contrast, the **CLS** representation assigns a low compositionality score to NP → DT NN, which is unexpected given that such phrases are generally seen as compositional. The reconstruction error for the most common phrase types is shown in [Figure 3.8](#).

Because different phrase types may be treated differently by the model, we examine the relative compositionality of phrases within each phrase type. Examples of the most and least compositional phrases from several phrase types are shown in [Table 3.8](#) for RoBERTa_{CLS}. Patterns vary for model and representation types, but long phrases are generally represented more compositionally.

Phrase type	Most compositional	Least compositional
PP → IN NP	("of", "two perilous day spent among the planters of Attakapas, ...") ("of", "the cloth bandoleers that marked the upper part of his body ...")	("of", "September") ("like", "the Standard & Poor 's 500")
S → NP-SBJ VP	("him", "to suggest it's the difference between the 'breakup' value ...") ("it", "was doing a brisk business in computer power-surge protectors ...")	("other things", "being more equal") ("less", "is more")
NP → NNP NNP	("M.", "Bluthenzweig") ("Dr.", "Volgelstein")	("Edward", "Thompson") ("Alexander", "Hamilton")

Table 3.8: Phrases rated most and least compositional using RoBERTa_{CLS} representations, from several syntactic phrase types. "..." indicates that a phrase continues but is too long to display. Long phrases and abbreviated names tend to have a higher compositionality score.

Named Entities We used SpaCy to tag and examine named entities [114], as they are expected to be less compositional. We find that named entities indeed have a lower compositionality score in all cases except RoBERTa_{CLS}, indicating that they are correctly represented as less compositional. A representative example is shown in [Figure 3.6](#). Full results can be found in [§ B.11](#). We break down the compositionality scores of named entities by type and find surprising variation within categories of named entities. For numerical examples, this often depends on the unit used. For example, in RoBERTa_{AVG} representations, numbers with "million" and "billion" are grouped together as compositional, whereas numbers with quantifiers ("about", "more than", "some") are grouped

together as not compositional. The compositionality score distributions for types of named entities are presented in [Figure 3.7](#).

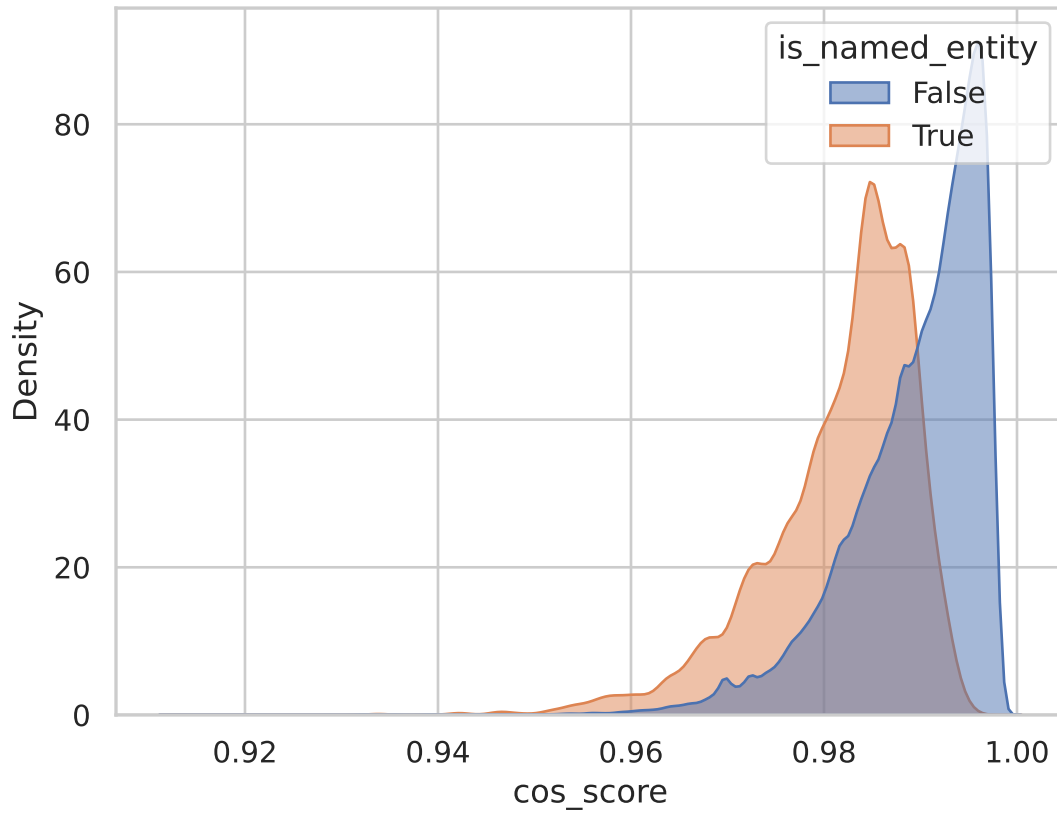


Figure 3.6: Density plot for compositionality scores of named entities and non-named-entities with RoBERTa_{AVG} representations. Higher means more compositional.

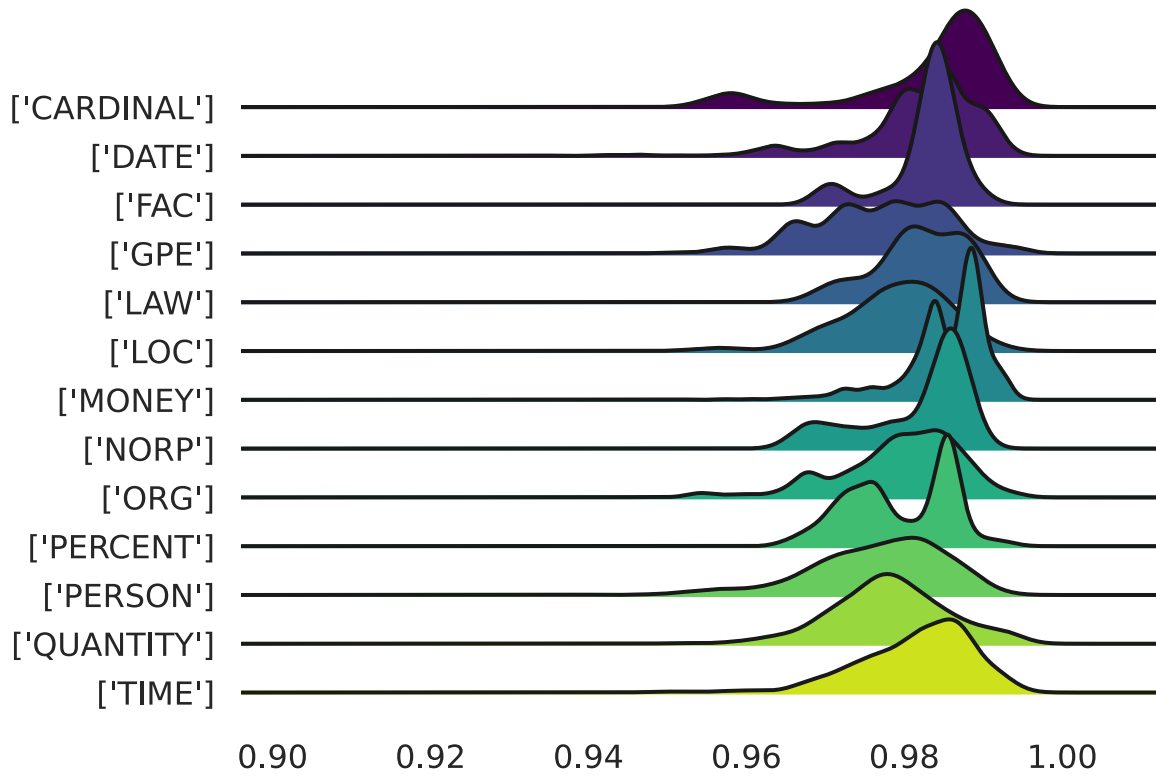


Figure 3.7: Density plots for compositionality scores of different named entity types with RoBERTa_{AVG} representations. Higher means more compositional.

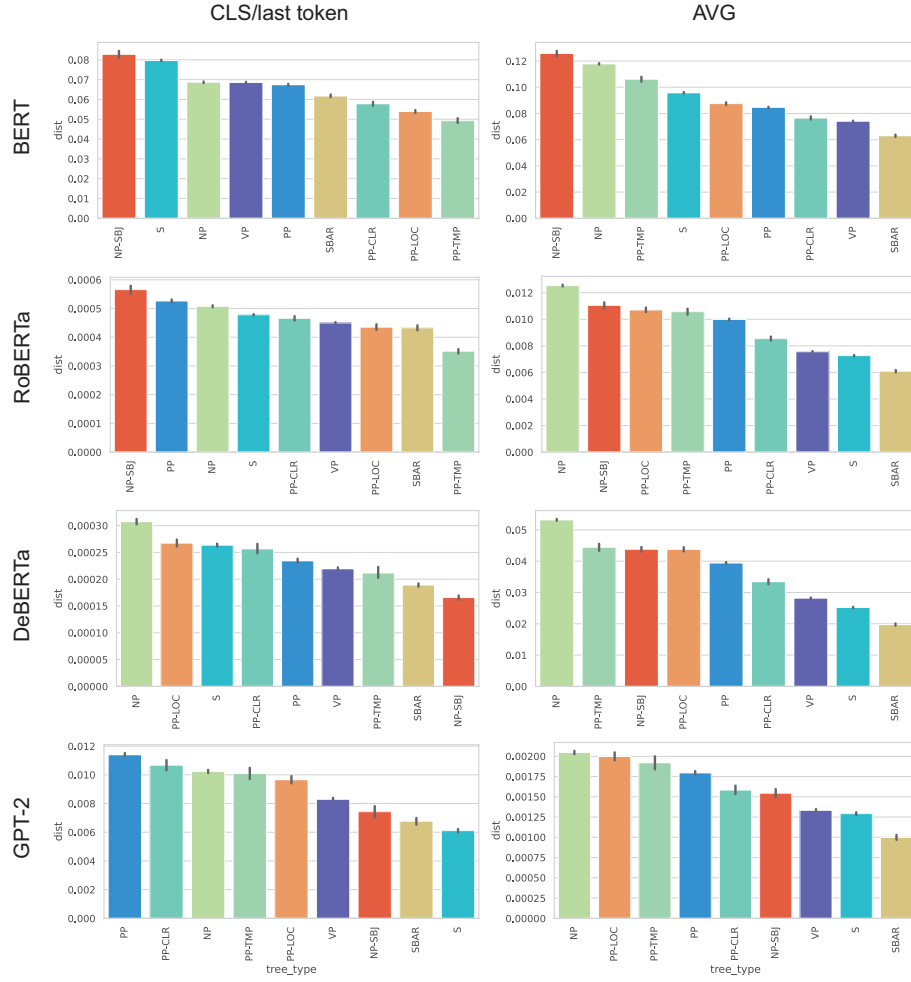


Figure 3.8: Tree reconstruction error (cosine distance) for each parent phrase type present in the Treebank, ordered from highest mean error to lowest. Based on the affine approximation for each model and representation type. Expanded version with all tree types is presented in § B.10.

Examining Compositionality and Phrase Length There is no consistent relationship between phrase length and compositionality score across models and representation types. However, **CLS** and **AVG** representations show divergent trends. There is a strong positive correlation between phrase length and compositionality score in the **AVG** representations, while no consistent trend exists for the **CLS** representations. This indicates that longer phrases are better approximated as an affine transformation of their subphrase representations. This trend is summarized in § B.4. All correlations are highly

significant.

3.3.6 Comparing Compositionality Judgments of Humans and Models

Methods

Human Annotation Human annotators assigned labels to each phrase in the matched dataset from § 3.3.3: 1 for not compositional, 2 for somewhat compositional, and 3 for fully compositional. They could also decline to answer if they felt that the phrase didn’t make sense on its own. Furthermore, they were asked how much each subphrase (left and right) contributed to the final meaning, from 1 for not at all, to 3 for a great deal. The Likert scale of 1-3 was chosen based on analysis of previous compositionality annotation tasks, which found that extreme values of compositionality were the most reliable [229].

Initially, six English-speaking graduate students were recruited. The six initial annotators all annotated the first 101 examples and the subset of three annotators with the highest agreement who agreed to continue (Krippendorff $\alpha = 0.5750$) were recruited for the full study, annotating 1001 examples. For the full study, the agreement was higher ($\alpha = 0.6633$). We took the mean of compositionality judgments to be the final score for phrases. The instructions shown to annotators are in § B.7. Examples judgments from an annotator can be found in Table 3.9.

Phrase	Idiom	Judgment	Subphrase contribution
Making heavy weather	Yes	1 - Not compositional	Making: 1 - Not at all Heavy weather: 2 - Somewhat
Chief part	No	2 - Somewhat compositional	Chief: 2 - Somewhat Part: 3 - A great deal
Portrait of Washington	No	3 - Fully compositional	Portrait: 3 - A great deal of Washington: 3 - A great deal

Table 3.9: Example judgments of one annotator on the pilot set. Annotators were asked to rate each phrase from 1 to 3, where 1 meant not compositional and 3 meant fully compositional. They were also asked how much each subphrase contributed to the meaning.

Model Comparison To compare human judgments to model compositionality scores, we use the best trained approximative probe for each model and representation type to predict a vector for the full phrase based on its left and right subphrases (taking the probe trained on the first fold). We use cosine similarity to the expected representation as the measure of how compositional a phrase is for a model and representation type.

We take the Spearman correlation between model compositionality scores and human compositionality judgments and observe differences between human judgments and compositionality scores from model representations.

Results

Correlation with human judgments There is a weak correlation between model and human compositionality scores. The most promising trend is found in RoBERTa, where both **CLS** and **AVG** representations have a significant positive correlation with human judgments. Results are in Table 3.10, with corrected p-values [113].

Model and representation	Spearman ρ	p-val	
BERT _{CLS}	-0.02308	0.9915	
RoBERTa _{CLS}	0.1913	9.7934 $\times 10^{-8}$ *	$\times$
DeBERTa _{CLS}	0.01466	0.9915	
GPT-2 _{last}	0.009428	0.02654*	
BERT _{AVG}	0.1283	8.594×10^{-4} *	
RoBERTa _{AVG}	0.1386	2.782×10^{-4} *	
DeBERTa _{AVG}	-0.03819	0.7792	
GPT-2 _{AVG}	-0.04598	0.6987	

Table 3.10: Spearman correlation between human judgments of compositionality and compositionality score generated by different model and representation combinations. P-values are corrected for multiple comparisons with the Holm-Bonferroni correction.

Subphrase Contribution Test Annotators indicated to what extent they believed each part of the phrase contributed to the final meaning. We examined examples in which annotators rated one part of the phrase, for example a , as contributing more to the final meaning, and checked how often $d_{cos}(r(x), r(a)) > d_{cos}(r(x), r(b))$. Models do surprisingly poorly at this test, with most performing below chance. Results are presented in Table 3.11. An error analysis on RoBERTa_{AVG} indicated that in many cases, errors were due to idiomaticity failures. For example, "noble gas" is a type of gas that was rated as being more similar to "gas" by humans, but "noble" by RoBERTa.⁸

Idiomaticity Test Because idioms were matched with non-idiomatic expressions, we tested for correctly identifying the idioms. We limited the analysis to pairs where the idiomatic expression was rated as less compositional than the matched expression. Results are shown in Table 3.11. Results are better than the subphrase contribution test, but models do not achieve good results, the best performing representation being RoBERTa_{CLS}.

⁸Similar errors were made for phrases such as "grandfather clock", "as right as rain", "ballpark estimate". A "grandfather clock" is a type of clock, "as right as rain" indicates that something is alright, and a "ballpark estimate" is a rough estimate.

Model and representation	Subphrase accuracy
BERT _{CLS}	49.71%
RoBERTa _{CLS}	45.91%
DeBERTa _{CLS}	45.61%
GPT-2 _{last}	43.86%
BERT _{AVG}	52.92%
RoBERTa _{AVG}	45.03%
DeBERTa _{AVG}	46.20%
GPT-2 _{AVG}	45.32%
Idiomatic accuracy	
BERT _{CLS}	45.60%
RoBERTa _{CLS}	60.03%
DeBERTa _{CLS}	56.67%
GPT-2 _{last}	59.15%
BERT _{AVG}	57.57%
RoBERTa _{AVG}	58.98%
DeBERTa _{AVG}	45.77%
GPT-2 _{AVG}	48.42%

Table 3.11: Accuracy of model representations on the subphrase test and idiomaticity test.

Correlations with Other Factors We examine correlations of model and human compositionality scores with the frequency and length of the phrase in words. As noted before, there is a strong correlation between length and compositionality score in models but not in human results. Results are in § B.12. A comparison of phrases rated as most and least compositional by humans, as well as RoBERTa, is presented in Table 3.12.

Model & representation	Most compositional	Least compositional
Human	"population growth" "few weeks away" "railroad monopoly"	"gravy train" "shrinking violet" "revolving door syndrome"
RoBERTa _{CLS}	"two small sticks" "dark glass bottle" "annual music festival"	"worse than none" "cases apart" "arch'd eyebrow"
RoBERTa _{AVG}	"look with open eyes" "be of equal importance" "come after breakfast"	"advertisement revenue" "taking it upon oneself" "all paces"

Table 3.12: Most and least compositional phrases in CHIP by human judgments and RoBERTa compositionality scores. Human scores are the average of 3 annotators.

3.4 Abductive Reasoning

3.4.1 Introduction

Suppose a friend teaches you to cook by heating olive oil and adding garlic. When you later cook without olive oil, you hypothesize that oil prevents burning and substitute butter instead. This exemplifies **abductive reasoning**—inferring the rule that best explains your experience [78, 209]. Alternatively, your friend could have given you explicit instructions (**deductive reasoning**) or you could have relied on remembered examples without explicit reasoning (**inductive reasoning**).

These reasoning types have close analogues in language model steering: **instruction following** (conditioning on explicit commands), **few-shot prompting** (learning from examples), and **instruction inference** (generating textual commands from examples before prediction). However, it remains unclear when to prefer each approach and how these capabilities relate in current models.

We examine these questions through numerical function learning and translation tasks, asking: (1) When does instruction inference improve over few-shot prompting?

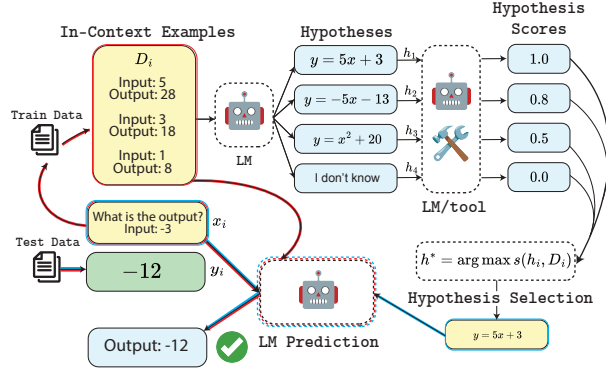


Figure 3.9: Diagram of abductive reasoning for an LM. Red arrows show data flow in inductive reasoning (few-shot prompting), while blue arrows show data flow in deductive reasoning (instruction following). Black arrows indicate data flow unique to abductive reasoning (instruction induction). Instruction inference generally improves on few-shot prompting and zero-shot chain of thought. However, success at inductive reasoning and success at instruction inference are not related.

Reasoning type	Analogue in LMs	Citations
Deductive	Instruction following	Sanh et al. [239], Wei et al. [285]
Inductive	In-context learning	Brown et al. [35]
Abductive	Chain-of-thought (one hypothesis)	Kojima et al. [142], Wei et al. [290]
Abductive	Externally aided instruction inference (many hypotheses)	Qiu et al. [223], Wang et al. [280], Yang et al. [302], Zhu et al. [316]
Abductive	LM-guided instruction inference (many hypotheses)	This paper

Table 3.13: Summary of reasoning types and analogues in language models. Citations to instruction following, in-context learning, and chain-of-thought are limited to the original paper due to the high number of papers on these topics, while we have tried to list all currently available papers on instruction inference (for abductive reasoning) with LMs.

(2) How do deductive, inductive, and abductive reasoning abilities relate? We find that instruction inference helps in simple cases but struggles with complex tasks due to incorrect hypotheses. Critically, abductive reasoning ability (proposing hypotheses) is generally unrelated to inductive reasoning (few-shot learning), while deductive reasoning shows strong performance when correct hypotheses are provided. These results highlight the diverse reasoning mechanisms triggered by different prompts in current language models.

3.4.2 Three Types of Reasoning in Language Models

We first give a formal definition of instruction following, in-context learning, and instruction inference, relating these processes to deductive reasoning, inductive reasoning, and abductive reasoning respectively. (Other ways of interacting with LMs may also

evoke one or more of these forms of reasoning, but we focus on important and widely used prompting strategies.) Figure 3.9 gives a schematic of the reasoning loop, while Table 3.13 gives examples of work falling in each category.⁹ Throughout this section, we assume that we have an input x , and we want to produce an output y with an autoregressive LM conditioned on some additional piece of data D that specifies the target task: $p_{LM}(y | x, D)$. We use in-context learning of linear functions as a running example.

3.4.3 Instruction Following

In instruction following, we want to map each input x to a y according to a general instruction or prediction rule D that is specified in the input. This may be viewed as a kind of deductive reasoning, which begins with one or more premises, and applies logically valid rules to reach a conclusion [245]. For instance, the instruction D might name a general function, such as *Apply this function to the input: $y = 5x + 3$* , followed by a specific query: *Input: -3*. This is illustrated in Figure 3.9.

3.4.4 Few-Shot Prompting

Few-shot prompting, by contrast, specifies the target task implicitly through examples. For each input x , the task specification D consists of a set of k training examples $D = \{(x_1, y_1), \dots, (x_k, y_k)\}$. We then sample from $p_{LM}(y | x, D)$. Few-shot learning requires LMs to perform inductive reasoning [104]. Unlike deductive reasoning, there is no explicit premise stated, but the model must complete the task in a similar way to the examples. In Figure 3.9, this is pairs of inputs and outputs, i.e. *Input: 5, Output: 28*. These examples help specify the task as a numeric prediction task, as well as the identity of the specific target function.

3.4.5 Instruction Inference

Instruction inference connects instruction following and in-context learning: given examples $D = \{(x_i, y_i)\}_{i=1}^k$ and input x , we can instruct the model to generate a hypothesis h about the identity of the task, i.e. an instruction describing the task associated with D . We may sample one hypothesis and immediately condition on it (a form of chain-of-thought prompting), or sample several and select the most promising one.

Compared to chain-of-thought, fewer studies have explored multi-hypothesis instruction induction; those that do typically rely on an external validation model rather than using the LM itself to evaluate. Our approach (in §3.4.6) has the following high-level form: after sampling n hypotheses $h_1, \dots, h_n$ from $p_{LM}(h | t, x)$, we evaluate each hypothesis by assigning a score $\text{score}(h_i, t)$ based on the hypothesis and in-context examples.

⁹We note that there is some disagreement in the philosophy literature about the exact distinction about inductive and abductive reasoning. Here we use commonly-cited definitions, which happen to distinguish in-context learning from the more explicit process of verbalizing and testing hypotheses.

We choose the best one $h^* = \arg \max \text{score}(h_i, D)$, and feed it back into the context as an instruction.¹⁰

Abductive reasoning is often called “inference to the best explanation” [66, 165]. Suppose that the model generates the hypotheses shown in Figure 3.9. We may parse hypotheses and apply them back to the in-context examples, then ranking them by prediction error. Then the model simply has to follow the instruction as in Section 3.4.3.

3.4.6 Methods

We describe the implementation of instruction inference with multiple hypotheses in this section. As the other settings (few-shot, zero-shot chain-of-thought, etc) are already well understood, we do not explain them further, but include the exact prompts used to implement methods in Appendix C.2, C.6, and C.10.

Instruction inference [Abduction with many hypotheses] After generating n hypotheses $\mathcal{H} = \{h_1, \dots, h_n\}$, we explore methods for reranking them. For all experiments, $n = 5$. Generally, these reranking methods capture “fit” to training data. External validator reranking was used only in the functions domain (see Section 3.4.9). The other reranking methods used a language model. Given scores of each hypothesis, we choose:

$$h^* = \arg \max_{h_i \in \mathcal{H}} \text{score}(h_i, D_i) .$$

Settings with instruction inference are referred to collectively as `instruction_inference`. We detail the score functions below:

Verbalized confidence We directly prompt the model to estimate the probability of the hypothesis given D_i . $\text{score}_{\text{Verbal}}(h_i, D_i)$ is set to the model’s probability estimation.

P(data) We use a separate LM,¹¹ to generate log probabilities for in-context examples given the hypothesis. $\text{score}_{\text{P(data)}}(h_i, D_i)$ is the sum of log-probabilities of tokens of D_i .

P(answer) This is similar to P(data) and uses the same template, but $\text{score}_{\text{P(answer)}}(h_i, D_i)$ only sums log-probabilities of tokens in answers y_i in the in-context examples.

External validator For structured hypotheses, it is possible to parse them and apply them back to in-context examples, with the score as the negative error. It may not be possible to parse all hypotheses, if they are in natural language or inconsistent formats.

3.4.7 Domains and Evaluation

We investigate LM behavior in three domains: linear function inference, an artificial language learning task, and vocabulary + typological feature learning in the Kalamang

¹⁰Several recent papers have recently proposed similar processes, but called this inductive reasoning. We believe that abductive reasoning may be a more apt term in any process that includes hypothesis evaluation and selection. See ?? for more discussion.

¹¹`text-davinci-002` was used as the logprobs-generating LM for all models.

Domain	In-context examples (D_i)	Query exam- ple (x_i)	Expected answer (y_i)	Hypothesis example (h_i)
Functions	(-10, -213), (9, 167), (4, 67), ...	15	287	$f(x) = 20x - 13$
Colours	(lug dax, blue green), (lug zup, blue yel- low), (lug bluf, blue blue), ...	lug walm dax bluf	blue blue blue green green	lug $\rightarrow$ blue
Kalamang vo- cab (ek)	(That is their place., Tompat ma me muin), (I'm getting pandandus, I want to make a mat.), ...	Sakina is pouching guavas.	Sakina sarimara lawat.	guava $\rightarrow$ sarim
Kalamang grammar (ek)	""	-	-	Order of subject and verb: SV

Table 3.14: Examples of in-context examples, queries, and hypotheses in each domain.

language. We refer to the underlying task we would like to improve (output prediction, translation) as the **base task**, and the task of inferring an explicit natural language hypothesis from few-shot examples as the **abductive task**. Examples of each task are in Table 3.14. We also evaluate hypotheses themselves in each domain. For the exact prompts used in each domain, refer to Appendix C.2, C.6 and C.10. Models used across all domains consisted of gpt family models (gpt-3.5-turbo; 35, gpt-4-turbo, 200) and llama family models (llama-2-7b-chat, llama-2-70b-chat, 270).

3.4.8 Linear Functions

Following investigations of language models' in-context learning abilities [6, 84], we construct a dataset of 40 linear functions $f(x) = ax + b$, where a and b are uniformly sampled from the integers $[-20, 20]$, along with 5 test examples for each function, yielding 200 test examples. For each test example, we randomly generate 5 in-context examples of the function $(x_i, y_i)_{i=1}^5$ and one test example. The test example and inputs for in-context examples are also uniformly sampled from the integers $[-20, 20]$. We refer to this as the **functions domain**. Prompts used are given in § C.2.

Hypotheses In this domain, hypotheses are proposed using all in-context examples for each test example. The model was presented with functions (and instructed to write them) in the form $y = ax + b$. For external validator reranking, we parse the generated hypothesis, and score it by its negative mean-squared error (MSE) when applied to the in-context examples.¹² If $h_i(x_{ik})$ represent executing the parsed hypothesis represented by h_i on example x_{ik} :

$$s_{\text{ground_truth}}(h_i, D_i) = \sum_{j=1}^k (h_i(x_{ij}) - f(x_{ij}))^2$$

Evaluation To evaluate the base task, we use 0-1 accuracy, as well as median squared errors. To evaluate hypotheses, we examine the accuracy of model-proposed a and b

¹²If a model produced an unparseable hypothesis (for instance, *I don't know*, or a generic answer like $y = ax + b$), that hypothesis was assigned a score of $-\infty$.

coefficients, as well as the Spearman correlation between proposed coefficients and the ground truth.

3.4.9 Simple Artificial Languages

Inspired by compositional instruction-following datasets, we generate a simple dataset where the inputs are nonce words such as *lug*, and the outputs are colour terms such as *blue* [148, 151]. We call the expanded dataset the **colours domain**. The ruleset for this domain can be found in § C.5. We generate 200 test examples and 800 training examples.¹³ Prompts used are in § C.6.

Hypotheses Following the original miniscan, we create a fixed minimal set of 5 in-context examples that contains each nonce word at least once, and from which the meaning of each nonce word can be reasonably inferred. During instruction inference, we isolate one nonce word at a time from the sentence, and have the model try to induce the vocabulary mapping for that word from 5 retrieved in-context examples containing that word. The ground truth grammar was written by the author, and consisted of production rules for each nonce term. Prompts used in the abductive and base tasks can be found in § C.6.

Evaluation We also use 0-1 accuracy to evaluate the base task of nonce word translation, as well as corpus-level chrF [215]. To evaluate the quality of the hypotheses themselves, we extract the production rules proposed by models and compare 0-1 accuracy against the ground truth. As a lenient evaluation on the “repeat” terms (see § C.5 for nonce terms and meanings), we marked a hypothesis for “repeat” terms as correct if it contained the term *repeat* or the numerals 2 and 3 respectively.

3.4.10 Kalamang Translation

For low-resource translation, we use the Machine Translation with One Book (MTOB) dataset, an English–Kalamang dataset with a grammar book [267]. Kalamang is an extremely low-resource language with fewer than 200 speakers, and virtually no text on the web. The base task is to perform sentence-level translation in both directions, while the abductive task is to infer correct grammar features and vocabulary mappings. The dataset consisted of 100 test sentences (50 in each direction) and 400 train sentences. A ground-truth bilingual dictionary was provided [276]. A grammar book was included as well, but to check correctness of high-level grammar inferences, we compiled a high level grammar sketch instead from WALS and GramBank features [67, 251]. More details about the grammar sketch can be found in § C.11. Otherwise, we use the same experimental settings as the baseline, summarized in § C.9.

Hypotheses We split hypotheses into vocabulary and grammar. The ground truth instruction included retrieved examples from the wordlist and the grammar sketch. To induce the grammar sketch, we posed each grammar feature as a question, for instance:

¹³We generate this dataset instead of using the original miniscan to mitigate memorization of the original nonce words, as well as to gain more test examples, as the original dataset has only 10.

Domain	Deductive reasoning works on average	Abductive reasoning works on average	Hypothesis proposal works on average	Abductive reasoning related to inductive reasoning
Functions	✓	✓	✗	✗
Colours	✓	✓	✓	✗
Kalamang	✓	✗	✗	✗

Table 3.15: Summary of results. A checkmark indicates that the property held for all or almost all language models, a half-checkmark indicates a partial success for all or almost all language models, while an X-mark represents lack of success for most language models.

What is the order of subject and verb in Kalamang?, and sampled 5 sentence pairs at a time with the requisite parts of speech until the model proposed an answer to the question. If the model responded that it was unclear, this would repeat up to a maximum of 10 iterations. We only performed this process once, for GPT-3.5-turbo and GPT-4-turbo respectively. Each model used its self-induced grammar sketch at test time.¹⁴ For vocabulary induction, we followed a similar process as in the colours domain.¹⁵

Evaluation We use corpus-level chrF. To evaluate the grammar sketch and vocabulary induction, we respectively compare to the ground truth wordlist and grammar sketch.

3.4.11 Results

In this section, we first evaluate models’ concrete performance across different domains. We highlight the significant improvement instruction inference offers in some cases in synthetic tasks, yet despite this, improvements are not uniform across tasks, and not attested in challenging domains like Kalamang. Our analysis also highlights an inconsistent correlation between the quality of generated hypotheses and few-shot learning success, meaning that the ability to generate or follow instructions doesn’t reliably predict task mastery or vice versa. These results (Table 3.15) suggest that while structured instructions can boost performance in simpler scenarios, their impact is less predictable in complex settings. Furthermore, the relationship between instruction induction, instruction following, and in-context learning is complex, and each capability may rely on separate unknown aspects of model architecture, training procedures, or data.

¹⁴Llama-2 models were found to be unable to propose grammar features with our prompts, so we used the GPT-3.5-turbo induced grammar sketch for these models.

¹⁵For Kalamang, due to computational costs, we cache the first parseable hypothesis proposed for each word and reuse it on subsequent sentences containing that word.

3.4.12 When Does Instruction Inference Improve Over In-Context Learning?

How effective is using the true instruction? In the domains we study, the ground-truth instruction tends to yield accurate results. Figure 3.11 displays mean accuracy in the linear functions and colours domains over six trials, sampled at different temperatures (see § C.1) for details. Notably, in linear functions (3.11A), GPT-4-turbo’s accuracy increases to 96% from a baseline of 30%, with GPT-3.5 also notably improving. In the colours domain, (3.11B), we see the true instruction also helps all models except for Llama-2-7b. However, this trend does not extend to the Kalamang task, where most models struggled to leverage the provided wordlist and grammar sketch effectively, indicated by chrF scores in Figure 3.12.

How effective are models’ induced instructions? Self-generated instructions also improve on the baseline in many cases, with variations by domain. In linear functions, models’ hypothesis induction markedly surpasses the few-shot baseline, with both the verbalized confidence and log-probability based reranking methods yielding comparable improvements (see Figure 3.11A again). For the colours domain (3.11B), instruction inference benefits performance, though not as strongly as in linear functions. chrF scores follow a similar trend, and are depicted in § C.7. Interestingly, for gpt-4-turbo and Llama-2-7b, using the models’ self-proposed hypotheses benefits performance more than using the ground-truth grammar for the colours language, despite the fact that self-proposed are not always correct. Unlike the two synthetic domains, inducing grammar and vocabulary items for Kalamang does not improve translation metrics in most cases.

3.4.13 How Does the Ability to Induce Instructions Relate to In-Context Learning?

How accurate are model-generated hypotheses? In assessing accuracy of models’ self-generated hypotheses across different domains, our findings reveal significant variations in accuracy. In the linear functions domain, we plot hypotheses generated by two models in Figure 3.10, and list the Spearman ρ [257] as well as p-values of model-predicted coefficients in Table 3.16. GPT-3.5-turbo and GPT-4-turbo’s proposed coefficients are positively correlated with the real coefficients, and this is statistically significant. However, this level of accuracy is not observed with Llama-2 models, indicating a disparity in model capabilities. In the colours domain, most models, except GPT-4-turbo, tend to generate inaccurate hypotheses. The exact accuracy of proposed hypotheses for colours is shown in § C.8. Mappings of simple vocabulary items such as *lug* tend to more accurate, with GPT-4-turbo achieving an 87% mean accuracy for hypotheses about this word. On the other hand, the difficulty increases with the terms which involved repeats, with GPT-3.5-turbo only achieving a 15% mean accuracy on *bluf*, the “repeat twice” term.

However, when extending these analyses to Kalamang, all models’ performance in predicting grammar features is relatively poor, with GPT-3.5-turbo predicting 5/18 and 4/18 features correct respectively. § C.12 shows the correctness of each grammar feature.

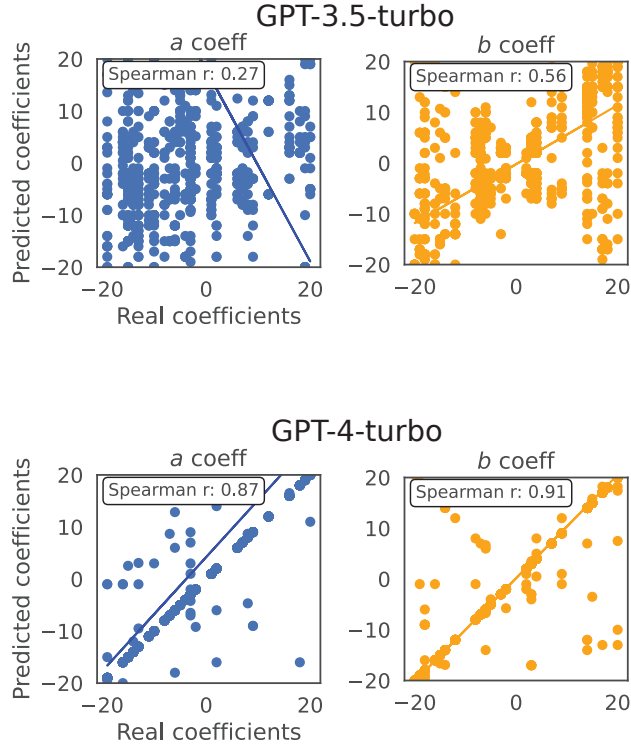


Figure 3.10: Real coefficients of linear functions and relationship to hypothesized coefficients for GPT-3.5-turbo and GPT-4-turbo. Remaining models can be found in § C.4. The x-axis has been truncated for visualization purposes (as there are some large outlier hypotheses). GPT-4-turbo is able to induce a reasonable function in-context, but other models struggle.

Model	b corr.	b p-val	a corr	a p-val
GPT-3.5-turbo	0.27	5.0×10^{-15}	0.56	2.7×10^{-67}
GPT-4-turbo	0.85	2.2×10^{-292}	0.91	0.0
Llama-2-7b	-0.0069	0.90	0.14	0.0081
Llama-2-70b	0.14	0.060	-0.018	0.82

Table 3.16: Spearman correlation and p-values of true vs predicted coefficients for each model in the functions domain.

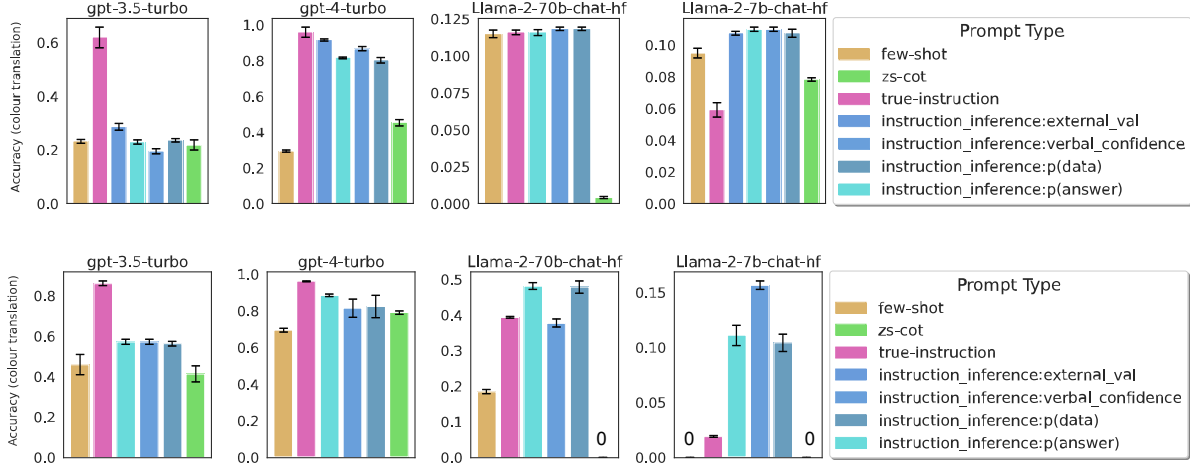


Figure 3.11: Accuracy of models in synthetic domains with and without hypothesis generation. Error bars indicate standard error. The top row shows results for the functions domain, while the bottom row shows results for the colours domain. Results are aggregated across 6 runs, and zero values are marked with '0'.

Vocabulary induction accuracy is also generally low, falling between 10-20% for most models in both the English to Kalamang as well as Kalamang to English directions. See § C.13 for details, as well as averaged segment-level chrF for the vocabulary hypotheses.

Is abductive reasoning related to in-context learning ability? In our final analysis, we examine whether the ability to induce correct instructions correlates with success in in-context learning. Specifically, we focus on the final hypotheses selected by the gpt model family, given that many hypotheses generated by llama models were incoherent or not formatted correctly. We use the point-biserial correlation [207] to assess the relationship between the accuracy of a model’s final hypothesis and its success in in-context learning.

In linear functions, we examine GPT-3.5-turbo because of GPT-4-turbo’s consistent accuracy in selecting correct hypotheses.¹⁶ The analysis reveals a chance-level agreement (0.0013), suggesting GPT-3.5-turbo may be able to predict outputs without identifying the underlying function well, or vice versa. This reveals a dissociation between prediction accuracy and instruction induction.

In the colours domain, we examine both gpt models and conduct a similar analysis for each nonce word. We divide the test examples into examples containing each word, and examine the point-biserial correlation between accuracy of induced word meanings and correct translations in the few-shot context. This correlation is generally low, and few p-values are significant after correcting for multiple comparisons with false discovery rate (FDR) [21]. See § C.8 for details.

In Kalamang, we repeat the process of computing point-biserial correlation between vocabulary induction correctness with segment-level chrF in the few-shot translations.

¹⁶This means agreement cannot be computed with the point-biserial correlation.

Unlike the other domains, there is a small positive correlation between correct vocabulary hypotheses and chrF in gpt models. See § C.13 for details. We note that chrF is a more fine-grained measure than accuracy, and the initial scores were low enough that copying some correct vocabulary items may have had a slight impact on otherwise completely incorrect translations.

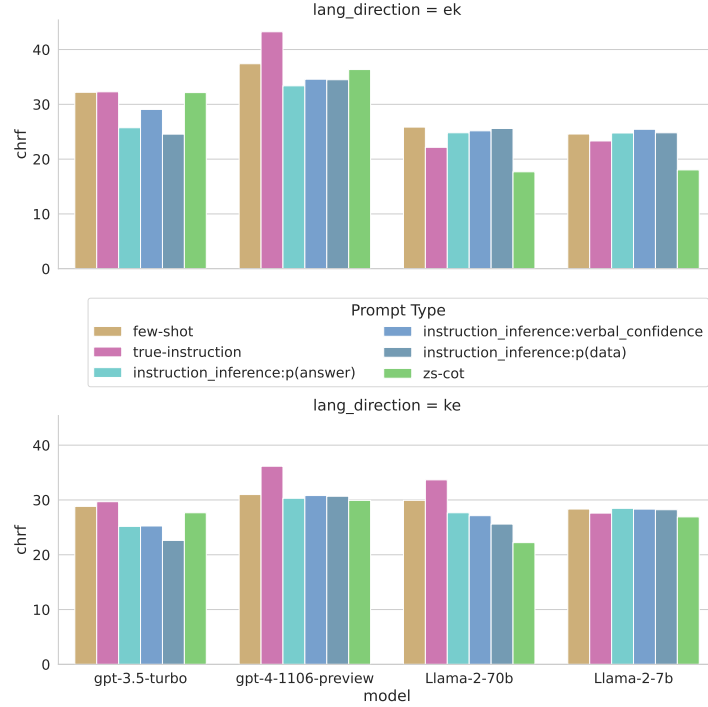


Figure 3.12: chrF scores for Kalamang under different methods, in English to Kalamang direction (top row) and Kalamang to English direction (bottom row)

3.4.14 Conclusion

We have examined the interplay between deductive, inductive, and abductive reasoning in LMs through the tasks of hypothesis proposal, in-context learning, and self-generated instruction following. Across three domains (linear function learning, artificial language translation, and Kalamang translation), we show that instruction inference is able to improve over few-shot prompting in simple synthetic domains, but that the relationship between these types of reasoning is complex, and they may not work together as expected presently when models are solving complex tasks. As abductive reasoning seems to be a relatively weaker capability in current language models as compared to instruction following, future work could develop more advanced mechanisms for natural-language hypothesis verification and correction. The use of hypothesis proposal during training remains under-explored, and joint training of models on question answering and hypothesis proposal with enforced consistency may help models display more consistent

behaviour. Enhancements in these areas could accelerate progress towards models capable of autonomous learning and self-improvement.

3.5 Conclusion

Overall, our investigation of core generalization capabilities from cognitive science in language models shows a consistent pattern: models do not necessarily fail catastrophically, but often diverge systematically from human cognition. These diagnostic insights motivate the central question of our next investigation: if we understand what capabilities are missing, can we also understand how they should develop during training?

CHAPTER 4

STUDYING THE DEVELOPMENT OF COMPLEX SKILLS IN TRAINING

4.1 Background

Having identified systematic gaps in trained models, we now investigate how models develop in the course of training. In §4.2, we examine the premise that scaling laws can predict downstream task performance, and find that across all examined tasks, adding information about design decisions apart from scale, such as data composition, improves performance prediction across all examined benchmarks. Importantly, this varies per task. For example, code may improve code-related tasks, while degrading natural language inference tasks. This suggests that strategic data composition could significantly improve downstream capabilities, but this is likely to be somewhat specific to the downstream task in question.

Building on this, we next examine the “midtraining” phase of training, a recent addition to many LM training pipelines. This practice introduces specialized data late in pretraining. We find that midtraining is most effective for domains more distant from general web text, such as math and code, and that effective midtraining effectively bridges distribution gaps between pretraining and posttraining, leading to better in-domain performance as well as less catastrophic forgetting after supervised fine-tuning.

This raises a deeper question: if both data composition and timing matter for capability development, can we understand capability acquisition at an even more granular level? In §4.4, we propose decomposing complex tasks into “elemental skills”, which are minimal tasks like copying, basic arithmetic, or pattern completion that serve as building blocks for more complex behaviors. By tracking when these atomic capabilities emerge during pretraining and how they combine, we can move from reactive data mixing to predictive capability engineering, potentially enabling synthetic data generation for precisely targeted skill development.

4.2 Linking Downstream Capabilities to Architecture and Data

4.2.1 Introduction

The effectiveness of language model training depends critically on pretraining decisions, yet current scaling laws based on parameter count and token number do not fully explain downstream performance [64, 133]. While controlled studies have examined individual factors—such as code data improving reasoning [13] or data age affecting QA performance [173]—these approaches are computationally expensive and typically vary only single aspects of training.

We instead ask: can we leverage the collective experience of open language models to understand how training decisions jointly impact downstream performance? We catalog architectural and data features from 92base pretrained models spanning 2019-2024, then train regressors to predict benchmark performance using both traditional scaling factors and design decisions.

Our analysis reveals that scaling alone does not determine performance—regressors using all features consistently outperform scaling-only baselines across 12 benchmarks. We reconfirm empirical findings like optimal code ratios in pretraining and discover that features extracted from model-generated text (such as question-word frequency) predict downstream performance, suggesting that generation patterns reflect pretraining biases that influence capabilities. This approach provides a practical resource for learning from the community’s collective training experience.

4.2.2 Scaling Laws

Definition

We define scaling laws here as a relationship between the number of parameters N and the number of tokens D of a language model family, and the expected language-modelling loss at convergence $L(N, D)$.¹ Importantly, these laws are typically examined while holding all other factors constant: keeping the same model architecture, training data, and model parameters. Originally, Kaplan et al. [133] showed that over a wide range of transformer-based models, this relationship can be expressed as a power law:

$$L(N, D) = \left(\left(\frac{N_c}{N} \right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_c}{D} \right)^{\alpha_D} \quad (4.1)$$

Later, Hoffmann et al. [111] introduced a similar law, which differed in the coefficients fitted, but was also based on a power law.

¹Please see §4.2.6 for more detailed discussion; scaling laws can and do take into account other factors in various works, but for simplicity we call N and D scale-related here, while all other decisions in §4.2.3 are contrasted with these.

However, scaling laws are not absolute, and the exact functional form and fitted coefficients may depend on the architecture type, size range [206], or other considerations such as inference costs. See (§4.2.6) for further discussion.

Maybe it’s Not Just Scaling?

Are parameter count and number of training tokens really all that are needed to accurately predict a model’s downstream performance? Intuitively the answer is “no” – there are a number of design decisions that go into model training, and all of them could have an effect on model performance.

Model Architecture Details While the majority of modern language models follow the transformer architecture, there are some details that differ. For instance, the variety [310] and position [296] of layer normalization, and the type of positional encoding [220, 261] make significant differences in model performance. Previous work, such as Gu and Dao [97], has demonstrated empirically that holding all other factors equal, models that make better architecture decisions [270] outperform those that make worse decisions [275].

Data Composition In addition, data composition and quality plays a major role in the final quality of a model. For instance, past work has demonstrated that training on some quantity of code improves performance on English reasoning tasks [178]. Also, work has demonstrated that filtering for “educational” content allows for more efficient learning and higher performance on knowledge-based question answering tasks [98].

Task Setting Finally, there is an interplay of all the above factors with how model performance is measured. While previous work on scaling laws has mostly measured loss values, downstream users usually care about task performance, rather than validation loss on a pretraining dataset. Although there is often a correlation between the two for many tasks, certain tasks may be harder to predict from a model’s loss alone [23]. Moreover, certain tasks exhibit pathological scaling behaviour, such as inverse or U-shaped scaling [36, 182, 289], or simply more unpredictable performance [120].

We ask: *can we more effectively predict the performance of LLMs by devising a new set of “laws” that are not just reliant on scaling-based factors?*

4.2.3 Building a Database of Publicly-Available Language Models

To approach our research question, we built a comprehensive database of publicly available language models. Our database encompasses models spanning a wide range of sizes, from 11M to 110B parameters,² and includes only distinct pretrained base models

²Including embedding parameters.

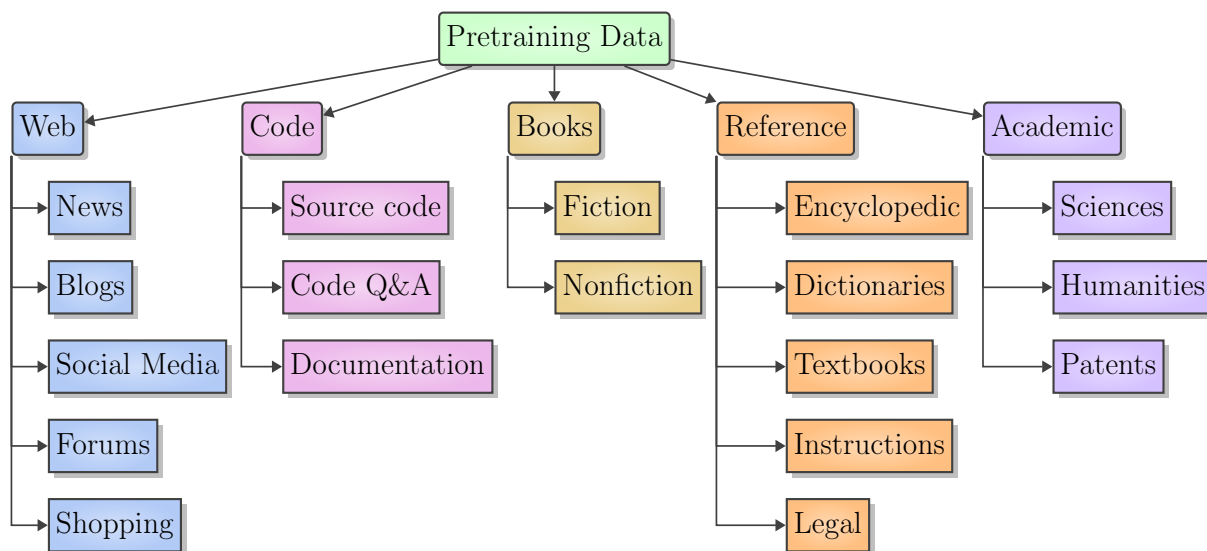


Figure 4.1: Taxonomy of pretraining data categories. We sorted data sources into this taxonomy based on model documentation.

with decoder-only architectures.³ In this section, we describe the criteria used for model inclusion, how we featurized the models, as well as the evaluation suite we used.

Data Collection

To ensure that our analysis was consistent, we applied the following criteria:

Pretrained-only: Only base models that were pretrained from scratch were included. Fine-tuned variants, merged models, and models with additional post-training were excluded.

Architecture: Only transformer-based decoder-only models were included to maintain uniformity. Mixture-of-experts (MoEs) or other architectures were excluded.

Publicly available information: Only models with publicly available metadata, documented through configuration files or papers, were included. In particular, both the total number of parameters and total number of tokens trained on were required for inclusion. A full list of models and model families can be found in § D.1.

Characterizing Models and Data

We represent each model by the architectural choices it makes, as well as its choice of pretraining data. Formally, let $\mathcal{A}$ be the set of features related to model architecture, and $\mathcal{D}$ be the set of features related to the model’s pretraining dataset. For each task T we want to approximate a model M ’s true score s_T with a prediction $\widehat{s}_T$:

$$\widehat{s}_T(M) = f_\theta([\mathcal{A}_M; \mathcal{D}_M]). \quad (4.2)$$

This reduces to typical scaling laws when $\mathcal{A} = \{\# \text{ params}\}$, $\mathcal{D} = \{\# \text{ tokens}\}$, and f_θ is a power law.

In total, we document 92 open models along the dimensions of model features, high-level dataset features, and features derived from that model’s no-context generations. For the full set of features and definitions, please see § D.2.

Features from Model Documentation We first collect information about each model by reading source papers/blogs when available (see § D.1 for original citations), as well as data listed on the Hugging Face Hub [292].

Architectural Features: These features capture design decisions that determine model structure. For example, *total parameters* (including embedding parameters), the number of transformer layers, the embedding and feed-forward dimensions, and details such as the type of layer normalization or attention variant used.

Data Features: These features summarize pretraining data composition. Representative examples include *total tokens trained on* and the percentage breakdown of tokens sourced from various domains defined in Figure 4.1, as well as the proportion of English-language tokens. Our pretraining data domains were derived from common subdomains in open pretraining datasets [82, 254]. We use the top level domains (web, code, books, reference, academic) as this tends to be the granularity at which data composition is described in papers.

Exploring Data Composition via Generation Although many models do document some details of their data composition, relatively few release their full pretraining corpus. Further, even when data composition statistics are provided, they are often presented at a high level of granularity. As a result, many models in our study have missing values for data composition.

To address this, we further propose an approach of estimating the data a model was trained on by generating from the model with a context containing just the beginning-of-sequence (BOS) token for that model (or end-of-sequence if that model lacked a BOS

³By distinct, we mean that each combination of training data and model architecture decisions should be unique. Variants of the same model trained on a deduplicated dataset are counted as separate, but not variants trained on the same data but with different curricula/initializations.

token). We use temperature-based sampling with $T = 1$, as this should in principle recover $P_{\text{raw}}(x_t \mid x_{<t}) \approx P(x_t \mid x_{<t})$ in the limit of sampling infinitely from an LM that captures its distribution perfectly.

We call this **free-generation**. However, this has important caveats: we cannot actually sample infinitely, and a model does not reflect its pretraining data perfectly. Therefore, we use this as a “fingerprint” of the model’s training, but do not claim that this reflects a model’s pretraining dataset exactly. In practice, we sample between 5-10k free-generations from each model (5k for larger models). In addition to categorizing them (see § D.5) by the domains in Figure 4.1, we also extract lower-level features:

Low-level data features: We aggregate per-generation statistics that reflect data quality and linguistic structure. Examples include the words per sentence, the depth of the constituency tree for natural language text, and dependency length.

Evaluation Datasets and Metrics

Commonsense Reasoning / NLI		
ANLI [195]	~ 163k	Brier Score
HellaSwag [308]	~ 70k	Accuracy
Winogrande [238]	~ 44k	Accuracy
XNLI [53]	~ 2.5k	Brier Score
Math / Logic		
GSM8K [52]	8 000	Accuracy
LogiQA2 [281]	~ 8k	Brier Score
MathQA [241]	~ 37k	Brier Score
General Knowledge		
ARC Challenge [50]	~ 2.6k	Accuracy
Lambada [205]	~ 10k	Accuracy
MMLU [108]	~ 2.85k	Accuracy
Other		
TruthfulQA [163]	817	Accuracy
HumanEval [45]	164	Accuracy

Table 4.1: Overview of LM evaluation datasets with approximate sample counts, citations, and evaluation metrics. Datasets ANLI, XNLI, LogiQA2, and MathQA use Brier Score, while the others use Accuracy.

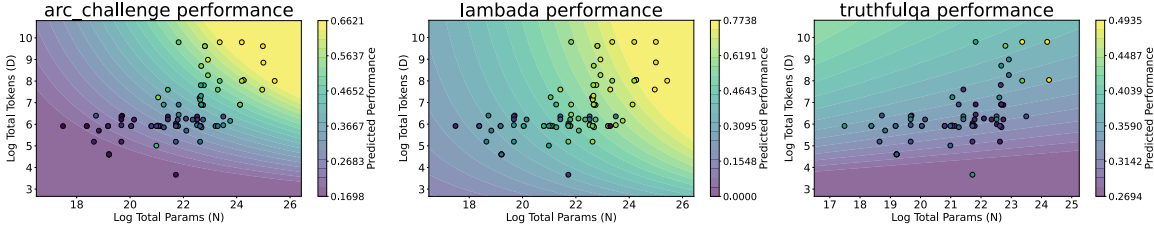


Figure 4.2: Performance of plotted against their total parameters and tokens. The background colour represents Equation 4.1 fitted to the task, and the marker colours indicate true performance. Some tasks have different performance trends with scale. Within each task, individual models may also perform unexpectedly.

To assess how design choices affect reasoning capabilities, we evaluated model performance on a curated suite of datasets derived from the first version of the Open LLM leaderboard [191], which was designed to capture diverse facets of reasoning (Table 4.1).⁴ We collect results for some models directly from the leaderboard, and for models not on the leaderboard we use the Eleuther LM eval harness [83] to conduct evaluations with exactly the same setting. In addition, if there were multiple versions of a task or sub-tasks, we evaluated all of them and averaged them to get the overall task score. For the full list of evaluation datasets and settings, see § D.3.

For an evaluation dataset T where the i -th sample is y_i and model M , we define $s_T(M)$ with:

Accuracy We use unnormalized, exact-match accuracy $s_{T,\text{acc}} = \frac{1}{|T|} \sum_{i=1}^{|T|} \mathbb{1}\{y_i = \hat{y}_i\}$ for the majority of tasks. We use pass@1 for Humaneval, but group it with accuracy tasks for convenience.

Brier score For tasks where smaller models struggle to achieve non-zero accuracy, we follow Schaeffer et al. [242] in using multiclass brier score as an alternate continuous metric for multiple-choice tasks [32]. For a task with K classes, let p_{ik} be the predicted probability for class k on sample i . Then $s_{T,BS} = \frac{1}{|T|} \sum_{i=1}^{|T|} \sum_{j=1}^K (p_{ik} - \mathbb{1}\{y_i = k\})^2$.⁵

Heterogeneity in Task-specific Scaling

Before adding in other factors, we examine differences in scaling along N and D between our selected tasks. We fit a Kaplan et al. [133] style law to each task. As seen in Figure 4.2, we see that different tasks may exhibit marked differences both in how well they follow scaling trends, as well as their individual scaling contours. For instance, TruthfulQA appears to exhibit U-shaped scaling, while Humaneval has more “outlier” models. A full list of R^2 values for tasks can be found in § D.4.

⁴Arithmetic and Minerva math ([33, 107]) tasks were also initially included in this leaderboard, but we excluded them as we focused solely on base (not instruction tuned) models, and very few were able to achieve non-zero scores.

⁵Note that lower is better for brier score. Multiclass brier score ranges between 0-2.

4.2.4 Predictive Modeling

Next, given our database we fit a regressor to try to predict performance. In traditional scaling laws, regressors are fit based on power laws. However, we are now dealing with a larger number of features, some of which may not be captured well by simple parametric forms. Hence, we follow previous work on performance prediction [294, 304] utilizing tree-based regressors based on XGBoost [46].⁶

For each evaluation benchmark, we train a model to predict the performance metric on that task based on architectural features $\mathcal{A}$ and data features $\mathcal{D}$. For each task setting, we perform 3-fold cross-validation due to the relatively small number of models, with a nested inner cross-validation over the training set in each fold. The inner cross-validation conducted grid search over a small set of hyperparameters, allowing the model to slightly vary per task. See § D.9 for more details.

Evaluation To evaluate the predictors, we use Mean Absolute Error averaged across all models and folds. In other words, for a task with N models evaluated, $\text{MAE}_T = \frac{1}{|T|} \sum_{i=1}^N |s_T(M_i) - \widehat{s}_T(M_i)|$. We compare the scaling-laws predictor as well as the all-features predictor against each other, but also against the **median baseline**, which simply predicts the median score of the models in the training set for each model in the test set of that fold.

Iterative Feature Selection As the full set of features is very large, we sequentially selected features from the full set greedily based on which reduced MAE the most, averaged across 5 random seeds. Features were added until no reduction of at least 1×10^{-4} was observed. We started using only the two scaling laws features, and refer to this as the **scaling-laws** model, though it does not have the form of a traditional power law.⁷ By then incorporating additional architectural or data features, we can then directly quantify the incremental predictive power afforded by these extra features. We refer to the model with the set of features as the **all-features** model. In all cases, we ran models with the same hyperparameter grid and the same random seeds and splits.

Significance Testing Because the relative difference between baselines is small, we test both predictors across many seeds (50). We then ran paired t-tests on the overall MAE values for each seed, and corrected for multiple comparisons across tasks with the False Discovery Rate [22].

⁶We also performed preliminary experiments with LightGBM [137] but it yielded very similar results in both prediction accuracy and feature importance.

⁷Typically, scaling laws are used to extrapolate the performance of larger models. Because we use a decision-tree based predictor, our approach is less likely to extrapolate, a trade-off we opted to take to incorporate the array of scale-independent features we have, not all of which are numeric. Therefore, we more so focus on interpolating performance within the size boundaries that we have (roughly 10M-100B parameters, and 50B-3T tokens). Examining results with a variety of other prediction methods is an interesting direction for future work.

Benchmark	Setting	Baseline MAE	Scaling Laws MAE	All Features MAE	p-val (corrected)
Accuracy					
Arc Challenge	25-shot	13.23%	4.36% $\pm$ 0.12%	3.67% $\pm$ 0.09%*	4.89×10^{-19}
GSM8k	5-shot	15.65%	6.04% $\pm$ 0.21%	5.10% $\pm$ 0.23%*	6.49×10^{-14}
Hellaswag	10-shot	12.26%	3.93% $\pm$ 0.13%	3.18% $\pm$ 0.09%*	6.66×10^{-20}
Humaneval	0-shot	11.79%	8.08% $\pm$ 0.22%	6.93% $\pm$ 0.22%*	1.46×10^{-12}
Lambada	0-shot	16.89%	9.51% $\pm$ 0.33%	6.85% $\pm$ 0.25%*	2.87×10^{-22}
MMLU (0-shot)	0-shot	11.98%	4.76% $\pm$ 0.20%	4.10% $\pm$ 0.17%*	6.02×10^{-13}
MMLU (5-shot)	5-shot	12.25%	3.97% $\pm$ 0.18%	3.54% $\pm$ 0.14%*	2.09×10^{-10}
TruthfulQA	0-shot	3.72%	2.75% $\pm$ 0.08%	2.29% $\pm$ 0.06%*	1.27×10^{-17}
Winogrande	5-shot	10.14%	3.39% $\pm$ 0.08%	3.09% $\pm$ 0.07%*	6.02×10^{-13}
Brier score					
XNLI	0-shot	7.22	6.68 $\pm$ 0.11%	6.30 $\pm$ 0.11%*	3.16×10^{-9}
ANLI	0-shot	5.90	6.74 $\pm$ 0.19%	6.53 $\pm$ 0.21%*	3.84×10^{-4}
MathQA	0-shot	7.57	2.83 $\pm$ 0.06%	2.75 $\pm$ 0.07%*	1.63×10^{-4}
LogiQA2	0-shot	12.74	4.74 $\pm$ 0.12%	4.60 $\pm$ 0.15%*	1.37×10^{-2}

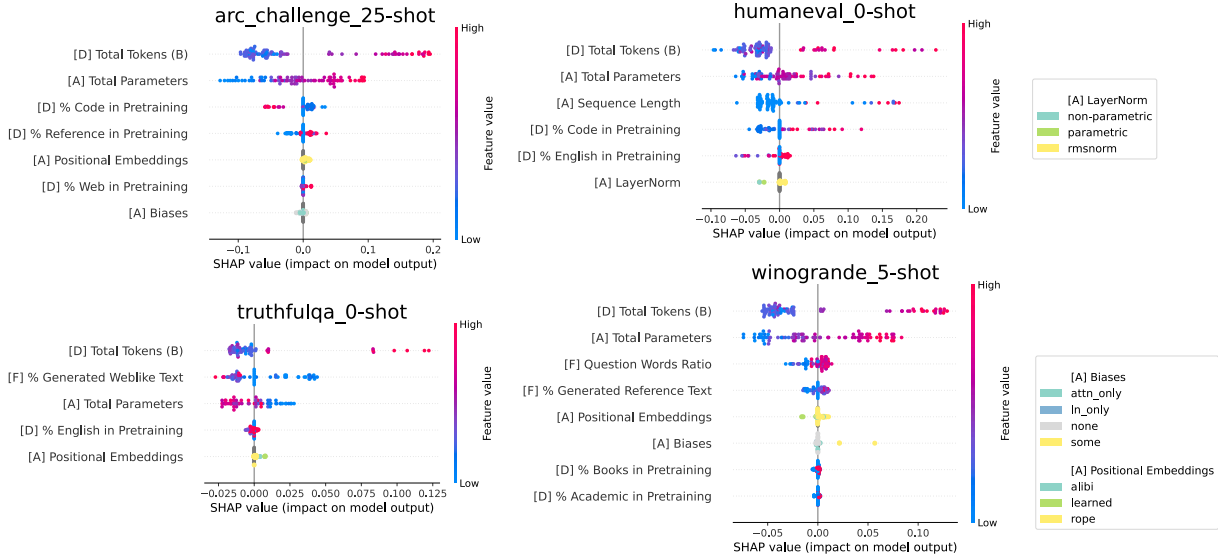
Table 4.2: Comparison of MAE values (mean $\pm$ 95% CI) for Scaling Laws and All Features predictors alongside Baseline MAE. Lower MAE is bolded; * indicates significance ($p < 0.05$). Brier score values are multiplied by 100 to be on a similar scale to accuracy.

4.2.5 Results

Predictor Performance

Incorporating scale-independent features consistently improves benchmark performance. We find that incorporating extra features alongside traditional scaling laws features leads to substantial improvements in prediction accuracy across multiple benchmarks, as seen in Table 4.2. The all-features predictor outperforms the scaling-laws-only predictor in all evaluated cases, with improvements ranging from approximately 3% (MathQA) to about 28% (Lambada) relative error reduction. Notably, the strongest improvements were observed in language modeling and common-sense reasoning tasks.

Certain tasks are more strongly dependent on non-scale features. This pattern of improvements suggests that architectural and training data features may be more informative for predicting performance on certain types of tasks more strongly linked to a particular “genre” of data. Large improvements were observed for both code generation (HumanEval, 15% improvement) as well as natural-language based reasoning tasks (e.g. Lambada, 28% improvement). Even tasks with narrower domains, such as mathematical reasoning (GSM8k, +16%) or knowledge-intensive evaluations (MMLU, +11–14%), see consistent, if more moderate, enhancements. The Brier score benchmarks, however, show smaller improvements (around 3–6%). This may be because the Brier score is inherently less sensitive to emergent effects in model performance, the specific choice of tasks limits the room for improvement, or a combination of both factors.



Feature Name	Description
[D] Total Tokens (B)	Total number of tokens used during pretraining, measured in billions (log scale).
[A] Total Parameters	Total number of parameters in the model (log scale).
[D] % Code in Pretraining	Percentage of pretraining data that consists of code.
[F] Question Words Ratio	Ratio of question-related words generated by the model.
[A] Dimension	Embedding dimension.
[A] Sequence length	Sequence length.
[A] LayerNorm	Type of layer normalization used (non-parametric, parametric, rmsnorm).
[A] Biases	Presence of bias parameters in the model.
[A] Positional Embeddings	Type of positional embedding used (alibi, learned, rope).
[F] % English Generated	Percentage of English text generated by the model.
[D] % Academic in Pretraining	Percentage of pretraining data from academic sources.
[D] % Reference in Pretraining	Percentage of pretraining data from reference sources.
[F] % Generated Weblike Text	Percentage of web-like text generated by the model.
[F] % Generated Reference Text	Percentage of reference-like text generated by the model.
[D] % Books in Pretraining	Percentage of pretraining data from books.
[D] % English in Pretraining	Percentage of English text in the pretraining data.

Figure 4.3: In all tasks, the number of parameters and pretraining tokens heavily influences the predictions made by the regressor. The percentage of code in pretraining often influences predictions negatively for NLI tasks but positively for Humaneval. [D], [A] and [F] denote features derived from data, architecture, or free-generations of a model respectively.

What Features Does Task Performance Depend On?

To better understand the factors that influence task performance, we examine Shapley [246] (SHAP) values of the predictor, which provide a local view of how individual feature values influence predictions. The results and feature descriptions for Arc Challenge, HumanEval, Winogrande, and TruthfulQA are shown in Figure 4.3, and results for remaining benchmarks are shown in § D.12.

A little code goes a long way, but too much is harmful to NLI. One of the most important non-scaling features is the percentage of code data in pretraining. Higher code

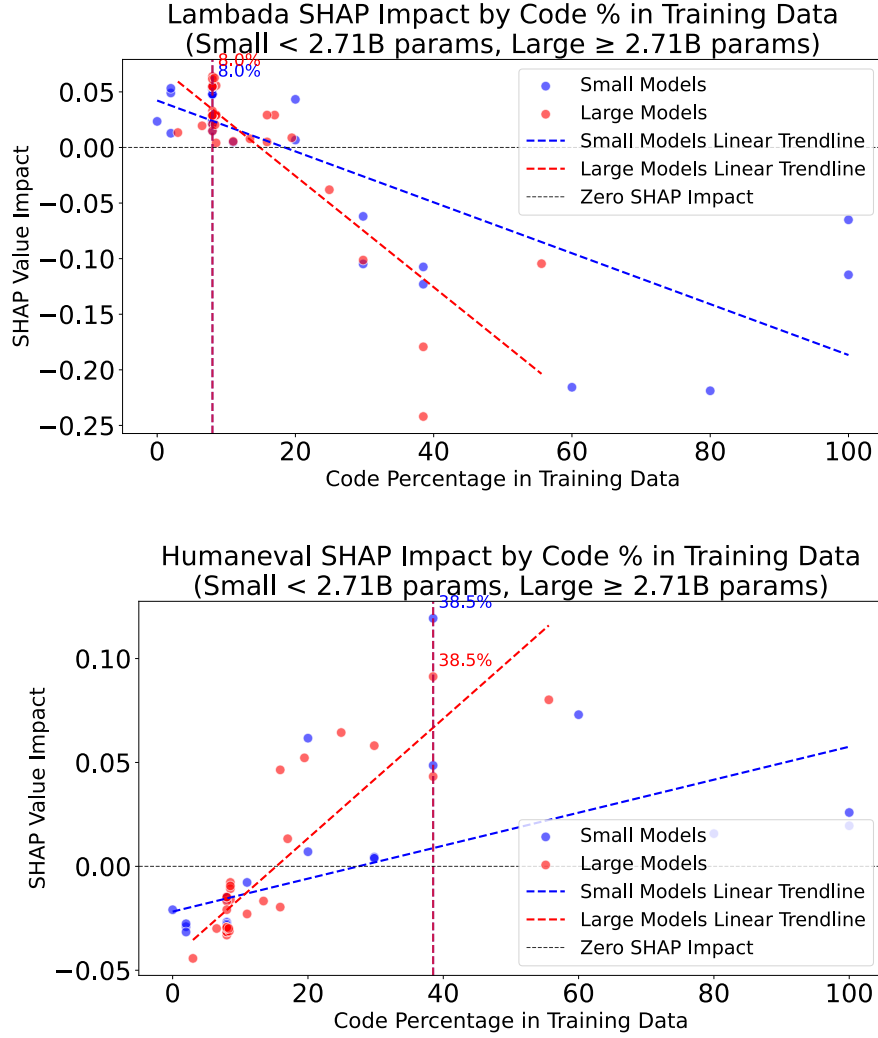


Figure 4.4: SHAP impact of code percentage on Lambada (representative NL task) and Humaneval on our regressors.

composition results in positive Shapley values (i.e. higher predicted performance) for Humaneval, but it negatively affects Arc Challenge, Hellaswag, Winogrande, and Lambada. In the scatterplots of Figure 4.4, we compare the code percentage against SHAP impact for both small and large models. We see that models trained with more than roughly 20–25% code are predicted to have large gains on tasks like Humaneval, but start to incur penalties on standard natural language benchmarks. By contrast, a moderate code proportion in the 15–25% range appears to balance these competing demands, yielding a more neutral or slightly positive effect overall.

Other domains of data can occasionally yield task-specific effects. While the percentage of code in pretraining is consistently selected as an impactful feature, with clear trade-offs, this other pretraining domains are selected less frequently. From the fine-grained features from free generations, we also observed that many recent models (particularly those trained on synthetic data such as the Phi [98] and SmolLM [9]) generate

a relatively large number of question words, indicating extensive training on data related to question answering. A higher percentage of reference-like or question-loaded generations resulted in better model accuracy on some tasks such as Arc Challenge and Winogrande. Additionally, models that generate more web-like data tend to do worse on TruthfulQA (Figure 4.3).

Non-scale architectural decisions have minor effects. Most highly influential features were data-related or architectural features related to scale (e.g., dimension). However, both the type of layer norm and the positional embedding were deemed to have a significant effect in some cases.

4.2.6 Related Work

Empirical Data Composition Results

Prior work has examined the effects of including code during pretraining [13, 178] and ablating domains such as C4 or books from The Pile [173]. Data filtering has also been shown to improve performance beyond scaling alone by removing low-quality data [95, 255]. Our results align with prior findings, indicating that code can enhance natural language reasoning at moderate proportions but degrades performance at higher percentages. We estimate an optimal code ratio of 15-25%, refining prior work suggesting 25% [13], though intermediate ranges were not tested in their study. Our approach—first pooling insights from existing models—complements empirical ablations by identifying useful axes of variation for further testing.

Observational and Task-Specific Scaling Law Fitting

Previous works have examined task-specific scaling laws. In machine translation, parameter allocation between encoder and decoder affects outcomes, and incorporating machine-translated data can be detrimental [90]. Multilingual studies reveal that language similarity doesn’t impact scaling trends; however, multitasking offers greater benefits when English is the target language [75]. Scaling laws for downstream tasks and transfer learning have been proposed, emphasizing that alignment between pretraining data and downstream tasks is crucial for performance prediction [109, 120]. Data repetition has been considered, especially in data-constrained domains [190], with extensions to multiple data domains [95]. Alternative scaling formulations address factors like sparsity [79], precision [147], and inference costs [112]. In contrast, some studies find stable performance across various batch sizes and learning rates [59].

Ruan et al. [234] also use observations from open-source models to predict task performance, but derive their predictions of one task’s performance from performance on other tasks. We find a similar result in identifying two axes of performance—general natural language ability and coding ability but are motivated instead by tracing these capabilities back to pretraining decisions.

Pretraining Data Selection

Domain mixing has been studied in pretraining, and other works have formulated this as a regression problem [168, 303] or used proxy models to select domain weights in the course of training [7, 124, 295, 305]. In contrast, we retrospectively analyze how domain composition and training decisions influence performance across tasks, which is a complementary perspective to optimizing data weights for a single model during training.

Tracing Capabilities to Data

Specific language model capabilities have been linked to patterns in pretraining data. Performance on numerical reasoning and syntactic rule learning depends on frequency of numerical terms in the training data [136, 286]. Ruis et al. [235] found that influential data for reasoning is dispersed across numerous documents and is associated with procedural content. Similarly, Chen et al. [47] observed that “parallel structures” are closely tied to in-context learning abilities. We currently focus on broader data domains, but our framework can be extended with more granular tasks or refined data features.

4.2.7 Conclusion and Future Work

We perform the first systematic analysis of the performance of open language models across diverse tasks and tie their performance to architectural and data-compositional design decisions. Looking into the future, there are a number of clear directions. First, our database (§4.2.3) can be further expanded as new models and benchmarks are released, and we will release the code and data to help spur community efforts for more systematic data documentation. Second, we hope our work will help discover hypotheses to be tested in more controlled settings – existing models intertwine a number of design decisions, and further controlled pre-training experiments that only involve one axis of variation could further clarify the effect of each feature. Finally, within our study, the great majority of pre-trained models focused on dense transformer architectures, while alternative architectures such as mixture-of-experts [60, 123] and state-space models [97] have also seen significant research interest. How to appropriately featurize these more various model architectures and use the information in performance prediction is an interesting challenge that may uncover further insights. Lastly, although pretraining data analysis and selection has mainly been focused on empirical findings so far, building a better understanding of how training impacts model capabilities through large-scale empirical studies could also facilitate interpretability experiments and possible interventions on learned representations, with controlled axes of variation providing case studies.

4.3 Midtraining Bridges Pretraining and Posttraining

4.3.1 Introduction

While scaling has driven most language model successes, a simple intervention has emerged across many models: breaking pretraining into two stages, with the second stage mixing in higher-quality specialized data from domains like mathematics and coding [5, 116, 198]. This “midtraining” approach shows better posttraining performance than single-stage pretraining [116], yet remains poorly understood scientifically. Not all specialized datasets work—instruction data alone yields no improvements, while mixtures including code and mathematical content provide substantial gains [198].

We conduct the first systematic investigation of midtraining through controlled experiments on models pretrained from scratch, asking: When does midtraining outperform continued pretraining? What data types are most effective? How does timing affect results?

Our investigation reveals three key findings: (1) Midtraining is highly effective for domains distant from general web text, particularly math and code; (2) It reduces catastrophic forgetting compared to both direct fine-tuning and continued pretraining; (3) Timing of specialized data introduction matters more than fine-grained mixture decisions. Together, these findings suggest midtraining effectively “bridges” the distribution gap between pretraining and specialized fine-tuning, providing smooth transitions at appropriate training stages.

4.3.2 Defining Midtraining

In this section, we define what we refer to as midtraining throughout this paper. Although this term has been used colloquially by model developers, it may be used slightly differently by different groups of people as it lacks a standard definition. However, as our paper investigates this phenomenon in detail, we introduce a formal definition in order to clearly delineate what counts as midtraining for the purpose of our experiments. We also focus on language models, though midtraining may also be applicable to other modalities.

Training Sequence Definitions

Modern language model training typically follows a sequential approach where models are trained on different data distributions in a specific order. Rather than attempting to define training phases by their data characteristics or objectives—which often overlap in practice—we adopt a sequence-based framework that defines phases by their order in the training process.

Let $\theta \in \Theta$ denote model parameters and $\mathcal{X}$ be inputs. A **training phase** is a tuple $(D, \mathcal{L}, n)$ where D is a data distribution over $\mathcal{X}$, $\mathcal{L} : \Theta \times \mathcal{X} \rightarrow \mathbb{R}$ is a loss function, and $n \in \mathbb{N}^+$ is the number of training steps taken in that phase. A **training sequence**

is an ordered collection $S = \{(D_i, \mathcal{L}_i, n_i)\}_{i=0}^N$ of training phases, where the parameters θ_i resulting from training up to phase i serve as the initialization for phase $i + 1$.

Pretraining is the initial training phase $(D_0, \mathcal{L}_0, n_0)$ in the sequence. In practice, pretraining often uses large, diverse sources of data, such as from the web. The objective for decoder-based language models, which we examine, is typically next-token prediction, $\mathcal{L}_0(\theta; x) = -\sum_{t=1}^{|x|} \log p_\theta(x_t | x_{<t})$. Pretraining typically consumes the majority of steps taken during training, so that $\forall i > 0, n_0 \gg n_i$.

Midtraining refers to intermediate training phases $(D_i, \mathcal{L}_i, n_i)$ where $0 < i < N$. Midtraining data is typically more specialized than general pretraining data, often including domain-specific content (code, math) and instruction-formatted data, while maintaining a mixture with general pretraining data. Typically, midtraining is a longer phase compared to finetuning, but shorter than the preceding pretraining phase, $n_0 > n_i > n_N$. There can potentially be multiple midtraining phases as well in the case of multi-stage pretraining curricula, but we focus on one-stage midtraining in this paper.

Finetuning is the final training phase $(D_N, \mathcal{L}_N, n_N)$ in the sequence. Finetuning usually uses task-specific datasets and may employ diverse objectives or preference-based methods rather than next-token prediction. Typically, finetuning uses the least number of steps, although this is not necessarily always the case.

Note that our framework reflects current popular practices for language model training, rather than prescribing an optimal training methodology. Additional edge cases include multi-stage finetuning, cyclic training schedules, and mixture schedules where distribution weights vary continuously.

Relationship to Curriculum Learning and Continued Pretraining

Midtraining relates to, but offers a different approach to domain adaptation compared to continued pretraining and relates to broader curriculum learning principles. While continued pretraining switches entirely to domain-specific data, midtraining maintains mixed distributions throughout the intermediate phase.

Curriculum learning The original definition of curriculum learning focused on gradually increasing the difficulty or diversity of training examples throughout the course of training [20, 69]. However, this term has evolved to generally mean any strategic ordering of training data [256]. In this general view, it encompasses any training strategy in which the sequence $S = \{(D_i, \mathcal{L}_i, n_i)\}_{i=0}^N$ is designed according to some principled ordering criterion—whether based on example difficulty, domain progression, task complexity, data quality, or other strategic considerations. Therefore, midtraining can be viewed as a type of curriculum learning that operates at a distributional level rather than over individual examples: instead of ordering specific training instances by difficulty or other criteria, midtraining strategically organizes different data distributions or domain mixtures across discrete training phases. This coarse-grained approach focuses on when to introduce entire types of data during pretraining rather than fine-grained sequencing.

Continued pretraining Continued pretraining typically refers to additional pretraining of an already-pretrained model on domain-specific data, where the data distribution shifts completely to the target domain [18, 101]. This approach commits entirely to domain-specific data after a certain point in training, potentially losing general capabilities. In contrast, midtraining maintains mixed distributions throughout the intermediate phase, preserving general pretraining data alongside specialized data. Formally, continued pretraining can be viewed as the limiting case where the midtraining distribution D_i consists entirely of domain-specific data (0% original pretraining data, 100% domain data), though our results suggest this pure approach may be suboptimal..

4.3.3 Experimental Setting

In order to systematically investigate the effect of midtraining, we design controlled experiments around four key research questions, which we introduce one by one along results:

RQ1: What downstream tasks benefit most from midtraining? We investigate which domains and tasks show the largest improvements from midtraining compared to standard pretraining followed by fine-tuning.

RQ2: What data is most effective for midtraining? We examine what characteristics make midtraining data effective, investigating the relationship between midtraining data and target tasks. We also compare midtraining with continued pretraining to understand the effects of mixed versus pure domain-specific distributions.

RQ3: When and how much midtraining data should be introduced? Current practice varies widely with respect to both timing and proportion of midtraining data. We systematically examine the effects of timing (when to introduce specialized data) and mixture weights (how much specialized data to include).

RQ4: How does midtraining change model representations? We analyze how midtraining affects model representations and activation patterns compared to non-midtrained models.

Training Setup

Pretraining We pretrain models from the Pythia family ranging in size from 70M-410M parameters on C4 web data [24, 228]. In all cases, we train for 128B tokens (approx. 61k steps) with a cosine learning rate schedule with a maximum learning rate of $3e-4$ and the AdamW optimizer [175]. We chose to fix the training budget at a point past Chinchilla-optimality for all models [111], in order to ensure that models have stabilized by the point at which midtraining data has been introduced, at least for later insertion points of midtraining data. We describe the exact training setup in Appendix E.1.

Midtraining We use five midtraining mixtures spanning popular domains: code (Starcoder), math, instructions (FLAN), general knowledge/QA, and high-quality web data (DCLM). Table 4.3 details each mixture’s composition and sources. All mixtures are introduced at varying start points (Starcoder: 6k steps, Math: 20k steps, others: 40k steps) based on data availability to prevent repetition. We compare against a control condition continuing C4 pretraining for the same number of tokens, keeping all other training details identical.

Starcoder (code) Our code mixture is a subset of the Starcoder pretraining dataset [159], which contains code in many languages. Note that this

Math The math mixture combines mathematical reasoning problems from the MAMmoTH ([306]) and OpenMathInstruct ([269]) datasets, featuring step-by-step explanations.

FLAN (instructions) Our instruction-formatted data comes from a processed version of the FLAN collection, which includes diverse task instructions and responses across natural language tasks [287].

KnowledgeQA (general knowledge and QA) The KnowledgeQA mixture is taken from Hu et al. [116]’s midtraining mix, and focuses on general knowledge and dialogue.

DCLM (high-quality web) Our high-quality web data is a subset of the DCLM pretraining dataset, representing web content with improved quality filtering compared to C4 [158].

Midtrain mix	Num. Samples (M)	Sources
Starcoder	96	[159]
Math	5.7	[269, 306]
FLAN	1.7	[287]
KnowledgeQA (-code)	4.7	[115]
DCLM	25	[158]

Table 4.3: Midtraining mixes used in our experiments and dataset(s) from which they were derived.

Downstream Evaluation We fine-tune models on the datasets GSM8k ([51]), SciQ ([291]), CodeSearchNet-Python ([118]), and LIMA ([315]) – chosen to span the domains covered by our midtraining mixtures. This allows us to test cases where the midtraining mixture is aligned or misaligned with the SFT dataset. We used standard language model supervised fine-tuning for all datasets. For information on the posttraining setup, see Appendix E.2.

Table 4.4: SFT and C4 validation losses for the 410M model across downstream datasets and midtraining mixtures, averaged across 5 seeds for each SFT dataset. Bold values indicate best performance within each dataset. Results demonstrate domain-specific benefits: Starcoder improves PyCode while Math improves GSM8K, with other combinations showing minimal benefits.

Downstream Dataset	Midtrain Mix	SFT Val Loss	C4 Val Loss
PyCode	C4	2.264	5.058
	Sarcoder (20%)	2.091	4.611
	Math (12%)	2.274	5.096
	FLAN (5%)	2.252	5.047
	KnowledgeQA (20%)	2.264	5.094
	DCLM (20%)	2.262	5.157
GSM8K	C4	0.956	4.937
	Sarcoder (20%)	0.944	4.951
	Math (12%)	0.918	5.038
	FLAN (5%)	0.958	4.970
	KnowledgeQA (20%)	0.961	4.966
	DCLM (20%)	0.949	4.966
LIMA	C4	3.490	3.166
	Sarcoder (20%)	3.422	3.155
	Math (12%)	3.507	3.168
	FLAN (5%)	3.513	3.164
	KnowledgeQA (20%)	3.501	3.163
	DCLM (20%)	3.496	3.158
SciQ	C4	3.271	6.873
	Sarcoder (20%)	3.261	6.831
	Math (12%)	3.257	6.894
	FLAN (5%)	3.277	6.771
	KnowledgeQA (20%)	3.273	6.814
	DCLM (20%)	3.270	6.874

4.3.4 Which downstream tasks benefit most from midtraining?

To identify which tasks benefit most from midtraining, we evaluate all midtraining mix and SFT dataset combinations. For each combination, we measure validation loss on the target domain (adaptation effectiveness) and C4 validation loss after fine-tuning (forgetting). We average results over 5 seeds after hyperparameter search for each checkpoint.

Results Our results in Table 4.4 for the 410M model’s final checkpoint demonstrate that **midtraining benefits are highly domain-specific**: code-focused midtraining (Sarcoder) substantially improves performance on the coding task (PyCode: 2.091 vs 2.264 validation loss, a 7.6% relative improvement), while math-focused midtraining shows the largest gains on mathematical reasoning (GSM8K: 0.918 vs 0.956, a 4.0% improvement). Crucially, mismatched domain pairs show minimal or negative effects—for instance, math midtraining provides no meaningful improvement on coding tasks. General instruction datasets (FLAN) provide little benefit across tasks, suggesting that domain-specific alignment is critical for midtraining effectiveness. Results for the other model sizes follow the same pattern and are shown in Appendix E.4.

4.3.5 What data is most effective for midtraining?

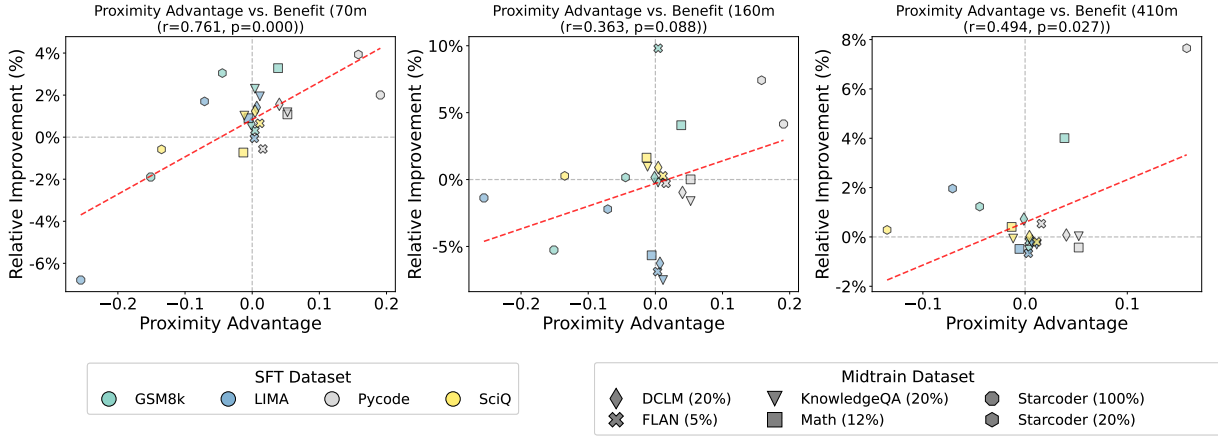


Figure 4.5: Relationship between proximity advantage and midtraining performance improvements for pairs of midtraining and SFT datasets. Each data point represents a (midtrain, SFT) pair, where the color indicates the SFT dataset and shape represents midtrain dataset. Proximity advantage ($\text{dist}(\text{C4}, \text{SFT}) - \text{dist}(\text{midtrain}, \text{SFT})$) indicates how much closer midtraining data brings the model to the target SFT dataset compared to the base pretraining data. Relative improvement is measured against the base model pretrained on C4. Strong positive correlations, particularly for smaller models (70m: $r = 0.761$, $p < 0.001$), support the hypothesis that effective midtraining data serves as a distributional bridge between pretraining and target domains.

Having established that midtraining effects are domain-specific, we now ask: *what determines the strength of these domain-specific effects?* Across midtraining-target pairs, we see improvements ranging from negligible (e.g. FLAN $\rightarrow$ coding) to strong (e.g. Starcoder $\rightarrow$ coding). Although the pairs that work the best seem obvious in hindsight, we ask whether we can quantify this intuition through measurable proximity between data distributions.

To test whether optimal midtraining data “bridges” the gap between pretraining and target distributions, we simply use token-distributional similarity between datasets, which we discuss in more detail in Appendix E.3. Figure 4.7 shows a reduced subset as an example: we can see that C4 and PyCode are fairly dissimilar (0.54 similarity), but the blend containing 29% Starcoder data brings this closer to Pycode (0.7 similarity).

We also compare the effect of midtraining mixes with continuous pretraining conditions in which the mixture weight switches to 100% specialized data. For code, we compare the default Starcoder midtraining mix (20% mixture weight, starting from 6k steps) with 100% Starcoder data starting from 40k steps. For math, we compare the math midtraining mix with 100% math data starting from 56k steps.⁸

⁸The different starting points are due to data availability, in order to ensure the midtraining mix does not repeat.

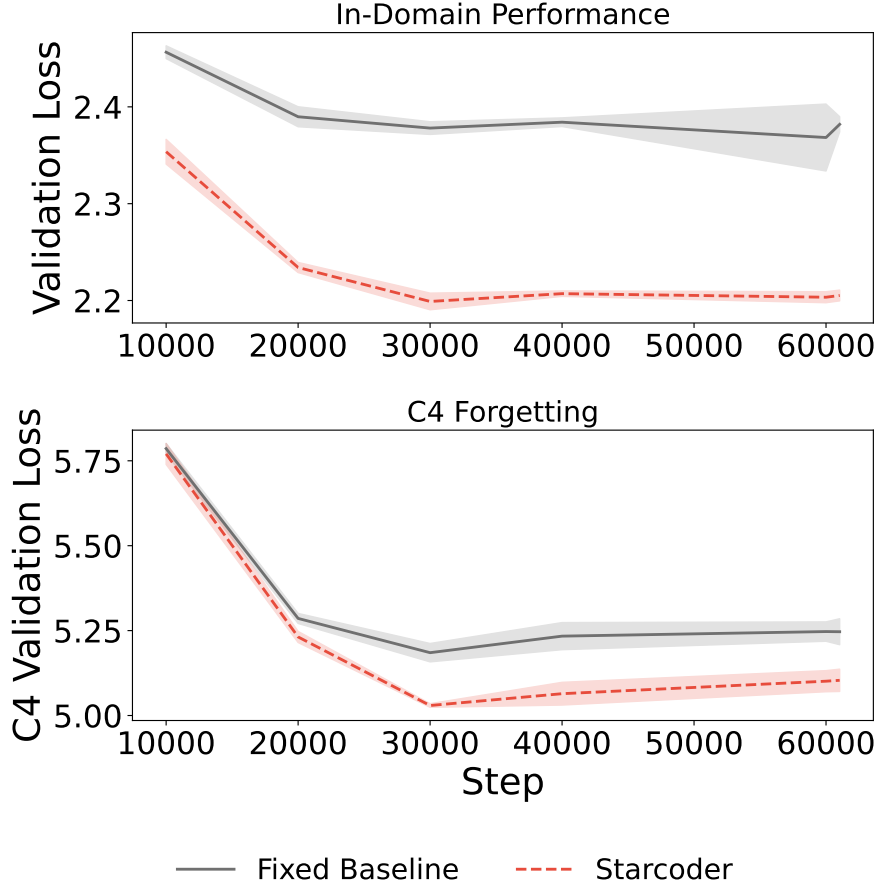


Figure 4.6: Validation loss and C4 loss for the starcoder-midtrained model and base pre-trained model *after* supervised fine-tuning on the Pycode dataset, with each point on the x-axis representing the number of steps the pretrained checkpoint was trained on.

Proximity and Bridging Effects

Results To understand why certain midtraining mixtures are effective, we test the hypothesis that **optimal midtraining data “bridges” the gap between pretraining (C4) and target SFT datasets**. Using the similarity measures developed in § 4.3.5, we measure the “proximity advantage” of each midtraining mixture—how much closer it brings the model to the target distribution compared to continuing with C4 alone. Results are shown in Figure 4.5.

We find a clear relationship between proximity advantage and downstream performance improvements across model sizes. The correlations are particularly strong for smaller models ($r = 0.761$, $p < 0.001$ for 70m), suggesting that effective midtraining data serves as a distributional stepping stone from general pretraining to specialized target domains. This bridging effect appears to be most beneficial when the gap between pretraining and target distributions is large, consistent with our hypothesis that mid-training helps models adapt gradually rather than requiring abrupt distributional shifts

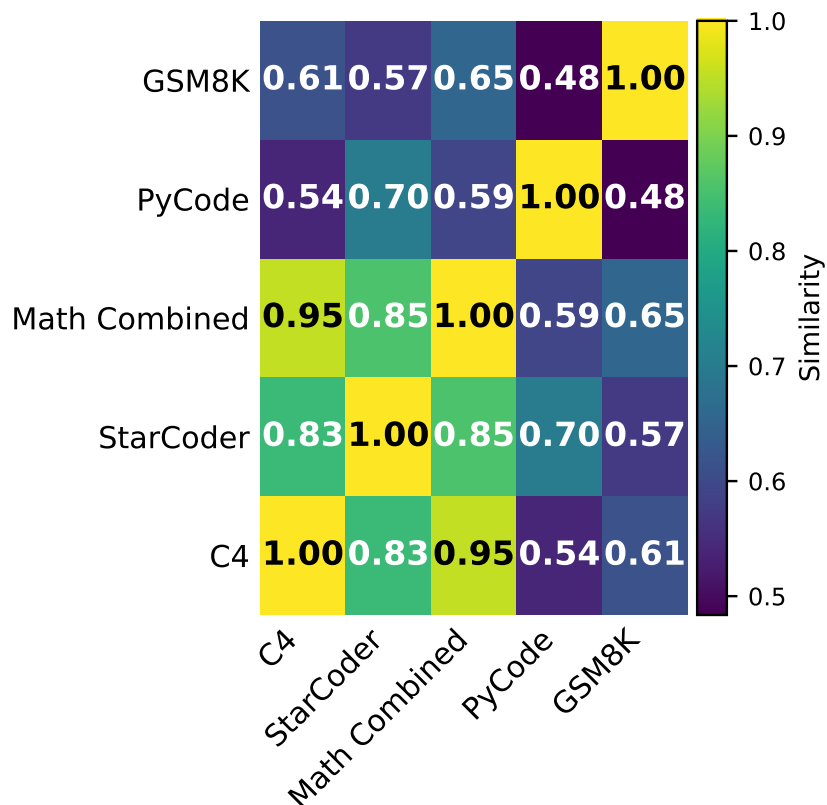


Figure 4.7: Example similarity matrix between pre/midtrain and posttraining datasets. For the complete matrix, see Appendix E.3.

during fine-tuning.

Midtraining vs. Continued Pretraining

Results Results in Table 4.5 show that midtraining consistently outperforms continuous pretraining across both domains and model sizes for both in-domain performance and C4 retention after fine-tuning. As this pattern holds for both code and math domains, this suggests that **maintaining some general pretraining data is useful during domain adaptation**, even for models specialized for a specific domain.

4.3.6 When and how much midtraining data should be introduced?

Given that a midtraining dataset’s proximity to the finetuning dataset matters, and that we want to balance midtraining data with general pretraining data, we next investigate when to integrate midtraining data and what mixture weight to use. Using the starcoder midtraining mixture, we vary both timing (10%-80% through pretraining) and mixture weights (10%-80% of training data). When testing the effect of timing, we use a fixed mixture weight of 20% for starcoder and vary the time at which the midtraining phase

Table 4.5: SFT and C4 validation losses for 70m and 160m models comparing default midtraining mixes to mixes of only the midtraining data (marked with 100%), averaged across 5 seeds for each SFT dataset. Bold values indicate best performance within each dataset and model size combination. Results reflect two continued pretraining trials in the code and math domains.

Model size	Downstream Dataset	Pre/midtrain mix	SFT val loss	C4 val loss
70m	Pycode	Pretrain-only	2.633	6.067
		Starcoder (20%)	2.530	5.996
		Ctd. pretrain (Starcoder)	2.580	6.160
	GSM8k	Pretrain-only	1.405	6.409
		Math (12%)	1.359	6.391
		Ctd. pretrain (Math)	1.401	6.422
160m	Pycode	Pretrain-only	2.382	5.247
		Starcoder (20%)	2.205	5.104
		Ctd. pretrain (Starcoder)	2.283	5.373
	GSM8k	Pretrain-only	1.188	5.339
		Math (12%)	1.139	5.198
		Ctd. pretrain (Math)	1.180	5.364

begins. When testing the effect of mixture weight, we fix the step at which midtraining starts at 40k training steps, while varying the mixture weight of starcoder as compared to C4.

As a special case of this, we also include a continuous pretraining baseline for the math and code midtraining mixes (100% midtraining mix), also starting at 40k steps, to distinguish midtraining from domain-specific continued pretraining.

Results When varying mixture proportions from a fixed starting point, effects were modest compared to varying the timing of introduction across different pretraining stages at a fixed 20% mixture weight (Figure 4.8). The timing effects are substantial: for the 160m model, early introduction at 6k steps achieves 2.205 validation loss compared to 2.374 when delayed to 50k steps. Similarly, the 70m model performs best with introduction at 20k steps (2.505) versus later introduction. This suggests that **when to introduce specialized data may be more critical than how much specialized data to include, at least within our experimental ranges**. See Figure 4.9 for mixture weight comparisons.

Relatedly, Figure 4.6 illustrates how midtraining benefits evolve over the course of pretraining for the 20% Starcoder mix (160m model). We finetune checkpoints from different pretraining steps on Pycode and measure both in-domain and C4 validation loss after fine-tuning. In-domain advantages emerge quickly after midtraining introduction (6k steps), while the C4 retention benefits develop more gradually, becoming apparent after approximately 20k steps. This temporal pattern suggests that early introduction of specialized data provides sufficient time for both immediate domain adaptation and

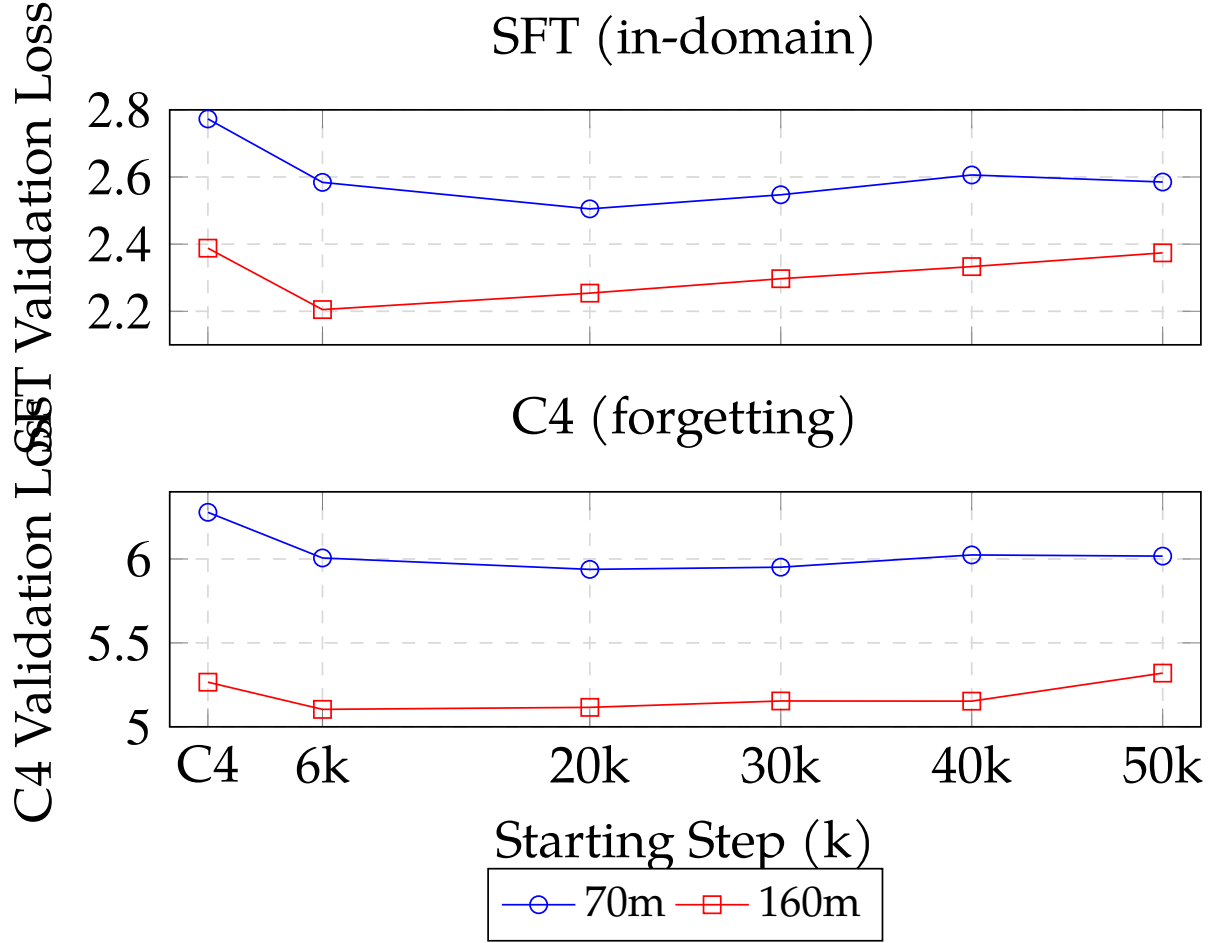


Figure 4.8: Timing ablation: validation losses (SFT and C4) by model size and starting step for the Pycode downstream dataset (in-domain). Top: SFT losses after Pycode finetuning; bottom: C4 losses showing forgetting after Pycode finetuning. In this setup, Starcoder is mixed at a mixed 20% ratio.

gradual integration with general capabilities.

4.3.7 How does midtraining change model representations?

Lastly, beyond performance improvements, we investigate how incorporating midtraining data changes the underlying model. We use linear Centered Kernel Alignment (CKA) to measure layer-wise similarity between model states in the code domain [143]. We extract activations from all layers using probe datasets (C4 and APPS [106]) and compute CKA similarity matrices between four key model states: base pretrained, midtrained (Starcoder), base fine-tuned, and midtrained fine-tuned. If midtraining creates better representations for downstream tasks, we expect to see smaller representational changes during fine-tuning for midtrained models compared to base models.

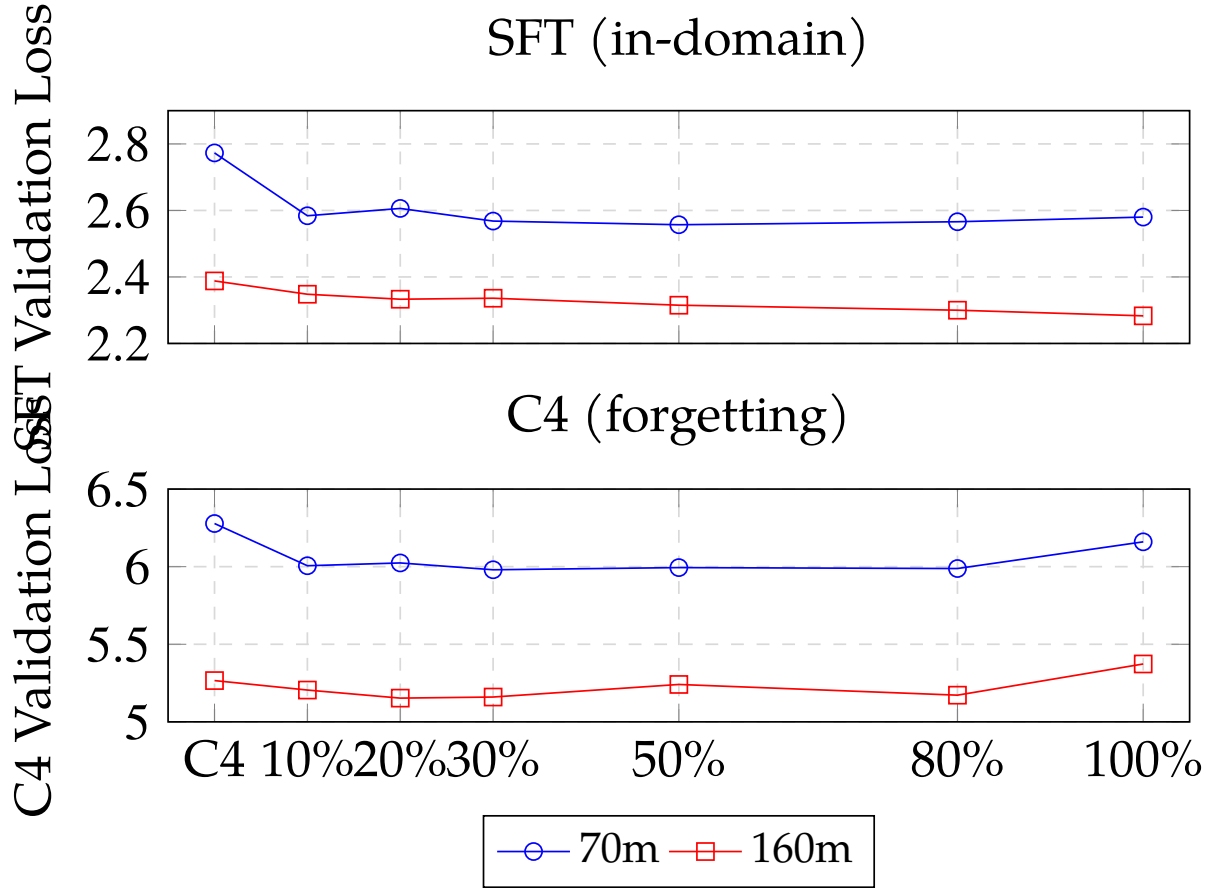


Figure 4.9: Mixture ablation: validation losses (SFT and C4) by model size and mix for the Pycode downstream dataset (in-domain). Top: SFT losses after Pycode finetuning; bottom: C4 losses showing forgetting after Pycode finetuning. The x-axis depicts the mixture weight of Starcoder mixed in at 40k steps.

Results Figure 4.10 shows the representational analysis for the 70m model. **The mid-trained model exhibits greater stability in the final layer after fine-tuning**, a pattern consistent across model sizes (see Appendix E.7 for the remaining results). However, the final fine-tuned models show high similarity regardless of whether models underwent midtraining. These effects are less pronounced for C4, which can be seen in Appendix E.8.

4.3.8 Related Work

Specific midtrained models Recently, several language model families have adopted midtraining approaches with varying implementation details [5, 43, 116, 198]. The mid-training phase duration varies from 2% [116] to 20% [43] of total training, motivating our systematic investigation of timing effects. Common midtraining domains include code, math, instructions, and higher-quality web data [198]—the domains we investi-

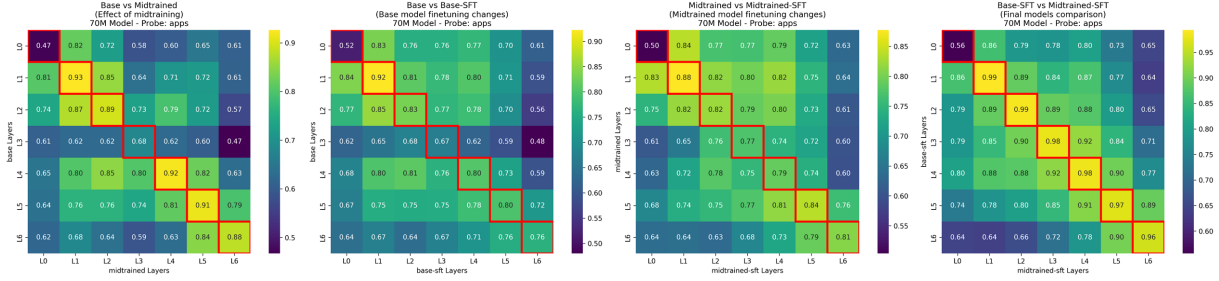


Figure 4.10: CKA analysis of model activations in the 70m model, probed with the APPS code dataset.

gate. Beyond general-purpose models, midtraining has shown benefits for specific tasks like RL [282] and GUI agents [312]. This widespread adoption motivates our questions of when and why midtraining provides downstream benefits.

Multi-stage pretraining Several works explore multi-stage pretraining, Blakeney et al. [31], Feng et al. [74] focusing on two-stage pretraining and Zhang et al. [313] proposing four-stage pretraining. These approaches demonstrate improvements over single-stage pretraining. However, these works evaluate base model performance after pretraining, whereas we focus on the post-finetuning setting to focus on benefits that also affect post-training.

Continual Learning Domain-adaptive pretraining (DAPT) and related approaches continue pretraining on domain-specific data [101]. Krishna et al. [145] show that pretraining on downstream data alone can rival full pretraining when evaluated after fine-tuning, suggesting pretraining-posttraining alignment matters—consistent with our findings. Mehta et al. [183] find pretraining reduces catastrophic forgetting during sequential fine-tuning; similarly, we observe midtrained models serve as better initializations with less forgetting.

Stability in Continued Pretraining Recent work addresses stability challenges during continued pretraining. Guo et al. [100] identify a “stability gap” where performance temporarily drops before recovering when shifting to new domains, Yang et al. [301] synthesize larger training corpora from small domain-specific datasets, and Lin et al. [164] introduce selective training on useful tokens only. While these works target training dynamics during continued pretraining, our approach examines how midtraining data selection affects post-fine-tuning performance, representing a complementary focus on end-task effectiveness.

Relationship between Pretraining and Finetuning Several recent works have explored incorporating instruction-formatted data during pretraining. Allen-Zhu and Li [10] show with an experiment on synthetic Wikipedia-style data that augmenting pretraining

data with QA-formatted data improves subsequent fine-tuning, and Jiang et al. [125] and Cheng et al. [48] demonstrate this in a practical context as well. Sun and Dredze [262] find continual pretraining benefits emerge only after fine-tuning, while Springer et al. [258] show extended pretraining causes catastrophic forgetting (“overtraining”), particularly on math/code domains least aligned with web data. It is possible midtraining may prevent overtraining by introducing specialized data earlier and providing a better initialization for posttraining.

4.3.9 Conclusion

We conduct the first systematic investigation of midtraining through controlled experiments. We demonstrate that midtraining benefits are domain-specific, with the most substantial improvements in math and code domains that are not well represented in standard web pretraining corpora.

Beyond domain-specific improvements, we also find that midtraining mitigates catastrophic forgetting of general language modeling abilities after specific supervised fine-tuning. Additionally, our comparison to continued pre-training with the same data reveals that maintaining some general pretraining data is helpful when adapting language models to new domains. Our ablations of timing and mixture weights show that at least in the case of code, the start point of midtraining may have a larger impact than the mixture weight, holding the other factor constant.

Rather than treating midtraining as a one-size-fits-all solution, our results suggest that its effectiveness depends critically on the alignment between midtraining domains and target applications. Future work remains in understanding these training dynamics, as well as systematically investigating midtraining’s effects at larger scales. Additionally, although we did not investigate RL-based posttraining in this work, understanding how midtraining interacts with methods such as RLHF and other preference-based approaches is an important avenue for future work. As language models continue to be used for more specialized and complex tasks, developing rigorous methods for pre-training and midtraining language models and understanding how to best adapt them to new domains will become essential for maximizing task-specific performance while maintaining knowledge from pretraining.

4.4 Elemental Skills and their Emergence (proposed)

4.4.1 Motivation

Previous sections have shown that pretraining is not as monolithic as assumed as different tasks exhibit distinct scaling properties and benefit from targeted data introduction at specific training stages. However, NLP tasks remain complex, requiring models to handle multiple subtasks such as memorization, logical reasoning, and copying mechanisms. Understanding whether a model lacks domain-specific data or fundamental skills like logical reasoning remains an ad-hoc process.

While tracking performance during pretraining provides guidance, it offers limited insight into whether models learn generalizable solutions. Mechanistic circuits have been identified for simple tasks, but scaling circuit discovery to concrete high-level tasks remains nontrivial, and it's unclear how to decompose tasks systematically in the first place. However, due to the release of recent interpretability tools [8, 12] and recent interest in and similar works in training dynamics [132, 218, 317], now is a good time to pursue this project.

We propose developing a suite of “elemental tasks” – simple and easy to evaluate tasks that serve as building blocks for complex behaviours models learn to perform. These intermediate-level abilities would be higher-level than attention operations but lower-level than downstream benchmarks. Examples include sorting lists, counting, or copying tokens from context. Crucially, these elemental skills should compose to form higher-level tasks, enabling both granular capability tracking and targeted synthetic data generation.

By isolating fundamental capabilities, we can analyze training dynamics to understand which tasks are foundational for learning certain abilities. This approach could accelerate learning on targeted capabilities while providing researchers tools to understand model capabilities at a more granular level than current benchmarks allow.

4.4.2 Research Questions

Our investigation focuses on four key questions:

- **Task Decomposition:** How can we systematically decompose high-level NLP tasks into individually benchmarkable subcomponents?
- **Emergence Timing:** When do elemental subtasks emerge during pretraining, and is there a consistent developmental order across models and datasets?
- **Capability Dependencies:** Do certain elemental tasks serve as prerequisites for others, and how do these dependencies affect overall task performance?
- **Targeted Interventions:** Can we accelerate training by strategically introducing synthetic data for rate-limiting elemental skills, and does this transfer to composite tasks?

4.4.3 Proposed Experiments

Our research plan consists of three integrated components:

(1) **Elemental Task Suite Development:** We will design minimal-functionality tasks spanning core cognitive abilities: copying mechanisms, in-context learning, basic arithmetic, pattern completion, and memory retrieval. Each task will have controlled synthetic variants enabling systematic evaluation. For instance, pattern completion ("a": "cat", "b": "dog", "abb":) isolates copying abilities that underlie more complex behaviors.

(2) **Training Dynamics Analysis:** Using checkpoints from open-weights pretrained models such as Olmo2 [198], we will track when each capability emerges and analyze co-occurrence patterns. This will reveal whether certain elemental skills consistently develop before others and identify potential bottlenecks in capability acquisition.

(3) **Synthetic Data Interventions:** For rate-limiting elemental tasks identified in phase 2, we will generate targeted synthetic data and test whether early introduction accelerates both elemental skill development and performance on composite downstream tasks that depend on those skills.

4.4.4 Expected Outcomes

We expect to contribute the following deliverables with this project:

- **Elemental Task Suite:** We plan to release these tasks as a benchmark for others to use, alongside synthetic data generators for each task and compositions of tasks.
- **Circuit Development Tracking Methodology:** There is currently no way to track the development of circuits that perform tasks in a granular way in the course of training. Part of this issue lies in the brittleness of current circuit discovery algorithms. We plan to experiment with and develop methodology to do this if possible.
- **Training Dynamics Characterization:** We plan to release empirical findings on developmental ordering of elemental skills and identification of common bottlenecks across models and datasets.

4.5 Conclusion

In this chapter, we have moved beyond evaluating cognitively relevant skills and towards linking final performance on downstream tasks to design decisions such as data composition as well as the timing of specialized data introduction during pretraining. We build on this to propose that we can enumerate and track elemental skills throughout the course of training which can be composed to predict the performance of language models on more complex tasks. In the next section, we focus on the natural culmination of this line of thought: systematic improvement of specific capabilities in practice that have been more resistant to improvement through model scaling, in many cases due to limited data.

CHAPTER 5

SCAFFOLDING DEVELOPMENT FOR REAL-WORLD TASKS

5.1 Background

In the previous chapters, we have diagnosed capability gaps motivated by cognitive science and also developed a more granular understanding of how pretraining data impacts downstream model capabilities. We now turn to targeted interventions for improving specific capabilities that are difficult to improve in the current training paradigm due to data scarcity or a lack of flexibility in training.

We begin with a concrete example of targeted capability improvement: addressing the challenge of translating idiomatic expressions (§5.2), which remain problematic despite massive training corpora due to their cultural specificity and long-tail nature.

Building on these insights about targeted interventions, we propose working on a more complex case: repository-level coding (§5.3), where the scarcity of high-quality training data (pull requests with tests) limits model capabilities. We propose decomposing this complex skill into elemental components that can be synthetically generated and composed.

Finally, but more speculatively, we also examine a natural extension of our capability-centered framework: rather than externally diagnosing capability gaps, could models learn to identify and address their own deficiencies during training? This preliminary investigation explores adaptive pretraining methods that could enable more autonomous capability development (§5.4).

5.2 Learning to Translate Idiomatically

5.2.1 Introduction

Idioms are conventionalized expressions whose intended meaning differs from their literal translation. Despite their abundance, they remain problematic for traditional machine translation systems, which tend to translate them literally, leading to incomprehensible outputs [58, 244]. While most work has focused on identifying this problem or interpreting model behavior [58, 103], fewer efforts have attempted systematic remedies.

We frame this as a non-compositional translation problem encompassing idioms and multi-word expressions that cannot be translated word-for-word. Through synthetic experiments, we find that transformer models translate word-for-word until a threshold proportion of sentences contain non-compositional expressions, then flip to correct translations. Commercial model evaluation across three languages confirms performance drops on idiomatic sentences, with better performance on common idioms—suggesting this follows patterns observed in other long-tail phenomena [131].

Building on this insight, we improve idiomatic translation in a pretrained model (Δ LM) using two simple methods: upweighting training loss on potentially idiomatic sentences and retrieval augmentation. Testing on French, Finnish, and Japanese idiom datasets, we achieve 10.4% average improvement in absolute accuracy without harming literal translation quality. Human evaluation shows 7.52% reduction in severe semantic errors, demonstrating potential for addressing this critical barrier to translation accessibility.

5.2.2 Non-Compositional Translation

Background on Idioms

Idioms are commonly understood to be fixed expressions that contradict the principle of compositionality in language, which is to say that their meaning cannot be predicted from the meanings of their parts [216, 227]. Idioms occur relatively frequently in all languages, and are often challenging for non-native speakers [55]. For instance, a literal translation of one Portuguese idiom is *"it is from little that you twist the cucumber"*. This is difficult to understand. However, an equivalent English expression is *"As the twig is bent, so is the tree inclined"*, which refers to actions during childhood influencing behaviours that people have as adults [273]. This example illustrates the importance of translating idioms using equivalent idioms from the target culture, or a paraphrase if there is no equivalent.

Idiomatic expressions are heavily shaped by the culture of language speakers, including religious beliefs, history, geography, and cuisine. For instance, food-related idioms in English tend to refer to foods such as beef and potatoes, while in Chinese, these idioms tend to refer more to rice and tofu [299]. Cross-cultural knowledge is important in choosing a translation that conveys the proper intent to readers in the target language [167]. Overly-literal translations and lack of broader context are two reasons why ma-

chine translation is still not at parity with human translators, particularly when translating literary text [181, 199, 213].

Formal definition

We use the idea of non-compositionality to frame idiomatic translation more precisely. Let $\mathcal{X} = \{x_1, \dots, x_N\}$ be the set of tokens in the source language, and $\mathcal{Y} = \{y_1, \dots, y_M\}$ be the set of tokens in the target language. Suppose that we have an oracle function $\text{TRANSLATE} : \mathcal{X}^* \rightarrow \mathcal{Y}^*$ that always produces a correct translation. We can imagine this to be a helpful speaker who is perfectly familiar with both languages and never misreads text. Then we can say that a multi-token string requires non-compositional translation if it can be translated correctly by the oracle as a whole, but it cannot be translated correctly by individually translating parts of the sentence and joining them (according to the target language’s word order). In other words, for a string of tokens $x_1, \dots, x_n$,¹

$$\bigoplus_{i=1}^n \text{TRANSLATE}(x_i) \neq \text{TRANSLATE}\left(\bigoplus_{i=1}^n x_i\right) \quad (5.1)$$

We note that this definition is very general and also includes other phenomena such as multi-word expressions and named entities. However, we can now use this definition to create a relevant synthetic task, allowing us to observe translation compositionality under different settings (§5.2.4).

5.2.3 Idioms and Data Collection

We can use the formal definition from the previous section to generate synthetic data for experiments. However, we ultimately want to improve translation of real idioms. To do so, we collect a dataset of natural sentences to evaluate commercial systems and the model we seek to improve.

Although a large corpus of potentially idiomatic expressions exists in English [102], there are no readily accessible equivalents in other languages. Therefore, we collected idioms in French, Finnish, and Japanese from language-learning sites, listed in §F.2. These languages were chosen for phylogenetic diversity, and due to availability of commercial translation systems. In total, there were 148 French idioms collected, 92 in Finnish, and 1336 in Japanese.

To collect sentences containing these idioms, we matched on lemmatized forms from the 2018 version of OpenSubtitles [166], where lemmatization was performed with Stanza [221]. In total, there were 85632 French sentences containing potentially idiomatic expressions, 51811 Finnish sentences, and 23018 Japanese sentences. To filter out unaligned

¹ $\bigoplus_X$ denotes string concatenation given the word order of language X , i.e. if the word order is SVO, the tokens belonging to the subject should be placed in front of the tokens belonging to the verb, and so on.

sentences, we scored each source and reference sentence using COMET-QE [232] and removed the bottom 10% of each language’s sentences by COMET-QE scores.

Some idioms have a plausible literal meaning (such as “kick the bucket” to mean kicking a physical bucket). To make sure that all examples in the idiomatic test set were actually idiomatic, we sorted sentences into an idiomatic test set where the idiomatic meaning of a phrase was used (e.g. “to die”) and a literal test set, where the literal meaning of the phrase was used (e.g. kicking a physical bucket). The first 100 examples containing each idiom’s lemmatized form were collected, and up to the first 3 (for Japanese) or 5 (for Finnish and French) literal and figurative examples in this set were collected to create the test set. This was to avoid dominance of very common idioms in the test set. This created two test sets related to the idiom list for each language, the *idiomatic* and *literal* test sets.

To validate these judgments, we hired native annotators in French and Finnish. They were presented with examples from the final literal and idiomatic test sets in a shuffled order, and asked to label them with idiomatic, literal, or N/A labels if they didn’t think it was an instance of either. Agreement (Krippendorff’s α [39, 144]) in both cases was moderately high (French $\alpha = 0.5754$, Finnish $\alpha = 0.6454$). Details can be found in § F.4.

Finally, we collect two random test sets, one which is in-domain and another which is out-of-domain. For the in-domain test set, we simply select sentences from the development set of OpenSubtitles (see § 5.2.6 for details on our split of OpenSubtitles). For the out-of-domain test set, we use the Ted Talks corpus [233]. This is to ensure that translation quality of other, unrelated sentences is not impacted by any modifications meant to improve translation of idioms. Topics discussed and vocabulary used in Ted Talks may be slightly different from what is discussed in movies or TV shows, so training the model on OpenSubtitles and testing on Ted Talks allows us to evaluate model generalization. For both test sets, to control for translation length as a source of difficulty, sentences were length-matched on the target side with corresponding sentences in the idiomatic set. This created the *random* set, which is the same size as the idiomatic test set. All three test sets are summarized in Table 5.1.

Language	Idiom matches	Idiomatic	Literal	Random (in)	Random (out)	Total
fr	85632	777	79	777	777	2410
fi	51811	449	81	449	449	1428
ja	23018	3253	389	3253	3253	10148

Table 5.1: Size of test sets for each language. The idiomatic and literal sentences contain strings matching known idioms (after lemmatization), and the in-domain random set contains unrelated sentences from OpenSubtitles, but the out-of-domain random set contains unrelated sentences from the Ted Talks corpus.

5.2.4 Evaluating Non-Compositional Translation

Artificial Language Translation

We first use the definition of non-compositional translation in (§5.2.2) to create a synthetic task. This allows us to gain an understanding of how much data is required to memorize non-compositional patterns. Although this experiment is not realistic to natural language (notably, there is no token-level ambiguity in this experiment), we note that using synthetic experiments allows us to easily extend the data generation setup and examine model behaviour along many different conditions, such as informativity.

The source language in these experiments was composed of tokens 0 through 9, $X = \{0, 1, 2, \dots, 9\}$. The target language was produced by adding 10 to each token, $Y = \{10, \dots, 19\}$. The translation rule was to add 10 to the value of each token in the source language, e.g. $0 \rightarrow 10$, $1 \rightarrow 11$. We add a single non-compositional rule that doesn't follow this trend, $0\ 1 \rightarrow 12$ (rather than $0\ 1 \rightarrow 10\ 11$). We limited the maximum sequence length to 6 tokens.

We generated synthetic training corpora of several sizes containing different numbers of occurrences of the non-compositional rule $0\ 1 \rightarrow 12$. The number of training sentences ranged from 100k to 10M, while the number of noncompositional occurrences ranged from 10 to 1M. We examined two informativity conditions, corresponding to the case where the context provides no information (tokens are randomized around the non-compositional expression), and the context being perfectly informative. The perfect informativity condition was achieved by adding the canary token “11” to the source vocabulary, and only inserting this token prior to the non-compositional pattern “0 1”.

We experimented with three different transformer sizes [274], each of which had a hidden dimension and embedding size of 512, as well as 16 attention heads. Only the number of encoder and decoder layers varied, such that the small transformer had 3 encoder and decoder layers, the medium transformer 8, and the large transformer 16. We fix the number of epochs for the small, medium and large models to respectively be 10, 20, and 30 in the non-informative case and 15, 15 and 25 in the informative case.² Further training details can be found in § F.1.

Although this may seem like a simple task, we found it surprisingly difficult for models to learn this non-compositional pattern. Results in each setting, averaged across 5 random seeds, are presented in Figure 5.1. Especially for the small model, there is a sharp gradation from translating none of the non-compositional expressions correctly to translating them all correctly, which occurs when roughly 10% of training data contains a non-compositional pattern. A similar trend exists for larger models, but the threshold is less distinct. This corroborates the tendency for transformers to translate non-compositional phrases literally [58]. Comparatively less data is required when the context is infor-

²The number of training epochs was determined by the number of epochs it took for the validation loss to plateau in the 100k size corpus with 1k non-compositional examples, rounded up to multiples of 5. This was done to mimic the typical training process for MT models, which are trained until loss or accuracy plateaus on a general dev set. Since idiomatic expressions tend to be uncommon compared to literal ones, there may not be many in the dev or train sets, and so the model's performance on idiomatic expressions may not be tracked.

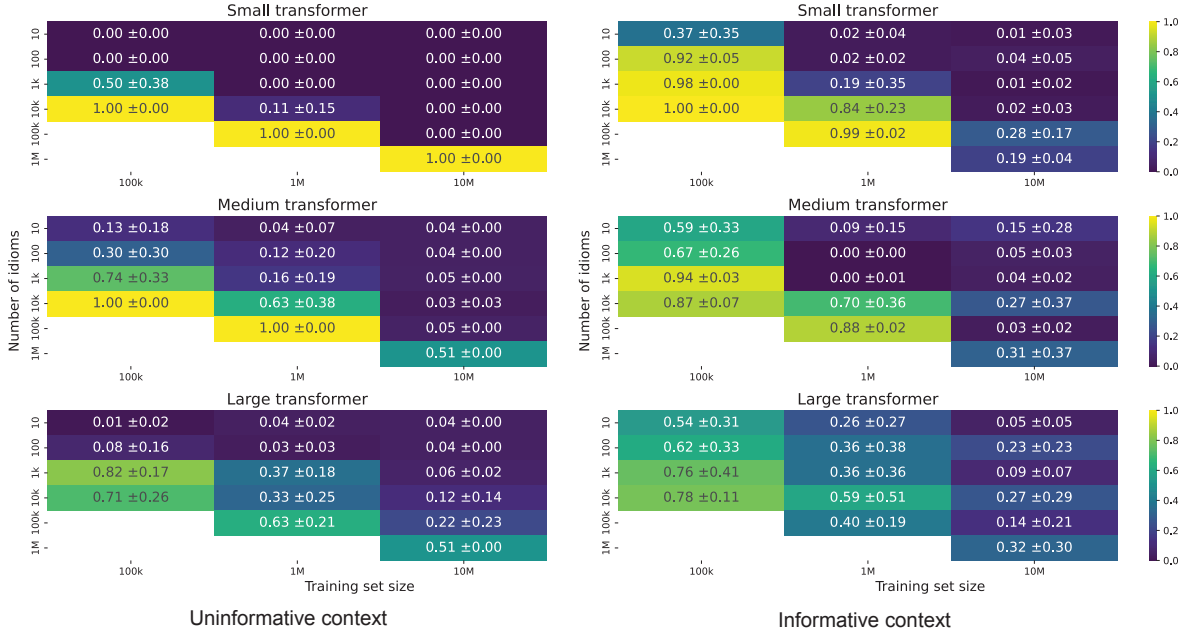


Figure 5.1: Accuracy of a transformer in translating a non-compositional phrase after training on datasets of different sizes, with different numbers of non-compositional patterns (only non-compositional translation accuracy is depicted). Results are averaged across 5 seeds, and standard deviation is shown.

mative, but the trends remain similar to the non-informative case. As model size and corpus size increase, the rate of correct translations for non-compositional examples actually drops, contrary to expectation.

It is unlikely that any individual idioms occur in 10% of sentences in natural language. Due to the highly regular translation rules in this synthetic language, there may be a stronger bias toward translating compositionally in this experiment. However, we gain the intuition that idioms can be translated effectively if they appear frequently, and that clear context clues reduce data required.

Evaluation of Commercial Systems

Although synthetic experiments provide intuition on the difficulty of translating idioms, one might ask whether similar results hold in natural language. To answer this, we examine the performance of commercial systems on the test sets in (§5.2.3). Namely, we examine Google Translate and DeepL on Finnish, French, and Japanese idiomatic, literal, and random sentences. Results are in Table 5.2. We observe drops in translation quality on idiomatic sentences in all languages, with lower automatic metrics overall.

Although it’s impossible for us to determine what data these commercial systems were trained on, we examine the frequency of each idiom within OpenSubtitles as a proxy for its overall frequency in the training data, and bucket idioms into quintiles based on their occurrence frequency in source text. As idioms become more frequent, the

quality of translations increases. An example of DeepL on the French idiom set is shown in Figure 5.2. Trends for other languages and systems are in § F.8. This indicates that like in the synthetic experiments, there may be strong frequency effects on translation quality of idioms.

Language	System	BLEU	METEOR	BERTScore
$f i_{\text{idiomatic}}$	DeepL	0.1001	0.2497	0.8866
	Google	0.0923	0.2250	0.8726
	Δ LM-base	0.1608	0.3592	0.9126
$f i_{\text{literal}}$	DeepL	0.1488	0.3908	0.9146
	Google	0.1398	0.3577	0.9017
	Δ LM-base	0.2093	0.5050	0.9350
$f i_{\text{random-out}}$	DeepL	0.2052	0.4082	0.9103
	Google	0.2288	0.4357	0.9062
	Δ LM-base	0.2365	0.4971	0.9145
$f r_{\text{idiomatic}}$	DeepL	0.1575	0.3278	0.9006
	Google	0.1261	0.2794	0.8808
	Δ LM-base	0.2001	0.4393	0.9211
$f r_{\text{literal}}$	DeepL	0.2219	0.4022	0.9122
	Google	0.2034	0.3830	0.9012
	Δ LM-base	0.2778	0.5504	0.9377
$f r_{\text{random-out}}$	DeepL	0.2854	0.4650	0.9125
	Google	0.3103	0.4922	0.9149
	Δ LM-base	0.2778	0.5504	0.9377
$j a_{\text{idiomatic}}$	DeepL	0.1172	0.2735	0.8932
	Google	0.0672	0.1839	0.8644
	Δ LM-base	0.09048	0.2998	0.9234
$j a_{\text{literal}}$	DeepL	0.1517	0.3440	0.9059
	Google	0.0937	0.2565	0.8829
	Δ LM-base	0.1416	0.4222	0.9222
$j a_{\text{random-out}}$	DeepL	0.1074	0.2934	0.8878
	Google	0.1079	0.2834	0.8829
	Δ LM-base	0.0948	0.3436	0.8946

Table 5.2: Performance of commercial systems on idiomatic, literal, and random test sets. There is a clear degradation in performance on idiomatic sentences.

5.2.5 Methods to Improve Non-Compositional Translation

We explore two methods to improve translation, loss weighting and kNN-MT. These two methods are relatively simple to use, where loss weighting only requires a list of

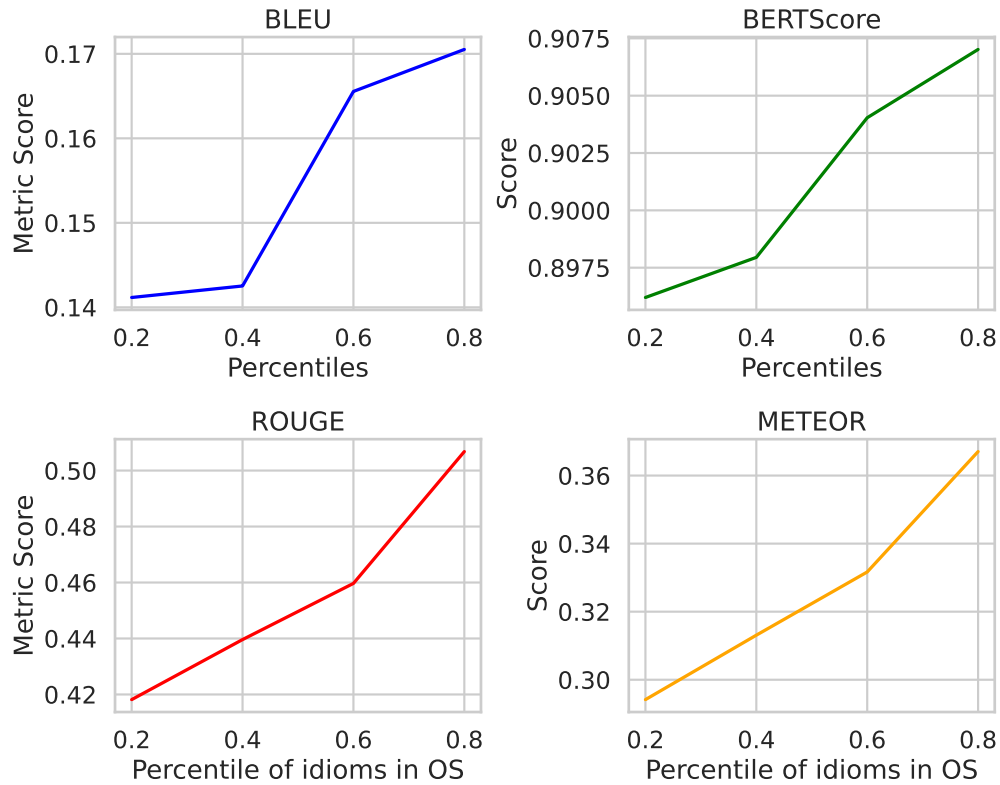


Figure 5.2: Automatic metrics – Quality of DeepL French translations on idiomatic test set bucketed by idiom frequency. The bottom 20% of least common idioms are excluded, as they may occur fewer than 3 times and not be in our test set.

potentially idiomatic phrases in the source language, and kNN-MT only requires enough space on disk to save the datastores.

More formally, we consider the basic case of autoregressive machine translation, with a set of parallel sentences in the source ($X = \{x^{(i)}\}_{i=1}^N$) and target ($Y = \{y^{(i)}\}_{i=1}^N$) language: $\mathcal{D} = \{(x^{(i)}, y^{(i)}), \dots, (x^{(N)}, y^{(N)})\}$. The model p_θ with parameters θ is trained by minimizing the loss:

$$\mathcal{L}(\theta, \mathcal{D}) = \sum_{i=1}^N \ell(y^{(i)}, p_\theta(x^{(i)})) \quad (5.2)$$

Upweighting here refers to sentence-level upweighting, where there is a set of sentences A that we'd like to upweight with a weight coefficient α . In this case, A would be potentially idiomatic sentences. We keep all other parameters for training the same as in the base model.

$$\mathcal{L}(\theta, \mathcal{D}) = \sum_{i=1}^N \alpha^{\mathbb{1}(x^{(i)} \in A)} \ell(y^{(i)}, p_\theta(x^{(i)})) \quad (5.3)$$

kNN-MT augments a translation model with a retrieval component [138]. Given each sentence (x, y) , we construct a datastore with keys based on hidden representations constructed from the translation model, and values being the next word in the target sentence.

During generation, a probability distribution over next words can be computed based on the retrieved next words and the distance of their keys to the current context. A parameter λ controls interpolation between the distribution over next words predicted by the base model, and the distribution predicted by the retrieved k neighbours.³

$$\begin{aligned} p(y_i^{(j)} | x^{(j)}, y_{1:i-1}^{(j)}) &= \lambda p_{\text{kNN}}(y_i^{(j)} | x^{(j)}, y_{1:i-1}^{(j)}) \\ &\quad + (1 - \lambda) p_\theta(y_i^{(j)} | x^{(j)}, \hat{y}_{1:i-1}^{(j)}) \end{aligned} \quad (4)$$

We also combine loss weighting with kNN-MT, where a model is trained with sentence upweighting and interpolated with a datastore based on representations from the upweight-trained model.

Intuitively, these methods make sense to use for idiom translation – we have previously seen that one problem with non-compositional phrases may simply be their rarity. Upweighting training examples that contain idioms may help with under-representation. Furthermore, retrieving similar examples may find occurrences of the same idiom which were translated correctly.

³We run a hyperparameter search using the validation set to find the best kNN-MT settings for each language. Further details are in § F.3.

5.2.6 Experimental Settings

Experimental Settings

We run experiments on Δ LM-base, a transformer encoder-decoder model with 360M parameters, a larger version of which ranked first in the WMT21 multilingual translation task [177, 250]. We train one Δ LM model for each language pair. Each model was trained for 2 million steps, and the checkpoint with the best loss on the validation set was kept. Further details are in § F.3. To decode, we used beam search with a beam size of 5.

Data

Models were trained on OpenSubtitles for each language pair. Data from test sets were removed, and 10% of the remaining data was used as a validation set. There were 33.8M sentences in the fr-en train set, 22.0M in fi-en, and 1.6M in ja-en.

Evaluation

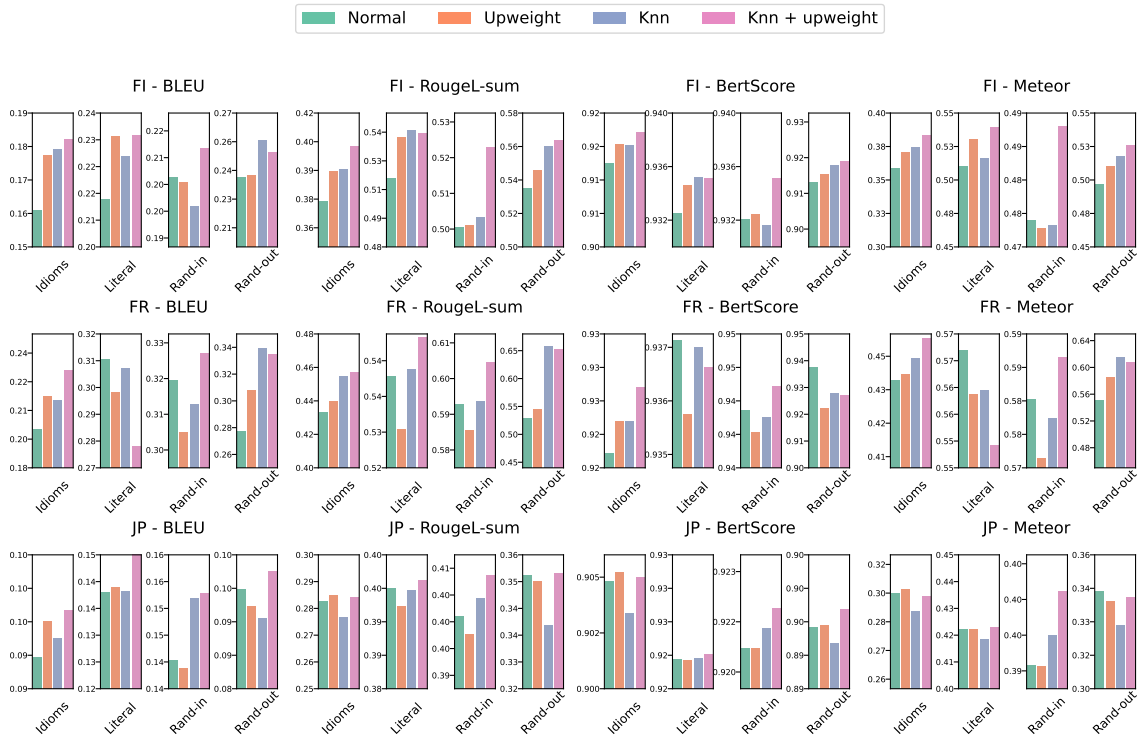


Figure 5.3: Results of automatic metrics. In most cases, combining loss weighting with KNN-MT improves automatic metrics the most on all three test sets, including the out-of-distribution (Random) test set.

We use multiple automatic metrics to evaluate translation quality. However, due

to the importance of accurate semantic evaluation, the authors (native English speakers and fluent in French and Japanese) conduct a human evaluation inspired by MQM [172]. Only errors that would fall under the “terminology” and “accuracy” error types are considered, as we are focused on severe semantic errors. We give a score of 0 for severe errors and a score of 0.5 for major errors. A score of 1 is given otherwise. Exact evaluation standards are in § F.5.

5.2.7 Results

Automatic and Human Evaluation

In most cases, as reported in Figure 5.3, using a combination of sentence upweighting and kNN-MT led to the greatest increase in automatic metrics on all three test sets, of up to 3.08 BLEU points on the idiomatic test set (fr), 2.69 BLEU points on the literal test set (fi), and 5.75 points on the random test set (fr). In all cases except ja-rand, using one or more of these methods improved over the baseline. Exact numerical results are in § F.10.

We evaluate the statistical significance of the results through a one-tailed permutation test [96]. Further details are in § F.6. Exact results are in § F.7. For Finnish, significance is achieved for all three test sets, and for French, significance is achieved for the idiomatic and random test sets. For Japanese, values achieved are not significant, but are borderline.

As our focus is on mitigating semantic errors, we mostly focus on the results of human evaluation, which are summarized in Table 5.3. Here, we also find that using both sentence upweighting and kNN is the best condition in most cases, increasing accuracy by roughly 13% in French and Finnish, and 4.5% in Japanese for idiomatic sentences. Encouragingly, this does not overly harm translation of literal sentences, as accuracy on the literal set either increases slightly (by roughly 4% in French and Finnish), or decreases very slightly (by roughly 0.4% in Japanese). For the random set, the combination of sentence upweighting and kNN-MT by around 7% accuracy. However, in Japanese, performance on the random test set decreases by 4%. In all cases except ja-rand, one or more of these methods improves over the baseline.

We note that the Japanese model was trained on roughly 1/10th of the data of the French and Finnish models, so its translations are not as high-quality. This also leads to the construction of a much smaller datastore, which may lead to weaker performance on the random set.

Error analysis

We repeat the frequency analysis performed on commercial systems (§5.2.4) for Δ LM, and find that adding upweighting and kNN-MT generally improves translations at all frequency levels. These increases are not concentrated in low-frequency idioms, so more

	base	knn	upweight	upweight + knn
fr-idioms	0.6177	0.6659	0.7010	0.7463
fr-literal	0.7039	0.7303	0.7105	0.7434
fr-rand-out	0.7526	0.8398	0.7477	0.8232
fi-idioms	0.4803	0.5562	0.5604	0.6194
fi-literal	0.7692	0.8462	0.8205	0.8141
fi-rand-out	0.7647	0.8235	0.7771	0.828
ja-idioms	0.4152	0.4286	0.4643	0.4598
ja-literal	0.6475	0.6516	0.6557	0.6434
ja-rand-out	0.6207	0.5560	0.5776	0.5862

Table 5.3: Human-judged accuracy on sentence-level semantics.

common idioms continue to be translated better.⁴ A representative example (for French) is in Figure 5.4. A complete set of plots are in § F.9.

We examine the rate of severe and major errors made in the base model and the upweight+knn model in Table 5.4. In French and Finnish, the rate of critical errors decreased greatly, particularly in the idiomatic and random test sets. This is true to a lesser extent in Japanese. Major errors also decreased to a lesser extent. The only test set where errors increase is again the ja-rand test set. We note that it’s possible for the rate of major errors to be higher in the upweight+knn model because some severe errors transitioned to major errors.

One question is why the error rate on out-of-distribution sentences drops for French and Finnish. In fi-rand, the severe error rate more than halves ($0.1317 \rightarrow 0.603$), and in fr-rand, it nearly halves ($0.1624 \rightarrow 0.09407$). However, it is unclear why this should be the case. We examined sentences where the original translation was incorrect but the upweight+knn translation was correct, and found that they tended to contain named entities. For instance, for the sentence “*La chirurgie à coeur ouvert au Nigeria, c’est un gros problème.* (Open heart surgery in Nigeria - big trouble.)”, the base model incorrectly produced the translation “Open-heart surgery in Forbes, that’s a big problem.”, while the upweight+knn model translated correctly. In some cases, words with multiple possible translations (e.g. *spectre*: ghost, spectrum) became correctly translated. “*Mais regardez le nombre de lignes noires dans ce spectre.* (But look at the number of black lines in that spectrum.)” was originally translated incorrectly as “But look at the number of black lines in that ghost”.

5.2.8 Related Work

Recent work has raised the issue of idiom handling in MT [16, 57, 58]. There is historical recognition of the problem, including of multi-word expressions [37, 236, 307].

⁴This trend is different in retrieval of long-tail facts in question answering, in which retrieval flattens out the difference between rare and common facts [131].

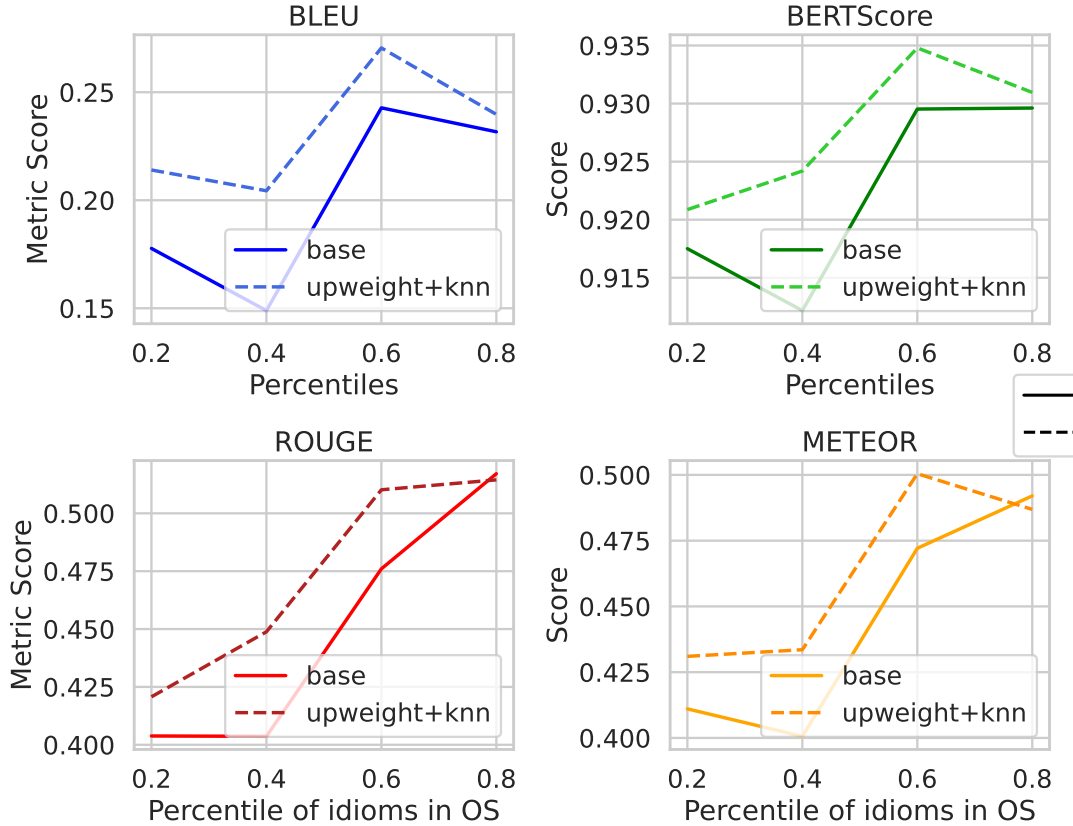


Figure 5.4: Automatic metrics for fr-idiom sentences, plotted by frequency, for base and upweight+knn.

This has historically motivated example-based machine translation [192]. Similar motivations underlie the use of kNN-MT. However, neural models may already be capable of translating idiomatic phrases if they appear often enough in training data.

Other works focus on data augmentation and creating new data resources [4, 71, 102, 110]. A related task is detection of conventionalized metaphors [156]. Automatic identification of idiomatic phrases, as well as data augmentation are promising avenues to improve performance in lower-resource languages.

Instance weighting has been explored previously in the MT literature, but has been mostly explored in the context of domain adaptation, rather than being used to improve translations of rare or non-compositional phrases in the same domain [77, 279].

Idiomatic phrases are a prototypical case of phrases that need to be memorized [103]. Many also occur infrequently in training data, which may make it difficult for transformer-based models to translate them [131]. This can be mitigated, as we have shown in this paper. However, more work is needed to effectively learn idioms and other infrequent linguistic elements with few repetitions.

	System	Severe ($\downarrow$)	Major ($\downarrow$)
fi-idioms	base	0.4258	0.1648
	upweight+knn	0.3242	0.0962
fi-literal	base	0.1728	0.1234
	upweight+knn	0.1234	0.1358
fi-random	base	0.1317	0.2009
	upweight+knn	0.0603	0.2188
fr-idioms	base	0.3042	0.1528
	upweight+knn	0.198	0.1092
fr-literal	base	0.2326	0.1047
	upweight+knn	0.2209	0.04651
fr-random	base	0.1624	0.1688
	upweight+knn	0.09407	0.1649
ja-idioms	base	0.4643	0.2411
	upweight+knn	0.4464	0.1875
ja-literal	base	0.2867	0.1311
	upweight+knn	0.2787	0.1557
ja-random	base	0.2931	0.1724
	upweight+knn	0.3190	0.1897

Table 5.4: Rate of major and severe errors in translations.

5.2.9 Conclusion

We highlight the challenge idiomatic expressions pose to machine translation systems and provide simple solutions to improve performance. Through synthetic experiments, we identify a threshold at which transformer-based models correctly default to idiomatic translations. We develop a dataset of sentences containing idiomatic expressions in French, Finnish, and Japanese, and introduce two techniques - upweighting training loss on potentially idiomatic sentences and augmenting models with kNN-MT - which enhance the idiomatic translation accuracy of a strong model, while offering potential benefits for non-idiomatic sentences.

Future research could extend these techniques to additional languages, and explore their effectiveness in dealing with other long-tail phenomena. We hope that this work contributes toward increasing the intelligibility of translations containing idioms or set phrases. Ultimately, for machine translation to be useful for everyone without causing misunderstandings, “last mile” problems involving cultural knowledge, long-tail phenomena, and complex semantic evaluation should be taken into account.

5.3 Skill-driven Model Training: A Study in Repo-level Coding (proposed)

5.3.1 Motivation

Building on the elemental tasks framework from § 4.4, we propose to apply this decomposition approach to a practical problem: training effective repo-level coding agents. Current approaches to agent training face significant scalability limitations as tasks become more complex and long-horizon. In the case of coding, high-quality repo-level coding examples are constrained by the availability of high-quality pull requests with tests. Specific datasets of trajectories have been generated for supervised fine-tuning use with models, but we propose to make this process more targeted and scalable [204, 300].

Rather than training exclusively on complete issue-solving trajectories, we want to first break down successful issue-solving into components such as file and function localization, documentation writing, and interacting with files. We plan to experiment with generating data for midtraining and SFT at scale from existing code repositories without requiring execution, though we will test our approach by generating smaller amounts of SFT data to begin with. Recent work has shown that midtraining on reasoning-intensive tasks can yield improvements on GUI agent performance [311], motivating our work on creating larger-scale agent training data in the code case as well.

Our hypothesis is that models trained on carefully designed elemental coding tasks will demonstrate superior performance on complex repo-level tasks compared to those trained solely on end-to-end trajectories, while providing greater interpretability about which capabilities limit overall performance.

5.3.2 Research Questions

- **Task Decomposition:** How can we systematically decompose repo-level coding tasks (issue resolution, feature implementation, debugging) into trainable elemental components that capture essential coding agent skills?
- **Training Efficacy:** Does targeted training on elemental tasks followed by integration training outperform direct training on complete trajectories for repo-level benchmarks like SWE-Bench and Commit0 [126, 314]?
- **Midtraining Benefits:** Given the domain gap between natural language pretraining and coding agent behaviors, does a midtraining phase provide benefits over simply performing supervised fine-tuning? If so, what makes good midtraining data?
- **Skill Transfer:** How do improvements in specific elemental tasks correlate with gains on composite downstream tasks, and can we predict overall coding agent performance from elemental task proficiency?

5.3.3 Expected Outcomes

- **Improved performance on coding benchmarks:** We will measure success by improvements on repo-level coding benchmarks using open source models such as Qwen3 [298].
- **Scalable data generation:** We plan to release our elemental tasks for coding agents so that others can generate synthetic training data corresponding to each task, given their own raw code files or Github data.
- **Understanding of bottlenecks in agentic tasks:** Open-source models currently lag behind closed-source models in complex coding tasks, but this is likely not due to not training on enough code during pretraining. It is likely that some format adaptation and targeted training on specific skills may eventually close this gap, but it is unclear what is specifically lacking. We will attempt to answer this through our experiments.

5.4 Active Exploration in Language Model Training (proposed)

5.4.1 Motivation

Lastly, we pivot from improving data use in the traditional training pipeline: although the general approach of breaking down complex tasks into elemental ones and strategically adding data may be applicable to many different tasks, there is still significant human effort involved in designing elemental tasks, noticing gaps, and retraining models. We have discussed how the typical pretrain to posttrain pipeline contrasts sharply with human learning, which involves continuous capability development, self-assessment, and self-directed learning throughout life. We ask whether we can improve models through a more self-directed learning process for use in more open-ended tasks, in which it might not be clear what data is necessary to for the model to train on.

Recent advances in self-play frameworks demonstrate that models can improve by generating their own training problems in closed domains with clear verification (arithmetic, coding) [44]. However, many consequential real-world tasks such as scientific reasoning, creative writing, and research ideation fall into semi-verifiable domains where feedback is subjective, sparse, or lacks binary correctness signals. These domains require models to develop sophisticated self-assessment capabilities and adjust their learning based on their self-assessment.

We propose extending current self-improvement paradigms to open-ended reasoning domains by developing frameworks for use cases such as scientific ideation and creative writing. Building on reinforcement pre-training approaches [], we investigate how models can: (1) engage in exploratory hypothesis formation during generation rather than simply predicting likely continuations, (2) develop internal reward mechanisms for evaluating reasoning quality in semi-verifiable domains, and (3) autonomously refine their reasoning capabilities through iterative self-assessment and targeted skill development.

This approach addresses a critical gap: while existing methods excel in domains with executable verification (mathematics, coding), this does not yet generalize to the complex reasoning patterns required for scientific discovery, creative problem-solving, and open-ended intellectual tasks that define much of human cognitive activity.

5.4.2 Research Questions

- **Reinforcement Pre-training for Open-Ended Generation:** How can we extend reinforcement pre-training beyond next-token prediction to encourage exploratory hypothesis formation and deeper deliberation in scientific reasoning domains?
- **Graded Reward Systems:** How can we develop reward mechanisms that capture partial correctness, logical coherence, and originality rather than binary ground-truth matching, enabling learning in semi-verifiable domains?

- **Autonomous Reward Model Induction:** To what extent can models autonomously discover latent evaluative principles (e.g., “claims must be supported by evidence”) from contrastive examples and distill these into trainable reward models?

5.4.3 Proposed Approach

This proposal is more speculative as we anticipate working on this project after the other proposed projects, but we plan to start by investigating whether models can learn to evaluate and improve their own reasoning in open-ended domains.

We will follow the general setup of [65], but targeted to open-ended domains. As such, we plan to use data from domains such as scientific papers from sources such as arXiv or PeS2o [253], and creative writing from sources like published fiction and creative writing datasets. For the first objective, we propose to assign graded rewards reflecting partial correctness or logical coherence, rather than based on matching the ground truth exactly. We will also explore ways to reward other qualities such as originality, as well as applying rewards on multiple levels of granularity (for instance, at the sentence level as well as at the individual token level).

For training reward models for open-ended tasks, we need granular rewards that are nonetheless based on clear concepts such as idea novelty or logical flow through ideas. We propose to have models such as the generative verifiers induce latent evaluative principles based on positive and negative examples (for instance, amateur vs. professional writing or accepted vs. rejected grant submissions), following approaches like SALMON which demonstrates how instructable reward models can generate scores based on arbitrary human-defined principles using synthetic preference data [263]. Such latent rules could look like “each claim must be supported by previously introduced evidence”, or “avoid extensive exposition when writing fictional stories”. These principles can then be distilled into separate reward models which can be combined for fine-grained feedback.

Notably, both aims can be combined in a loop: the generative verifier used in reinforcement pre-training can be used to mine heuristics which can be distilled into reward models, similar to the approaches where RL-trained generative verifiers are trained in an interleaved manner with the generator [309]. Overall, this allows us to iteratively refine model behaviour in open-ended domains, based on principled feedback, but without having to collect preference data from humans explicitly.

5.5 Conclusion

In this chapter, we demonstrate how breaking down more complex tasks into simpler forms and designing training interventions can translate into real-world improvements. The proposed work on repository-level coding and adaptive exploration represent applications of the same principle: we can systematically decompose complex tasks into simpler ones and create targeted training interventions for tasks that are not easily learned through pretraining.

CHAPTER 6

CONCLUSION

Note: This section will be expanded in the final dissertation document.

In this proposal, we evaluate capabilities that enable flexible reasoning in humans in Chapter 3, identifying figurative language understanding, compositionality, and the relationship between different forms of inference. Next, in Chapter 4 we investigate how models develop specific capabilities during pretraining, showing that data composition and training stage design, particularly midtraining, can strongly influence downstream performance, and proposing “elemental skills” as a unit for tracking and accelerating development. Finally, in Chapter 5 we apply these insights to guide targeted interventions in machine translation, code generation, and creative reasoning.

APPENDIX A

APPENDIX FOR CHAPTER 3.2

A.1 Crowdsourcing Details

We crowdsource metaphorical expressions and their interpretations through Amazon Mechanical Turk. Workers were recruited from the United States and were limited to those who had a $> 98\%$ approval rating on the platform, and who had also completed more than 1000 Human Intelligence Tasks (HITs). Data collection was split into two stages: in the first stage, 1458 train examples, and all the dev and test examples were collected. In the second stage, the remaining 6558 training examples were collected. We identified some workers who created especially good examples in the first stage, and recruited them back for more examples in the second stage. Workers were paid \$0.33 for each pair of sentences and were asked to generate 3 pairs at a time. An author of this paper wrote an initial pilot set of sentences, and timed themselves while writing some sentences. They found that each pair took around 1 minute to write, though this varied (less creative examples took less time, while more creative examples took more time). This extrapolates to an hourly rate of 19.80 USD, which is above the minimum wage in all US states, where workers were located.

Our HIT task was structured as follows: At the top of the page, the workers are shown the following instructions: "Your task is to generate three pairs of sentences with opposite or very different meanings, both of which contain rare/creative metaphors, which means metaphors that would not appear often in text on the Internet, books, social media, or news sites, but that can still be easily understood by people. For each metaphor, you should also provide a literal (non-metaphorical) sentence with the same meaning." Then, we display one example of a valid sentence pair. There is a button that opens a modal with more detailed instructions and some more valid/invalid examples for reference. Below that, we display three random words, which workers are encouraged to use in their sentences if they get stuck. Finally, we display three sets of 5 text fields for workers to fill in: one for the start phrase, two for each metaphorical phrase, and two for each literal interpretation. As the user types in each start phrase, we prepend a copy

of their phrase before the corresponding metaphor fields in the UI using some embedded JavaScript, which we found helped reduce confusion and resulted in less improperly formatted responses.

We launched many batches of these HITs until we had collected the desired quantity of data. Then, we converted the form responses into sentence pairs and validated each pair by hand before adding it to our dataset.

A.2 Invalid Examples

Figurative language examples collected from crowdworkers were excluded if they (a) did not make sense given the meaning and metaphorical expression, (b) had grammar or spelling errors that rendered them unintelligible, or (c) did not follow the format specified by the task template.

Examples are given below:

1. Do not make sense given the meaning and the metaphorical expression

Paired sentences	Possible answers
He was resourceful like toilet paper He was resourceful like a mess	He was very resourceful. He wasn't resourceful at all
The night was as long as a spool of thread The night was as long as a winding road	The night is long The night dragged on
the concert of the lesson is a main and a major the concert of the lesson features	we concert everyone we concert our loved one

Table A.1: Examples that were rejected due to being nonsensical.

2. Significant grammar or spelling errors

Paired sentences	Possible answers
fallten data are very much trusted fallten data are very valuable	fallten are nice flatten are safe
CAR IS BIRD FEATHEAR CAR IS COTTON	CAR SITE IS ROUGH CAR SITE IS HARD
Inflation is as natural as Minnesota rainfall in June Inflation is as natural as Minnesota snowfall in June	Inflation is perfectly natural Patient is in a natural result of other things

Table A.2: Examples that were rejected due to having significant spelling or grammar errors.

3. Do not follow format

Paired sentences	Possible answers
This attack is as weak as a feather This attack is as weak as a breeze	The attack is useless The attack doesn't work
My car motor is dusty like old cave My car motor is dusty like abandon building	Car motor is very rusty car motor is very dusty
the writer is stuck between a rock And another hard place the writer is stuck between a rock And a pebble	He is just stuck doesnt have a choice The writer can get over the pebble

Table A.3: Examples that were rejected due to not following the specified format.

Efforts were made to ensure that the final dataset contains no offensive content or personally identifiable information. WorkerID and other potentially personally identifying information were not included.

A.3 Backward accuracies

Model	Zero-shot	Fine-tuned (L)
GPT-2	52.18	52.00
GPT-neo 1.3B	54.36	63.44
GPT-3 Curie	58.46	74.83

Table A.4: Zero-shot and finetuned backward auto-regressive model accuracies on the test set

A.4 Paired accuracies

Model	Accuracy (pairs correct)
GPT-2 zero-shot	6.63
GPT-2 finetuned	5.06
GPT-neo zero-shot	10.3
GPT-neo finetuned	10.3
GPT-3 Curie zero-shot	17.4
GPT-3 Curie finetuned	50.0
BERT finetuned	70.6
RoBERTa finetuned	80.4
Human	89.7

Table A.5: Accuracy for models on the test set, counted in terms of pairs of sentences in which both are correct (%). Results are from one run.

A.5 Accuracy breakdown by Part-of-Speech

A.5.1 Subject

Part of speech	Accuracy	Frequency
NN	0.8569	538
PRP	0.8526	156
PRP\$ NN	0.9	110
NN NN	0.8889	63
DT NN	0.8182	44
NN NN NN	0.9375	32
JJ NN	0.9167	12

Table A.6: Accuracy breakdown and frequency of parts of speech in metaphor subjects. Only part-of-speech patterns with greater than 10 occurrences are shown.

A.5.2 Relation

Part of speech	Accuracy	Frequency
VBZ NN IN	0.8421	152
VBD RB JJ IN	0.8904	146
VBZ RB JJ IN	0.8889	99
VBZ	0.8352	91
VBD NN IN	0.8806	67
VBD	0.9180	61
VBN IN	0.9545	22
NN IN	0.8636	22
VBD JJ IN	0.9048	21
NNS IN	0.8889	18
VBD IN	0.8462	13
VBZ IN	1.0	13
VBD RB VBN IN	0.8182	11

Table A.7: Accuracy breakdown and frequency of parts of speech in metaphor relations. Only part-of-speech patterns with greater than 10 occurrences are shown.

A.5.3 Object

Part of speech	Accuracy	Frequency
NN	0.8788	429
NN NN	0.8992	129
JJ NN	0.8352	91
NN IN NN	0.8372	43
JJ NN NN	0.8710	31
NN NN NN	0.9130	23
VBG NN	0.9545	22
NN IN JJ NN	0.6154	13
PRP\$ NN	1.0	11
JJ	0.6364	11
NN IN NN NN	0.8182	11

Table A.8: Accuracy breakdown and frequency of parts of speech in metaphor objects. Only part-of-speech patterns with greater than 10 occurrences are shown.

A.6 Accuracy breakdown by hypernyms

A.6.1 Subject

Synset	Accuracy	Frequency
adult.n.01	0.8736	182
male.n.02	0.8684	152
woman.n.01	0.7391	46
female.n.02	0.9130	46
show.n.03	0.875	24
product.n.02	0.8636	22
motor_vehicle.n.01	0.9048	21
activity.n.01	0.8421	19
emotion.n.01	0.6667	18
publication.n.01	0.8333	18
feline.n.01	0.9375	16
being.n.01	0.7143	14
performer.n.01	0.8333	12
canine.n.02		12
body_covering.n.01	0.8333	12
vessel.n.03	0.8333	12
sound.n.01	1.0	12
domestic_animal.n.01	0.9167	12
person.n.01	0.8	10
scheme.n.01	0.9	10
contestant.n.01	1.0	10

Table A.9: Accuracy breakdown and frequency of WordNet hypernyms in metaphor subjects. Only hypernyms with 10 or greater occurrences are shown.

A.6.2 Object

Synset	Accuracy	Frequency
time_period.n.01	0.85	20
natural_object.n.01	0.8947	19
person.n.01	0.8824	17
large_integer.n.01	0.9286	14
adult.n.01	1.0	14
solid.n.01	0.9167	13
male.n.02	1.0	13
child.n.02	0.8333	12
body_of_water.n.01	0.75	12
body_covering.n.01	0.8333	12
digit.n.01	0.9167	12
region.n.01	0.8182	11
beverage.n.01	0.8182	11
juvenile.n.01	0.8182	11
container.n.01	0.9	10
rodent.n.01	0.9	10
feline.n.01	0.7	10
building.n.01	0.8	10
time_unit.n.01	0.9	10
travel.v.01	0.7	10

Table A.10: Accuracy breakdown and frequency of WordNet hypernyms in metaphor objects. Only hypernyms with 10 or greater occurrences are shown.

Startphrase	Completion	Label
His temper was a marshmallow. That is to say,	he was very easily angered, and once he was, he was very hard to cool down	Incorrect
He's got the swimming ability of a pack of dolphins. That is to say,	he's got none	Incorrect
The villain is as beautiful as the Joker. That is to say,	he's not	Correct
The child has the energy of a sloth. That is to say,	he is lazy	Correct
The girl moved as fast as a turtle. That is to say,	she moved at a turtle's pace	Literal
The sadness of her death was a flea. That is to say,	It was a flea that was a sadness	Literal

Table A.11: Examples of completions generated by GPT-3 Davinci.

A.7 Generation examples

Generation examples can be found in [Table A.11](#).

APPENDIX B

APPENDIX FOR CHAPTER 3.3

B.1 Treebank dataset tree types

Due to space constraints, we only show the top 20 tree types. This can be found in [Table B.1](#).

Tree type	Count	Example
PP → IN NP	77716	((in) (american romance))
S → NP-SBJ VP	62948	((he) (said simultaneously, "i wish they were emeralds"))
NP → DT NN	40876	((the) (way))
NP → NP PP	35743	((the temporal organization) (of the dance))
S → NP-SBJ S <VP-.>	24467	((the partners) (said they already hold 15 % of all shares outstanding.))
VP → TO VP	21833	((to) (be the enemy))
PP-LOC → IN NP	18005	((in) (the marketplace))
NP → DT NP <JJ-NN>	14898	((a) (professional linguist))
VP → MD VP	13575	((could) (make up his mind))
VP → VB NP	11838	((evaluate) (the progress of therapy))
PP-TMP → IN NP	11032	((for) (almost a year))
PP-CLR → IN NP	10054	((from) (the most sympathetic angle))
NP → NNP NNP	9863	((honolulu) (harbor))
NP → JJ NNS	9477	((recent) (years))
VP → VBD VP	8356	((was) (salted))
SBAR → WHNP-1 S	8332	((what) (to look for))
SBAR → IN S	7848	((that) (it exceeds the company 's annual sales and its market capitalization
NP-SBJ → DT NN	7600	((the) (rebound))
S → NP-SBJ-1 VP	7486	(draperies) (could be designed to serve structural purposes)
NP → NP SBAR	7317	((the " culture shock ") (they might encounter in remote overseas posts))

Table B.1: Counts of the top 20 grammatical tree types found in the WSJ and Brown sections of the Penn Treebank, with some examples given.

B.2 Treebank dataset phrase lengths

B.3 Probe learning curves

Learning curves of the approximative probes (across 10 folds) are shown in [Figure B.2](#).

B.4 Length Correlation

The correlations of the phrase length (in words) and compositionality scores in Treebank are shown in [Table B.2](#).

Model and representation	Spearman ρ	p-val
BERT _{CLS}	-0.0700	0.0
RoBERTa _{CLS}	0.1659	0.0
DeBERTa _{CLS}	0.1166	0.0
BERT _{AVG}	0.7143	0.0
RoBERTa _{AVG}	0.7086	0.0
DeBERTa _{AVG}	0.7866	0.0

Table B.2: Spearman ρ correlation between phrase length (in words) and compositionality score in the treebank.

B.5 Idiom and matched pair examples

Examples of idioms and the phrases they were matched with can be found in [Table B.3](#).

Idiom	Matched phrase	Syntactic pattern	Log f
Devil’s advocate	Baker’s town	JJ/dep/2 NN/pobj/0	2
Act of darkness	Abandonment of institution	NN/dobj/0 IN/prep/1 NN/pobj/2	4
School of hard knocks	Field of social studies	NN/pobj/0 IN/prep/1 JJ/amod/4 NNS/pobj/2	6

Table B.3: Examples of idioms with their matched phrases, selected based on having the same syntactic pattern and most similar log frequency in the Syntactic Ngrams dataset. Examples depicted here have the same log frequency. Note that the frequency is based on the most common dependency and constituency pattern found in Syntactic NGrams.

B.6 Error ratio of probes

Model/representation	Probe	Mean err. ratio ($\downarrow$)
BERT _{CLS}	ADD	0.4668
	W1	0.7806
	W2	0.3903
	LIN	0.3940
	AFF	0.3908
	MLP	0.3830
RoBERTa _{CLS}	ADD	0.4152
	W1	0.7946
	W2	0.2980
	LIN	0.3063
	AFF	0.3013
	MLP	0.3065
DeBERTa _{CLS}	ADD	0.7577
	W1	0.4661
	W2	0.7090
	LIN	0.6777
	AFF	0.9373
	MLP	0.5856
GPT-2 _{last}	ADD	0.4668
	W1	0.7806
	W2	0.3903
	LIN	0.3940
	AFF	0.3908
	MLP	0.3830
BERT _{AVG}	ADD	0.3873
	W1	0.8060
	W2	0.2167
	LIN	0.2327
	AFF	0.2098
	MLP	0.2283
RoBERTa _{AVG}	ADD	0.4504
	W1	0.8422
	W2	0.2431
	LIN	0.2471
	AFF	0.2095
	MLP	0.2181
DeBERTa _{AVG}	ADD	0.4472
	W1	0.8886
	W2	0.3202
	LIN	0.3143
	AFF	0.3044
	MLP	0.2952
GPT-2 _{AVG}	ADD	0.5013
	W1	0.9074
	W2	0.4226
	LIN	0.4041

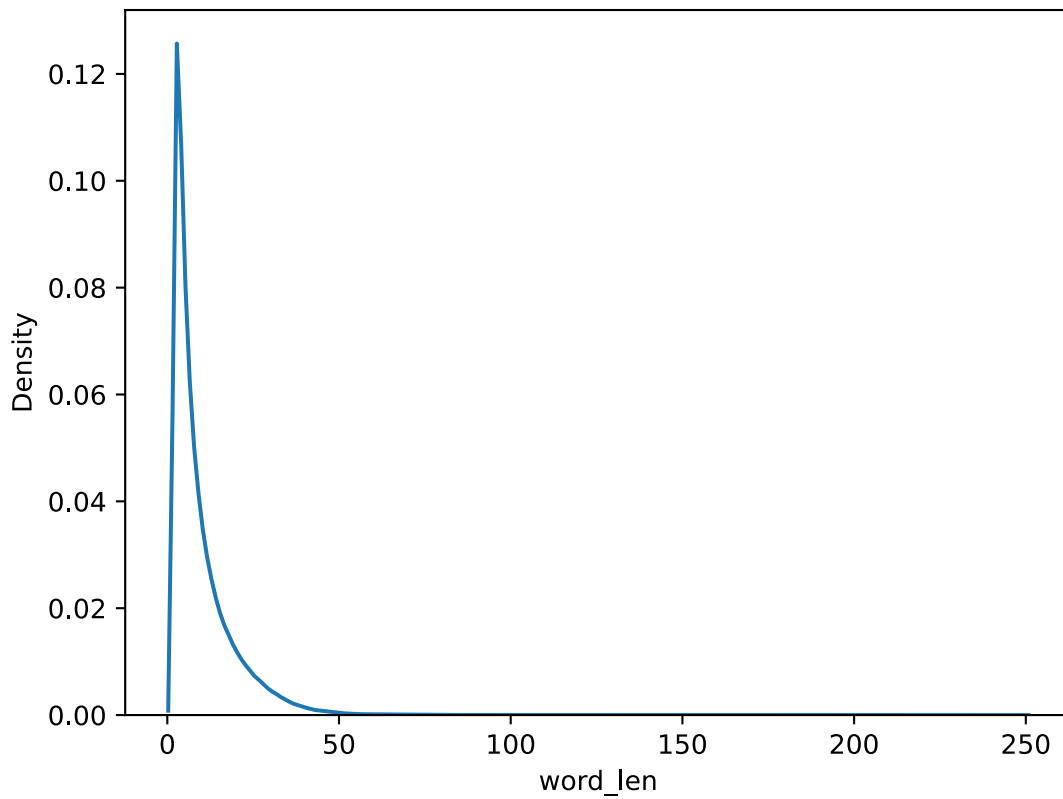


Figure B.1: Length distribution of phrases mined from the treebank, in number of words. The modal length was 3 words, followed closely by 2 words. Few phrases contained more than 50 words.

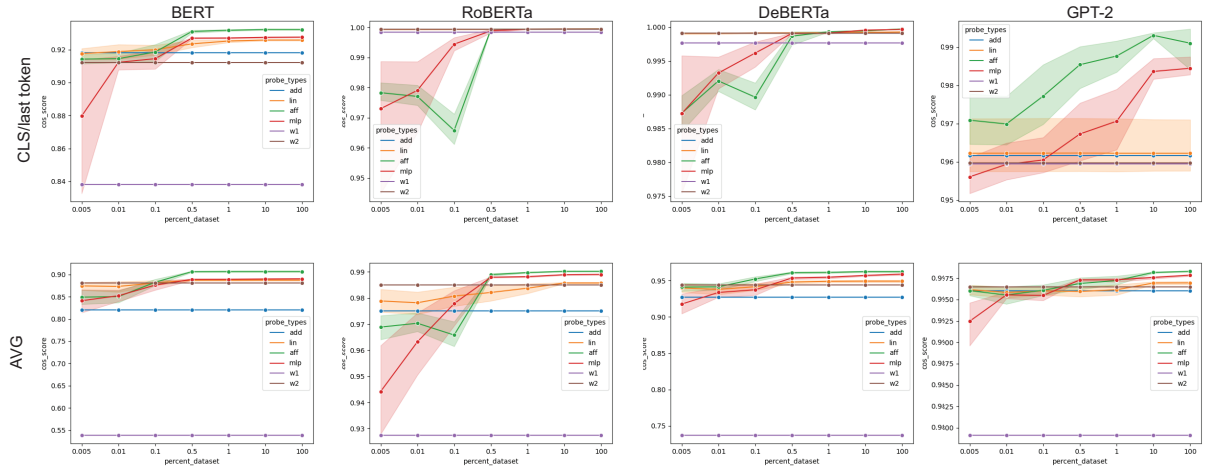


Figure B.2: Learning curves of approximative probes trained on differing percentages of train data.

B.7 Annotation setup and instructions

Annotators were recruited from a population of graduate students. Initially, 6 annotators completed the pilot experiment, which consisted of 101 examples. The subset of three annotators with highest agreement was asked if they would like to complete the full study. One annotator in the highest-agreement group could not continue to the full study, so this annotator was excluded, and the next group with highest agreement was chosen. The agreement values in § 3.3.6 are for the final group of annotators chosen.

The experiment was implemented on the Qualtrics platform, and participants were first presented with a consent form, linking to more background information on the study, and informing them that their participation was entirely voluntary. After agreeing to the terms, participants were shown some examples and went through 3 practice questions. The example given are shown in Figure B.3, and the annotation interface is shown in Figure B.4 and Figure B.5. After completing the practice section, annotators began annotating the real examples, which followed the same interface as the practice examples.

Annotators were all located in the United States, paid approximately \$15 per hour for their work.

Examples

The following examples illustrate some examples of compositionality. Compositionality means that you can understand the meaning of a phrase from its parts.

Ivory tower - This means that someone or something is out of touch with ordinary people. It doesn't mean "a white tower", and you wouldn't know what this means unless you came across it before, so it should be marked as **non-compositional**.

Balance sheet - This means a spreadsheet that someone calculates ("balances") their finances on. Its meaning can be inferred once you know what it is, but it might not be obvious right away, so it should be marked as **somewhat compositional**.

Brown dog - This is a dog which is brown. You can fully figure out the meaning just from the two words, so it is **fully compositional**.



Figure B.3: Examples of compositionality judgments shown to annotators

Consider the following phrase:

Raining cats and dogs

If the phrase has both a literal and idiomatic meaning, please consider the idiomatic meaning. E.g. “raining cats and dogs” could mean literal cats and dogs falling out of the sky, but you should consider the usual meaning. Some of the phrases are quite rare, so you should search up the meaning if you don’t know it.

Please consider the two parts of this phrase individually:

Raining

Cats and dogs

Consider the most typical meaning of the two parts of the phrase.

How well can you understand the phrase by combining **the most typical meaning** of the two parts of the phrase?

1. Not at all - you cannot understand the phrase from the typical meanings of its two parts.
2. Somewhat - you can understand the phrase somewhat. You may have to guess its meaning, or one of its parts is used in an atypical or figurative way.
3. Fully - you can completely understand the phrase by understanding the typical meanings of its two parts.

1 - Not compositional

2 - Somewhat compositional

3 - Fully compositional

Figure B.4: First page of annotation interface for a practice phrase

Knowing the final meaning of the phrase **Raining cats and dogs**, how much do you think **Raining** contributes to the final meaning?

1 - Not at all

2 - Somewhat

3 - A great deal or fully

Knowing the final meaning of the phrase **Raining cats and dogs**, how much do you think **Cats and dogs** contributes to the final meaning?

1 - Not at all

2 - Somewhat

3 - A great deal or fully

Figure B.5: Second page of annotation interface for a practice phrase

B.8 Compositionality scores without anisotropy correction

The raw compositionality scores can be found in [Table B.5](#).

Model and representation	Probe	Mean reconstruction score	Standard dev.
BERT _{CLS}	ADD	0.9178	0.001159
	W1	0.8382	0.003599
	W2	0.9117	0.0007133
	LIN	0.9258	0.0002285
	AFF	0.9322	0.0002033
	MLP	0.9276	0.0002108
RoBERTa _{CLS}	ADD	0.99935	3.895×10^{-6}
	W1	0.99850	2.612×10^{-5}
	W2	0.99937	6.866×10^{-6}
	LIN	0.99946	4.735×10^{-6}
	AFF	0.99950	6.093×10^{-6}
	MLP	0.99947	4.719×10^{-6}
DeBERTa _{CLS}	ADD	0.99908	4.070×10^{-5}
	W1	0.99762	2.900×10^{-5}
	W2	0.99911	1.399×10^{-4}
	LIN	0.99928	8.963×10^{-5}
	AFF	0.99972	1.542×10^{-5}
	MLP	0.99965	2.323×10^{-5}
BERT _{AVG}	ADD	0.8205	0.0003836
	W1	0.5383	0.007471
	W2	0.8893	0.03071
	LIN	0.8873	0.003071
	AFF	0.9069	0.002566
	MLP	0.8904	0.002988
RoBERTa _{AVG}	ADD	0.9752	0.0001306
	W1	0.9274	0.001695
	W2	0.9850	0.0005092
	LIN	0.9858	0.0004573
	AFF	0.9902	0.0003076
	MLP	0.9890	0.0003981
DeBERTa _{AVG}	ADD	0.9275	0.002634
	W1	0.7368	0.001575
	W2	0.9438	0.003321
	LIN	0.9493	0.003036
	AFF	0.9625	0.001814
	MLP	0.9590	0.002145
GPT-2 _{AVG}	ADD	0.9960	0.0002833
	W1	0.9791	0.0001214
	W2	0.9965	0.0003359
	LIN	0.9970	0.0003036
	AFF	0.9984	0.0002617
	MLP	0.9979	0.0001634

Table B.5: Mean reconstruction score (cosine similarity) and standard deviation of each approximative probe across 10 folds. Not corrected for anisotropy in each representation/model type.

B.9 AUC of approximative probes

Model and representation	probe	AUC
BERT _{CLS}	ADD	91.80
	W1	83.82
	W2	91.20
	LIN	92.57
	AFF	93.20
	MLP	92.74
RoBERTa _{CLS}	ADD	99.93
	W1	99.84
	W2	99.93
	LIN	99.94
	AFF	99.93
	MLP	99.94
DeBERTa _{CLS}	ADD	99.90
	W1	99.75
	W2	99.90
	LIN	99.92
	AFF	99.94
	MLP	99.95
GPT-2 _{last}	MLP	99.95
	ADD	96.16
	W1	95.94
	W2	95.97
	LIN	96.21
	AFF	99.18
BERT _{AVG}	MLP	98.32
	ADD	82.04
	W1	53.83
	W2	88.10
	LIN	88.68
	AFF	90.63
RoBERTa _{AVG}	MLP	88.96
	ADD	97.51
	W1	92.73
	W2	98.49
	LIN	98.56
	AFF	99.00
DeBERTa _{AVG}	MLP	98.88
	ADD	92.74
	W1	73.67
	W2	94.38
	LIN	94.89
	AFF	96.21
GPT-2 _{AVG}	MLP	95.75
	ADD	99.60
	W1	97.90
	W2	99.64

B.10 Mean deviation of phrase types by tree type

The mean deviation of the most common tree types can be found in [Figure B.6](#).

B.11 Further named entity results

Named entity results can be found in [Figure B.7](#) and [Figure B.8](#).

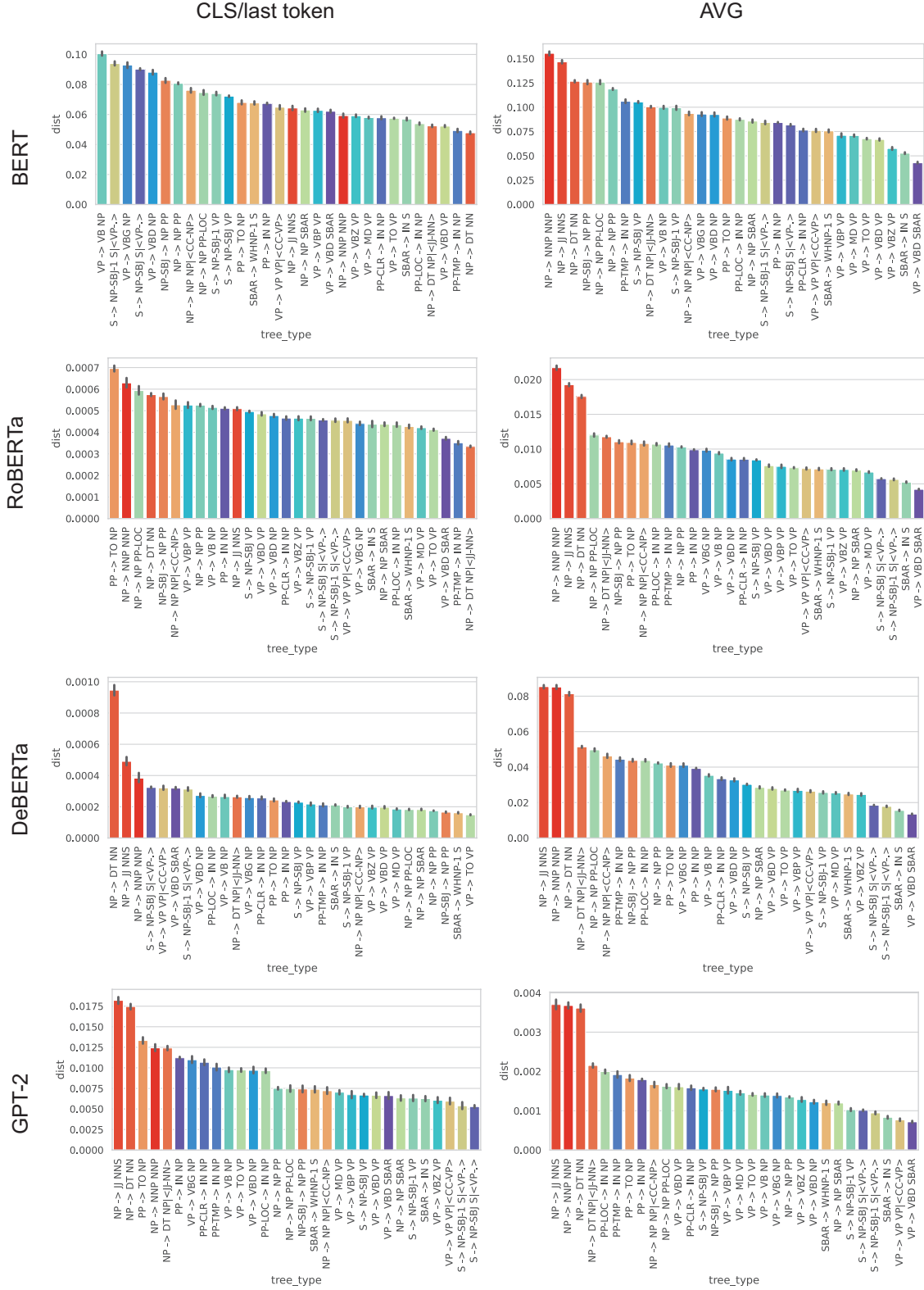


Figure B.6: Mean deviation from predicted representation across full tree types.

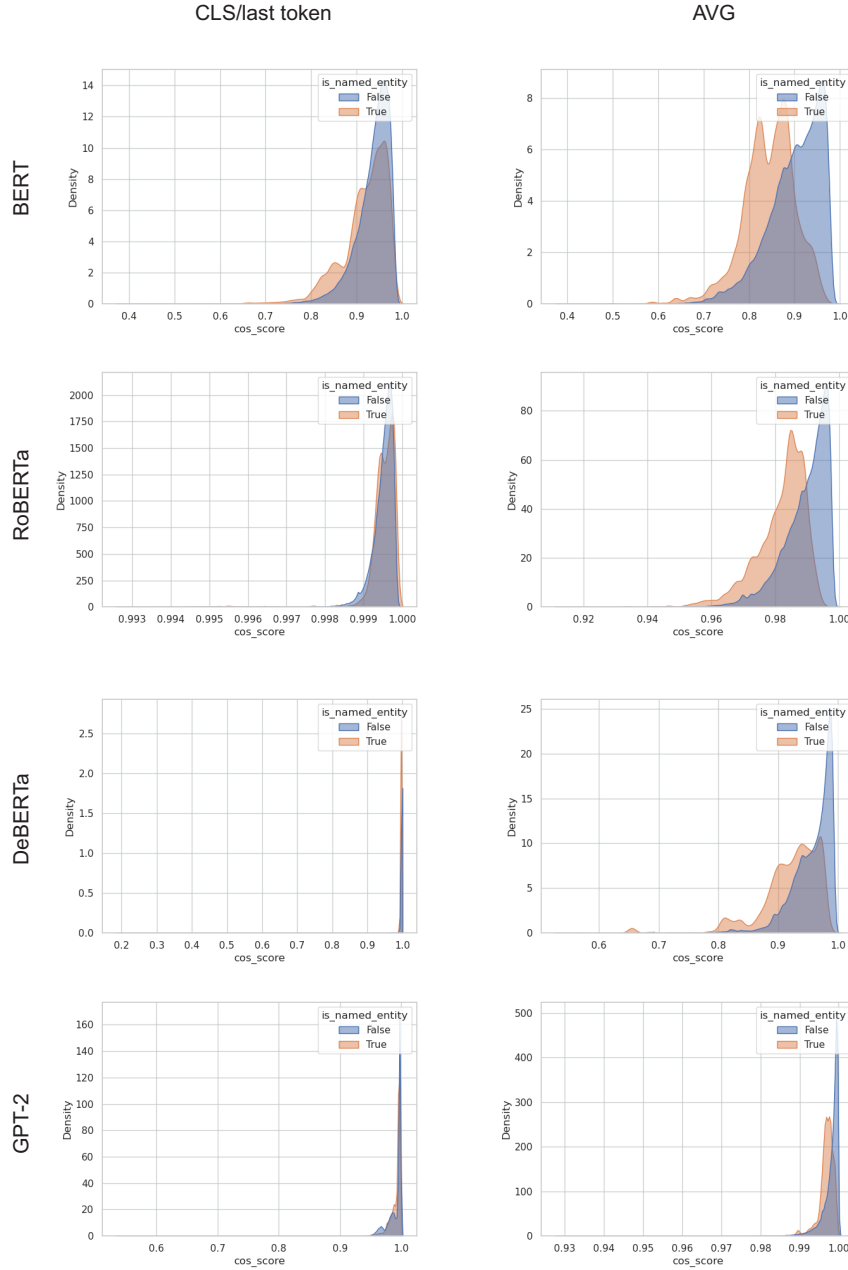


Figure B.7: Distributions of compositionality score for named entities and non-named entities across model types and representation types. The AVG representation matches the intuition that named entities are usually less semantically compositional, as they point to an entity in the real world that may not relate to their name.

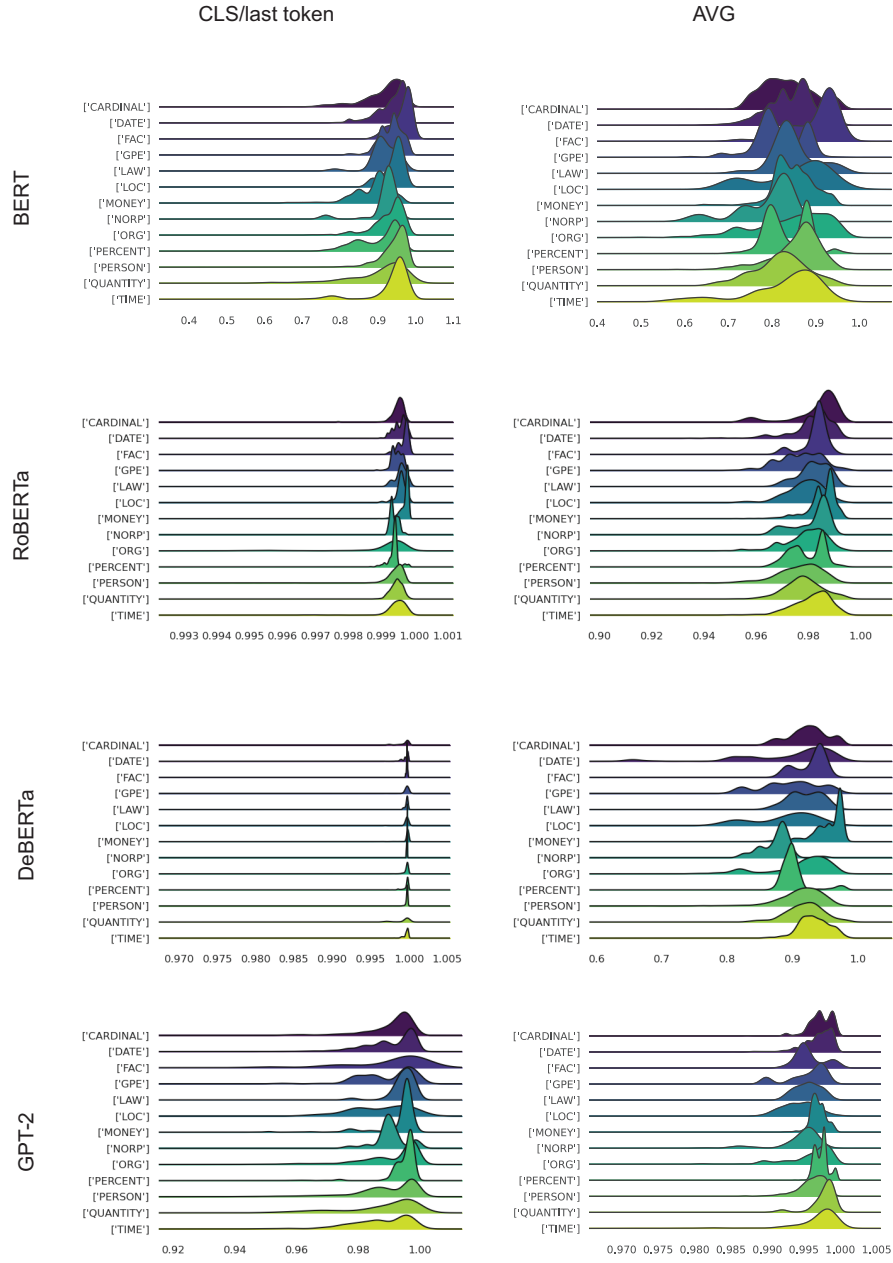


Figure B.8: Visualization of distribution of compositionality scores across different types of named entities.

B.12 Frequency and length correlations

Model and representation	Feature	Spearman ρ	p-val
BERT _{CLS}	Word length	0.2182	$3.055 \times 10^{-10*}$
BERT _{AVG}		0.007396	0.08722
RoBERTa _{CLS}		0.01686	0.6193
RoBERTa _{AVG}		0.3653	$4.773 \times 10^{-28*}$
DeBERTa _{CLS}		0.4087	$1.709 \times 10^{-35*}$
DeBERTa _{AVG}		0.4484	$1.340 \times 10^{-42*}$
GPT-2 _{last}		0.3228	$8.481 \times 10^{-22*}$
GPT-2 _{AVG}		0.03125	$1.719 \times 10^{-20*}$
Human	Word length	0.05666	0.1894
BERT _{CLS}	Frequency	0.2182	0.08193
BERT _{AVG}		-0.08582	0.07899
RoBERTa _{CLS}		0.02548	0.9053
RoBERTa _{AVG}		-0.08354	0.08193
DeBERTa _{CLS}		-0.1265	0.001459*
DeBERTa _{AVG}		-0.2185	$6.455 \times 10^{-10*}$
GPT-2 _{last}		-0.05750	0.3595
GPT-2 _{AVG}		0.04382	0.5891
Human	Frequency	0.008363	0.9053

Table B.7: Correlations of frequency and length with human and model compositionality scores. Corrected with Holm-Bonferroni correction.

APPENDIX C

APPENDIX FOR CHAPTER 3.4

C.1 Model Inference Settings

Generation settings, as well as other details, used in the three domains are detailed here. Each model was tested 6 times for each setting (*few-shot*, *zs-cot*...) at different temperatures, and the aggregate results are shown in figures and tables throughout the paper. The exact model versions used for gpt models were gpt-3.5-turbo-0613 and gpt-4-1106-preview.

Functions When answering questions in the base task, gpt models were tested 3 times each at temperature $T = \{0, 1\}$ (note that $T = 0$ is nondeterministic in gpt models because of hardware). No max number of tokens was set for the generation. For llama models, $T = \{0.1, 1\}$ was used, also with no max number of tokens.

When generating hypotheses, we always used the same model as performed the base task. That is to say, gpt-3.5-turbo would generate hypotheses used by gpt-3.5-turbo, and so on. Hypotheses were always generated with a temperature above 0 to encourage generation of diverse hypotheses. For gpt models, hypotheses were generated with $T = 1$ and for llama models, with $T = 1.0625$. The specific value for llama models was because llama would usually generate the same hypothesis 5 times at $T = 1$, but higher values greatly increased the number of nonsensical and badly formatted hypotheses.

When self-evaluating hypotheses, the verbalized confidence score was generated at $T = 0$. When using the log-probabilities from text-davinci-002 to rerank hypotheses, the model was also set to $T = 0$.

Colours As in the functions domain, gpt models and llama models were respectively tested 3 times each at temperature $T = \{0, 1\}$ and $T = \{0.1, 1\}$ with no max number of tokens.

When generating hypotheses, $T = 1$ was used for all models.

When self-evaluating hypotheses, settings were the same as in the functions domain.

Kalamang For the base translation task, the same settings were used for generation as in Tanzer et al. [267]. Due to cost constraints, we ran only once in each translation

direction on each setting with $T = 0.05$.

Vocabulary hypotheses for all models were generated with $T = 1$. Grammar feature hypotheses were generated independently of translations, and the first non-null hypothesis was chosen due to cost constraints. $T = 0.7$ was used for gpt grammar hypotheses, and $T = 1$ was used for llama grammar hypotheses.

When self-evaluating hypotheses, $T = 0$ was once again used for all models.

C.2 Prompts for Linear Functions Domain

Table C.1: Prompts for the functions domain. Newlines are depicted visually for ease of reading. Variables substituted per question are marked like {this}. The wording for the “prompt with self-induced hypothesis” and “zero-shot chain-of-thought prompt” was slightly changed from the few-shot examples prompt to reduce formatting errors.

Prompt Type	Usage	Prompt Text
Base system prompt	For reasoning with in-context examples	You are a problem solving system. Your job is to use the input-output pairs to solve the problem as well as you can.
Hypothesis proposal system prompt	For proposing hypotheses based on in-context examples	You are a pattern recognition system. Your job is to come up with a function that describes the data as well as you can.
Instruction following system prompt	For applying a proposed hypothesis or ground-truth hypothesis to the input	You are a problem solving system. Your job is to apply the function to the data in order to produce an answer.
Few-shot examples prompt	For reasoning with in-context examples only	Return the output preceded by 'Output:' Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Input: {query input}
Prompt with ground-truth hypothesis	Used when prompting the model to directly apply the correct hypothesis to the input. In-context examples are also included.	Function: {The real function} Examples: Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Return the output preceded by 'Output:' Input: {query input}

Table C.1: Prompts for the functions domain. Newlines are depicted visually for ease of reading. Variables substituted per question are marked like {this}. The wording for the “prompt with self-induced hypothesis” and “zero-shot chain-of-thought prompt” was slightly changed from the few-shot examples prompt to reduce formatting errors.

Prompt Type	Usage	Prompt Text
Prompt for hypothesis induction	Have the model propose a single hypothesis for the function based on in-context examples	Write the function that captures the relationship between inputs and outputs. You should write it in the form $y = a x^0 + b x^1$. Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Function (please write explicitly in the exact form 'Output: $y = a x^0 + b x^1$')
Prompt with a self-induced hypothesis	Same as ground-truth version, but with a self-generated hypothesis; wording adjusted to reduce formatting errors	Use this function to apply to the input example to get the correct output. {model's hypothesis} However, just write the output like what's shown in these examples. Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Return the output preceded by 'Output:' Input: {query input}

Table C.1: Prompts for the functions domain. Newlines are depicted visually for ease of reading. Variables substituted per question are marked like {this}. The wording for the “prompt with self-induced hypothesis” and “zero-shot chain-of-thought prompt” was slightly changed from the few-shot examples prompt to reduce formatting errors.

Prompt Type	Usage	Prompt Text
Prompt for zero-shot chain-of-thought	Encourage step-by-step reasoning	Return the output preceded by 'Final Output:' Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Let's think step by step about what the function could be. Remember to write down 'Final Output:' before your final answer. Input: {query input}
Prompt for hypothesis probability estimate	Prompt a language model to directly score hypotheses for reranking	How likely is this hypothesis about the function to be true given the data? Examples: Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Function explanation: {model's hypothesis} Please give a probability between 0 and 1 inclusive, and only answer with a number. Probability:

Table C.1: Prompts for the functions domain. Newlines are depicted visually for ease of reading. Variables substituted per question are marked like {this}. The wording for the “prompt with self-induced hypothesis” and “zero-shot chain-of-thought prompt” was slightly changed from the few-shot examples prompt to reduce formatting errors.

Prompt Type	Usage	Prompt Text
Prompt for data logprobs given hyp estimate	Rerank hypotheses using log-probabilities under the data	These are examples of applying this function: {model's hypothesis} Examples: Input: {input1} Output: {output1} Input: {input2} Output: {output2} ...

C.3 Prediction Fit of Other Models On Linear Functions

?? shows predictions made by each model when using a few-shot prompt, true instruction, or self-induced instruction.

C.4 Predicted Coefficients compared to Real Coefficients in Linear Functions

?? plots model hypotheses about the coefficients of x^0 and x^1 in a linear function against the actual coefficients

C.5 Grammar and Details of Colours Domain

The rules of the colours domain are expressed through production rules below. Models were found to respond better to verbal instructions than formal ones in initial testing (e.g. a verbal statement "repeat twice" rather than $[[x]]\text{bluf} \rightarrow [[x]][[x]]$)

Listing C.1: Grammar of the colours language, as presented to LMs.

```
lug -> blue
dax -> green
wif -> red
zup -> yellow
bluf -> repeat the last action twice
walm -> repeat the last action three times
```

Training and test data was automatically generated by generating sentences of up to 5 nonce words on the source side, with shorter sentences being more likely (the respective probabilities for sentence lengths from 1 to 5 are $[0.4, 0.3, 0.15, 0.1, 0.05]$). Each colour could also be repeated with the repeat actions, and whether repeat nonce words were inserted was also random, though skewed toward no repetition (the probabilities were respectively $[0.8, 0.1, 0.1]$ for no repeats, one repeat, or two repeats). A repeat term never followed another repeat term, and the same colour word never appeared twice consecutively to make it easier to learn the repeat terms.

Additionally, we provided a fixed set of few-shot examples that covered all the nonce terms:

Listing C.2: Few-shot examples in the colours translation task

```
Input: lug dax
Output: blue green

Input: wif zup
Output: red yellow
```

Input: lug bluf
Output: blue blue

Input: wif walm
Output: red red red

Input: lug walm dax bluf
Output: blue blue blue green green

C.6 Prompts for Colours Domain

Table C.2: Prompts for the colours domain. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The “prompt with a self-induced hypothesis” was slightly modified from the base prompt in order to encourage models to follow the correct formatting, while the “prompt for zero-shot chain-of-thought” was slightly modified to encourage models to generate a concrete chain of thought, and also to follow the correct formatting.

Prompt Type	Usage	Prompt Text
Base system prompt	For reasoning with in-context examples	You are a problem solving system. Your job is to use the input-output pairs to solve the problem as well as you can.'
Hypothesis proposal system prompt	For proposing hypotheses based on in-context examples	You are a rule induction system. Your job is to figure out the rules underlying a problem and report on them. Use the examples to guide your thinking.
Instruction following system prompt	For applying a proposed hypothesis or ground-truth hypothesis to the input	You are a parser. Carefully use the grammar to parse inputs to determine the correct output.
Few-shot examples prompt	For reasoning with in-context examples only	Return the output preceded by 'Output:' Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Input: {query input} Remember to start your answer with 'Output:'

Table C.2: Prompts for the colours domain. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The "prompt with a self-induced hypothesis" was slightly modified from the base prompt in order to encourage models to follow the correct formatting, while the "prompt for zero-shot chain-of-thought" was slightly modified to encourage models to generate a concrete chain of thought, and also to follow the correct formatting.

Prompt Type	Usage	Prompt Text
Prompt with ground-truth hypothesis	Used when prompting the model to directly apply the correct hypothesis to the input. In-context examples are also included.	Use this grammar to parse the input example to get the correct output. Grammar: {colours domain grammar} Examples: Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Return the output preceded by 'Output:' Input: {query input}
Prompt for hypothesis induction	Used to have the model propose a single hypothesis for the translation of a word.	The below examples contain the nonce word {word}. Using the examples, deduce what {word} means. Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Write your answer like this: {word} -> meaning. Meaning can be a word or a general rule dependent on the context. Rule:

Table C.2: Prompts for the colours domain. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The "prompt with a self-induced hypothesis" was slightly modified from the base prompt in order to encourage models to follow the correct formatting, while the "prompt for zero-shot chain-of-thought" was slightly modified to encourage models to generate a concrete chain of thought, and also to follow the correct formatting.

Prompt Type	Usage	Prompt Text
Prompt with a self-induced hypothesis	Used similarly to the "prompt with ground-truth hypothesis", except with a self-generated hypothesis. The wording is slightly changed.	Use this grammar to parse the input example to get the correct output. {model's hypothesis grammar} However, just write the output like what's shown in these examples. Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Return the output preceded by 'Output:' Input: {query input}
Prompt for zero-shot chain-of-thought	Used to encourage the model to generate a chain of thought.	Return the output preceded by 'Final Output:' Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Let's think step by step about what the translation could be. Work through your answer step by step and show your work. Remember to write down 'Final Output:' before your final answer. Input: {query input}

Table C.2: Prompts for the colours domain. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The "prompt with a self-induced hypothesis" was slightly modified from the base prompt in order to encourage models to follow the correct formatting, while the "prompt for zero-shot chain-of-thought" was slightly modified to encourage models to generate a concrete chain of thought, and also to follow the correct formatting.

Prompt Type	Usage	Prompt Text
Prompt for hypothesis probability estimate	Used to prompt a language model directly for reranking hypotheses	These are examples of the translation of the word {word}. Examples: Input: {input1} Output: {output1} Input: {input2} Output: {output2} ... Given these examples, how likely is this hypothesis about the meaning of {word}? {model's word hypothesis} Please give a probability between 0 and 1 inclusive, and only answer with a number. Probability:
Prompt for data logprobs given hyp estimate	Used to rerank hypotheses based on logprobs.	These are examples of applying this function: {model's word hypothesis} Examples: Input: {input1} Output: {output1} Input: {input2} Output: {output2} ...

C.7 chrF For Colours Domain

?? depicts the mean chrF score over 6 trials for each model in each setting. Reranking methods are averaged for instruction-induction.

C.8 Vocabulary Induction Accuracy for Colours Domain

We show in [Table C.3](#) the mean accuracy for each colour term for gpt models, computed over all selected (highest-ranked) word hypotheses. Point-biserial correlation between hypothesis correctness (in the instruction-induction case) and few-shot correctness is also shown. P-values are corrected with FDR in the trials for each model. We performed the correction per model rather than across all models because we were separately interested in the behaviour of each model. We parsed the production rule generated by the model in order to determine hypothesis correctness.

Table C.3: Vocabulary induction correctness and relation to few-shot translation correctness for gpt models. Significant correlations after correction for multiple comparisons are bolded.

Model	Word	method	Mean Correctness	Correlation	p-value	p-value corrected
gpt-3.5-turbo	lug	verbal_conf	0.604859335038363	-0.0370824643671235	0.300350893877687	0.52561406
gpt-3.5-turbo	dax	verbal_conf	0.233072916666666	0.0669078409779957	0.0638449077681366	0.29564626
gpt-3.5-turbo	wif	verbal_conf	0.329268292682926	0.0360864838621108	0.646415727402187	0.82271093
gpt-3.5-turbo	zup	verbal_conf	0.61	0.154599744147798	0.0288284171902121	0.20179892
gpt-3.5-turbo	bluf	verbal_conf	0	–	–	–
gpt-3.5-turbo	walm	verbal_conf	0	–	–	–
gpt-4-turbo	lug	verbal_conf	0.875763747454175	0.0439535251603629	0.168735841765082	3.58e-1
gpt-4-turbo	dax	verbal_conf	0.921487603305785	-0.0418339829001646	0.193444842357041	3.58e-1
gpt-4-turbo	wif	verbal_conf	0.865853658536585	0.0447789368805147	0.569118855838926	7.32e-1
gpt-4-turbo	zup	verbal_conf	0.87	0.0579317588377811	0.415170703498878	5.75e-1
gpt-4-turbo	bluf	verbal_conf	0.153846153846153	0.130289326661761	0.255554806693343	3.83e-1
gpt-4-turbo	walm	verbal_conf	0.171428571428571	0.0489476613672	0.687394764355678	7.55e-1
gpt-3.5-turbo	lug	p_data	0.28772378516624	0.0591580864392629	0.0983076746979401	0.29564626
gpt-3.5-turbo	dax	p_data	0.10546875	0.00996669750420649	0.782730731634431	0.84294079
gpt-3.5-turbo	wif	p_data	0.195121951219512	0.24388397206115	0.00164950240178991	0.02309303
gpt-3.5-turbo	zup	p_data	0.4	-0.114775306453679	0.105587948530317	0.29564626
gpt-3.5-turbo	bluf	p_data	0.0769230769230769	-0.1673596703488	0.143038746170574	0.33375707
gpt-3.5-turbo	walm	p_data	0	–	–	–
gpt-4-turbo	lug	p_data	0.731160896130346	-0.0803208788270973	0.0118067271484766	5.31e-2
gpt-4-turbo	dax	p_data	0.733471074380165	-0.0411584340901287	0.200743028038944	3.58e-1
gpt-4-turbo	wif	p_data	0.768292682926829	-0.0255095864281423	0.745759257358691	7.55e-1
gpt-4-turbo	zup	p_data	0.76	0.0221788843765235	0.755246389745868	7.55e-1
gpt-4-turbo	bluf	p_data	0.102564102564102	-0.140859042454752	0.218662613923779	3.58e-1
gpt-4-turbo	walm	p_data	0.142857142857142	-0.263591223386182	0.0274677998646683	9.89e-2
gpt-3.5-turbo	lug	p_answer	0.436061381074168	-0.0128717673177528	0.719304414221268	0.83918848
gpt-3.5-turbo	dax	p_answer	0.190104166666666	0.00399825404003988	0.9119154429019	0.91191544
gpt-3.5-turbo	wif	p_answer	0.353658536585365	-0.0743944887571608	0.343771340895407	0.53475542
gpt-3.5-turbo	zup	p_answer	0.59	-0.0879883939992145	0.21536453378319	0.43072907
gpt-3.5-turbo	bluf	p_answer	0.0512820512820512	-0.0551410966570355	0.631586163340137	0.82271093
gpt-3.5-turbo	walm	p_answer	0	–	–	–
gpt-4-turbo	lug	p_answer	0.70468431771894	-0.0661341939950676	0.0382590601605542	1.15e-1
gpt-4-turbo	dax	p_answer	0.712809917355371	0.188125007589795	3.67e-9	6.61e-8
gpt-4-turbo	wif	p_answer	0.817073170731707	-0.249727938065423	0.00126030696185871	1.13e-2
gpt-4-turbo	zup	p_answer	0.68	-0.142141439177509	0.0446656422362514	1.15e-1
gpt-4-turbo	bluf	p_answer	0.23076923076923	-0.294147299400922	0.00894893196851883	5.31e-2
gpt-4-turbo	walm	p_answer	0.0571428571428571	-0.0578005412233204	0.634584373040489	7.55e-1

C.9 MTOB Experimental Settings

We generally used the same experimental setup as in mtob. For the few-shot condition, we used two reference sentences for each word on the source side, selected via longest subsequence. For the true-instruction setting, we used reference sentences, in addition to the wordlist and true grammar sketch. For each word in the source sentence, the most similar word from the wordlist was also retrieved based on the longest common substring. In the instruction-inference settings, the model’s self-induced grammar sketch was substituted for the true grammar sketch, and the model was also prompted to first create hypotheses about the translation of each word in the source sentence.

At the time that we conducted experiments, the Kalamang to English training set was not available, so we created this training set by reversing the source and target in the English to Kalamang training set. This may have slightly harmed translations in this direction, as this caused the training set to be different from the one reported in the benchmark results.

C.10 MTOB Prompts

Table C.4: Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. “” indicates that the prompt is the same as in § C.6.

Prompt Type	Usage	Prompt Text
Base system prompt	For reasoning with in-context examples	“ ”
Hypothesis proposal system prompt (Grammar)	For proposing grammar feature hypotheses based on in-context examples	“ ”
Hypothesis proposal system prompt (Vocab)	For proposing vocabulary mappings based on in-context examples	“ ”
Instruction following system prompt	For applying a proposed hypothesis or ground-truth hypothesis to the input	“ ”

Table C.4: Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. "" indicates that the prompt is the same as in § C.6.

Prompt Type	Usage	Prompt Text
Few-shot examples prompt	For reasoning with in-context examples only	Human: Kalamang is a language spoken on the Karas Islands in West Papua. Translate the following sentence from English to Kalamang: {query input} To help with the translation, here is a translated sentence with words similar to '{word}' in a list of translated Kalamang-English reference sentences: English sentence: {eng sentence} Kalamang translation: {kgv sentence} <Repeated for all words> Now write the translation. English: {query input} Kalamang translation:

Table C.4: Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. "" indicates that the prompt is the same as in § C.6.

Prompt Type	Usage	Prompt Text
Prompt with ground-truth hypothesis	Used when prompting the model to directly apply the correct hypothesis to the input. In-context examples are also included.	Human: Kalamang is a language spoken on the Karas Islands in West Papua. Translate the following sentence from English to Kalamang: {query input} To help with the translation, here is one of the closest entries to '{word}' in the Kalamang=English bilingual dictionary: English word or phrase: '{word}' Kalamang translation: '{translation}' <Repeated for all words> To help with the translation, here is a translated sentence with words similar to '{word}' in a list of translated Kalamang-English reference sentences: English sentence: {eng sentence} Kalamang translation: {kgv sentence} <Repeated for all words> To help with the translation, here's a grammar sketch of Kalamang: {grammar sketch} Now write the translation. English: {query input} Kalamang translation:

Table C.4: Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. "" indicates that the prompt is the same as in § C.6.

Prompt Type	Usage	Prompt Text
Prompt for hypothesis induction	Used to have the model propose a single hypothesis for the translation of a word.	The following sentences contain the word `{word}` in English. Examples: English sentence: {eng sentence} Kalamang translation: {kgv sentence} ... What is the Kalamang translation of the word `{word}`? Write your answer like this: {English word} -> {Kalamang translation}.

Table C.4: Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. "" indicates that the prompt is the same as in § C.6.

Prompt Type	Usage	Prompt Text
Prompt with a self-induced hypothesis	Used similarly to the "prompt with ground-truth hypothesis", except with a self-generated hypothesis. The wording is slightly changed.	Human: Kalamang is a language spoken on the Karas Islands in West Papua. Translate the following sentence from English to Kalamang: {query input} To help with the translation, here is one of the closest entries to '{word}' in the Kalamang=English bilingual dictionary: English word or phrase: '{word}' Kalamang translation: '{hypothesis translation}' <Repeated for all words> To help with the translation, here is a translated sentence with words similar to '{word}' in a list of translated Kalamang-English reference sentences: English sentence: {eng sentence} Kalamang translation: {kgv sentence} <Repeated for all words> To help with the translation, here's a grammar sketch of Kalamang: {hypothesis grammar sketch} Now write the translation. English: {query input} Kalamang translation:

Table C.4: Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. "" indicates that the prompt is the same as in § C.6.

Prompt Type	Usage	Prompt Text
Prompt for zero-shot chain-of-thought	Used to encourage the model to generate a chain of thought.	Human: Kalamang is a language spoken on the Karas Islands in West Papua. Translate the following sentence from English to Kalamang: {query input} To help with the translation, here is a translated sentence with words similar to '{word}' in a list of translated Kalamang-English reference sentences: English sentence: {eng sentence} Kalamang translation: {kgv sentence} <Repeated for all words> Let's think step by step about what the translation could be. Now write the translation. English: {query input} Kalamang translation:
Prompt for hypothesis probability estimate	Used to prompt a language model directly for reranking hypotheses	These are examples of the translation of the word {word}. Examples: {example translation pairs} ... Given these examples, how likely is this hypothesis about the meaning of {word}? {model's word hypothesis} Please give a probability between 0 and 1 inclusive, and only answer with a number. Probability:

Table C.4: Prompts for Kalamang translation with MTOB. We follow the same prompt formatting as in Tanzer et al. [267], with the exception of replacing retrieved grammar book passages with the grammar sketch in § C.11. Newlines are depicted visually for ease of reading. Variables that are substituted depending on the question are marked like {this}. The English to Kalamang direction is depicted in the examples. "" indicates that the prompt is the same as in § C.6.

Prompt Type	Usage	Prompt Text
Prompt for data logprobs given hyp estimate	Used to rerank hypotheses based on logprobs.	These are examples of sentences that contain the word {word}: This is the translation of the word: {model's word hypothesis} Examples: {example translation pairs}

C.11 Kalamang Grammar Sketch

Listing C.3: Grammar sketch for Kalamang, compiled from WALS and Grambank features.

```

=== Start of grammar sketch ===
Basic Word Order: SV (Subject-Verb), OV (Object-Verb)
Noun Phrase Construction: Postpositional, Genitive-Noun, Noun-Adjective,
    Noun-Demonstrative, Noun-Num, Possessed-Possessor
Articles: No definite/specific or indefinite articles
Morphological Marking: No productive singular, dual, or plural marking on
    nouns; Possession marked by suffix on the possessed noun; Tense marked by
    auxiliary particle
Syntactic Alignment: Accusative
Negation: Standard negation marked clause-finally; Distinct imperative
    negation
Reduplication: Both verbs and nouns can be reduplicated
=== End of grammar sketch ===

```

C.12 Predicted Kalamang Grammar Sketches and Accuracy

Listing C.4: Grammar sketch for Kalamang, generated by GPT-3.5-turbo. Each feature was predicted individually based on randomly retrieved examples from the training data.

```

=== Start of grammar sketch ===

```

Basic Word Order: SV (Subject-Verb), OV (Object-Verb)
Noun Phrase Construction: Adposition-Noun Phrase, Genitive-Noun, Adjective-Noun, Noun-Demonstrative, Noun-Num, Possessor-Possessed
Articles: Unsure if there are definite/specific or indefinite articles
Morphological Marking: Unsure if singular, dual, or plural marking on nouns; Unsure if possession marked by suffix on the possessed noun; Unsure if tense marked by auxiliary particle
Syntactic Alignment: Unsure
Negation: Standard negation marked clause-initially; Unsure if distinct imperative negation
Reduplication: Unsure if verbs and nouns can be reduplicated
=== End of grammar sketch ===

Listing C.5: Grammar sketch for Kalamang, generated by GPT-4-turbo. Each feature was predicted individually based on randomly retrieved examples from the training data.

=== Start of grammar sketch ===
Basic Word Order: SV (Subject-Verb), VO (Verb-Object)
Noun Phrase Construction: Unsure on order of adposition and noun phrase, Genitive-Noun, Adjective-Noun, Unsure on order of noun and demonstrative, Num-Noun, Possessor-Possessed
Articles: No definite/specific or indefinite articles
Morphological Marking: Unsure if productive singular, dual, or plural marking on nouns; Unsure if Possession marked by suffix on the possessed noun; Unsure if tense marked by auxiliary particle
Syntactic Alignment: Unsure
Negation: Standard negation marked clause-initially; Unsure if distinct imperative negation
Reduplication: Unsure if verbs and nouns can be reduplicated
=== End of grammar sketch ===

Model	Grammar feature accuracy
GPT-3.5-turbo	27.78%
GPT-4-turbo	22.22%
Llama-2-7B	0%
Llama-2-70B	0%

Table C.5: Overall grammar sketch accuracy of models. Unsure answers were marked as incorrect.

C.13 Vocabulary Induction Accuracy for MTOB

The overall vocabulary induction accuracy for gpt models is shown in [Table C.6](#). Accuracy and chrF were calculated overall when exclude morphology is marked No, and morphologically rich words in the dictionary (marked with * or with - to represent suffixes/prefixes) were excluded when this value was marked Yes. For words with multiple translations in the dictionary, the translation was marked correct if it matched any of the possible translations. Words without an entry in the dictionary were skipped. If a null hypothesis such as *I don't know* was proposed, it was marked as incorrect.

Table C.6: Accuracy and segment-level chrF for vocabulary hypotheses in Kalamang translation. If morphology was excluded, all words with a * or - symbol in the translation (corresponding to prefixes/suffixes) were excluded, whereas otherwise we matched whether or not the model's hypothesis began/ended with the correct characters.

direction	model	rerank method	exclude morphology	acc	chrF
ek	gpt-4-turbo	p_data_given_hyp_guess	No	0.1528	20.33
ek	gpt-4-turbo	p_data_given_hyp_guess	Yes	0.0813	11.86
ek	gpt-4-turbo	p_answer_given_hyp_logprobs	No	0.16	20.69
ek	gpt-4-turbo	p_answer_given_hyp_logprobs	Yes	0.08621	12.18
ek	gpt-4-turbo	p_data_given_hyp_logprobs	No	0.166	20.79
ek	gpt-4-turbo	p_data_given_hyp_logprobs	Yes	0.1015	13.21
ke	gpt-3.5-turbo	p_answer_given_hyp_logprobs	No	0.1033	15.33
ke	gpt-3.5-turbo	p_answer_given_hyp_logprobs	Yes	0.8244	13.11
ke	gpt-3.5-turbo	p_data_given_hyp_guess	No	0.11	16.42
ke	gpt-3.5-turbo	p_data_given_hyp_guess	Yes	0.09018	14.33
ke	gpt-3.5-turbo	p_data_given_hyp_logprobs	Yes	0.0738	12.34
ke	gpt-3.5-turbo	p_data_given_hyp_logprobs	No	0.1014	15.28
ke	Llama-2-70b	p_data_given_hyp_guess	Yes	0.1344	21.69
ke	Llama-2-70b	p_data_given_hyp_guess	No	0.155	23.62
ek	gpt-3.5-turbo	p_data_given_hyp_guess	Yes	0.0922	15.05
ek	gpt-3.5-turbo	p_data_given_hyp_guess	No	0.1217	19.29
ke	gpt-4-turbo	p_answer_given_hyp_logprobs	Yes	0.08127	13.54
ke	gpt-4-turbo	p_answer_given_hyp_logprobs	No	0.0979	15.25
ke	gpt-4-turbo	p_data_given_hyp_logprobs	Yes	0.07641	12.82
ke	gpt-4-turbo	p_data_given_hyp_logprobs	No	0.0924	14.46
ke	Llama-2-70b	p_answer_given_hyp_logprobs	Yes	0.1339	21.92
ke	Llama-2-70b	p_answer_given_hyp_logprobs	No	0.1447	22.92
ke	Llama-2-70b	p_data_given_hyp_logprobs	Yes	0.1334	21.36
ke	Llama-2-70b	p_data_given_hyp_logprobs	No	0.1442	22.38

Due to hypotheses for this task being more difficult to automatically evaluate, an author also annotated the correctness of results for a single setting (instruction-inference: verbalized confidence) in order to calculate the association between translation quality and hypothesis correctness more accurately. Results can be found in [Table C.7](#).

Model	Correlation	p-val
GPT-3.5-turbo	0.1746	0.006331
GPT-4-turbo	0.1201	0.05088

Table C.7: Point-biserial correlation between segment-level chrF of each translation with the correctness of vocabulary hypotheses induced for that question (in the instruction-inference:verbalized confidence setting). No correction for multiple comparisons was done because we examined hypotheses at the model level (see § C.8 for how this was done in the colours domain)

C.14 Example Kalamang Translations

Table C.8 shows example translations made by GPT-3.5-turbo in both directions.

Table C.8: Example translations by the GPT-3.5-turbo model in different settings.

Source	Reference	Translation (few-shot)	Translation (instruction- induction:verbalized confidence)	Translation (true- instruction)
Who asked you? kiemara paruo?	Namana gerket?	kat Naman kiemara kowam?	Naman era?	Naman mua
The rope is wrapped around it.	Karek kokarap.	mat Kainasu kiwat bo dalang- dalang.	esun Sar tok esun et dalangdalang na.	Karek mam- bon guang- garien ma he me.
Before, they said Japan and The Nether- lands were at war.	Wiseme toni Japang- bon Beladar- bon nausair.	mu Wiseme Jafaris emun Nabil emun kaden kiesun ning.	mu Wise me mu salai Japan an Sar Be- ladargara din jambat mara.	Tikninda, termus siram panci tubak Beladar tamatko pang.
Mu bo Tamisen- gga som dande yecie.	They are going (on a return trip) to bury someone in Antalisa.	They go to Tamisengga and return.	They go to Antalisa, someone died, and they re- turn.	They will return to Tamisengga until they bury the person.
Suo me kene arkon metko.	At Suo is a kene tree.	I washed my body in the river.	Back me gives birth canoe.	Wash your face near the Narkon tree.
An istrat ka- hendengoa marmar ba temun.	I'm walking on a long wide road.	I put the long stick on the ground.	I have a long drying rack but I am walk- ing.	I walked far down the street to my uncle.

APPENDIX D

APPENDIX FOR CHAPTER [4.2](#)

D.1 List of all models

All models are listed in [Table D.1](#).

Table D.1: Model Parameter Counts by Organization (sorted by size)

Organization	Model Name	Parameters
EleutherAI [24]	pythia-14m	14M
EleutherAI	pythia-70m-deduped	70M
EleutherAI	pythia-70m	70M
facebook [184]	opt-125m	125M
EleutherAI [30]	gpt-neo-125m	125M
HuggingFaceTB [9]	SmolLM-135M	135M
EleutherAI	pythia-160m	160M
EleutherAI	pythia-160m-deduped	160M
None (this paper)	llama2_220M_nl_100_code_0	220M
None (this paper)	llama_220M_nl_80_code_20	220M
None (this paper)	llama2_220M_nl_40_code_60	220M
None (this paper)	llama2_220M_nl_20_code_80	220M
None (this paper)	llama2_220M_nl_0_code_100	220M
Salesforce [197]	codegen-350M-mono	350M
Salesforce	codegen-350M-multi	350M
Salesforce	codegen-350M-nl	350M
facebook	opt-350m	350M
HuggingFaceTB	SmolLM-360M	360M
EleutherAI	pythia-410m-deduped	410M
EleutherAI	pythia-410m	410M
facebook [185]	xglm-564M	564M
EleutherAI	pythia-1b-deduped	1B
bigscience [26]	bloom-1b7	1B
EleutherAI	pythia-1b	1B
cerebras [40]	Cerebras-GPT-1.3B	1.3B
microsoft [160]	phi 1.5	1.3B
EleutherAI	gpt-neo-1.3B	1.3B
EleutherAI	pythia-1.4b	1.4B
EleutherAI	pythia-1.4b-deduped	1.4B
HuggingFaceTB	SmolLM-1.7B	1.7B
Salesforce	codegen-2B-mono	2B
Salesforce	codegen-2B-nl	2B
Salesforce	codegen-2B-multi	2B
google [86]	gemma-2-2b	2B
cerebras	Cerebras-GPT-2.7B	2.7B
EleutherAI	gpt-neo-2.7B	2.7B
NinedayWang [297]	PolyCoder-2.7B	2.7B

Continued on next page

Table D.1 – Continued from previous page

Organization	Model Name	Parameters
facebook	opt-2.7b	2.7B
microsoft [2]	phi 2	2.7B
EleutherAI	pythia-2.8b	2.8B
EleutherAI	pythia-2.8b-deduped	2.8B
facebook	xglm-2.9B	2.9B
Qwen [224]	Qwen2.5-3B	3B
cerebras [63]	btlm-3b-8k-base	3B
openlm-research [87]	open_llama_3b_v2	3B
rinna [240]	bilingual-gpt-neox-4b	4B
Dampish	StellarX-4B-V0	4B
facebook	xglm-4.5B	4.5B
Salesforce	codegen-6B-multi	6B
EleutherAI [278]	gpt-j-6b	6B
Salesforce	codegen-6B-nl	6B
Salesforce	codegen-6B-mono	6B
cerebras	Cerebras-GPT-6.7B	6.7B
facebook	opt-6.7b	6.7B
EleutherAI	pythia-6.9b-deduped	6.9B
EleutherAI	pythia-6.9b	6.9B
Qwen [14]	Qwen-7B	7B
aisingapore [176]	sea-lion-7b	7B
bigscience	bloom-7b1	7B
google [85]	gemma-7b	7B
mosaicml [189]	mpt-7b	7B
openlm-research	open_llama_7b	7B
tiiuae [119]	falcon-7b	7B
allenai [76]	OLMo-7B-hf	7B
huggyllama [270]	llama-7b	7B
LLM360 [171]	Amber	7B
LLM360	CrystalCoder	7B
facebook	xglm-7.5B	7.5B
meta-llama [5]	Meta-Llama-3-8B	8B
google	gemma-2-9b	9B
01-ai [1]	Yi-9B	9B
EleutherAI	pythia-12b	12B
EleutherAI	pythia-12b-deduped	12B
cerebras	Cerebras-GPT-13B	13B
meta-llama [271]	Llama-2-13b-hf	13B
Qwen	Qwen1.5-14B	14B
Qwen	Qwen2.5-14B	14B

Continued on next page

Table D.1 – Continued from previous page

Organization	Model Name	Parameters
Salesforce	codegen-16B-nl	16B
Salesforce	codegen-16B-mono	16B
EleutherAI	gpt-neox-20b	20B
mosaicml	mpt-30b	30B
Qwen	Qwen2.5-32B	32B
Qwen	Qwen1.5-32B	32B
AbacusResearch	Jallabi-34B	34B
01-ai	Yi-34B	34B
01-ai	Yi-34B-200K	34B
meta-llama	Llama-2-70b-hf	70B
meta-llama [5]	Meta-Llama-3.1-70B	70B
meta-llama	Meta-Llama-3-70B	70B
Qwen [224]	Qwen2-72B	72B
Qwen	Qwen2.5-72B	72B
Qwen	Qwen1.5-110B	110B

D.2 List of all architectural and data features

D.2.1 Architectural Features

Note that features in this section are collected from official documentation (e.g. huggingface model/data cards or original papers).

- **Total parameters** - the total number of parameters (embedding included) in the model. Note that we only include decoder-only dense models.
- **Dimension** - the embedding dimension.
- **Num heads** - the number of attention heads.
- **MLP ratio** - the ratio of $\frac{\text{FFN dimension}}{\text{embedding dimension}}$.
- **Positional Embeddings** - the type of positional embedding. This is either non-parametric (sinusoidal or fixed embeddings), learned (just learned as a vector per position), rope (rope embeddings), or alibi (technically not an embedding, but included here due to its functional purpose)
- **LayerNorm** - the type of layernorm applied. This is either non-parametric (just an arithmetic based normalization), parametric (similar, but with some learnable parameters such as scaling/biases), and RMSNorm (a simplified version of parametric)
- **Attention variant** - The broad type of attention used. This is either full (vanilla attention), local (each token position only attends to positions around it), mqa (multi-query attention), or gqa (grouped-query attention)

- **Biases** - whether or not bias terms are present in parts of the model. Either none (no biases), attn only (only in attention layers), ln only (only in layer norm)
- **Block type** - whether or not the transformer blocks are computed in parallel at all. Sequential indicates not, while parallel indicates some parallelism in attention or FFN layers.
- **Activation** - the activation function used. Either relu, gelu/gelu variations, silu, or swiglu.
- **Sequence length** - the sequence length.
- **Batch instances** - the batch size used during pretraining.

D.2.2 Data Features

Note that features in this section are collected from official documentation (e.g. huggingface model/data cards or original papers).

- **Total tokens (B)** - total number of tokens used during pretraining, measured in billions (converted to log scale)
- **% Web in Pretraining** - Percentage of pretraining data from general web sources.
- **% Code in Pretraining** - Percentage of pretraining data that consists of code.
- **% Books in Pretraining** - Percentage of pretraining data from books.
- **% Reference in Pretraining** - Percentage of pretraining data from reference sources.
- **% Academic in Pretraining** - Percentage of pretraining data from academic sources.
- **% English in Pretraining** - Percentage of English text in the pretraining data.

D.2.3 Freegen-derived Features

These features are derived from model generations. For each model, 5–10k generations are extracted and the following metrics are aggregated (by mean and standard deviation). However, bigram entropy, the educational classifier score, and domain classifications are exceptions, as they are computed once across all generations.

We use Stanza [222] to generate the parse-based features after classifying generations by language. We only include languages that are supported by stanza in the final set of generations that the parse features are based on.

Generation Length & Basic Statistics

- **Mean Character Length** – Average number of characters per generation (capped at 2048).
- **Mean Tokens Generated** – Average number of tokens per generation.
- **Mean Sentences** – Average number of sentences per generation.
- **Mean Words** – Average number of words per generation.

- **Mean Words per Sentence** – Average number of words per sentence.

Constituency Parse Features

- **Mean Depth of Deepest Parse Tree** – Average maximum constituency tree depth per generation.
- **Mean Depth of Parse Trees** – Average constituency tree depth across all sentences/phrases.
- **Mean Word Depth** – Average depth of words within constituency trees.
- **Mean Word Depth Variation** – Average standard deviation of word depths across sentences/phrases.

Dependency Parse Features

- **Mean 90th-Percentile Dependency Head Distances** – For each generation, compute the 90th-percentile of the linear distances between words and their dependency head, then average these values.
- **Mean Maximum Dependency Head Distances** – Average maximum distance from any word to its dependency head per generation.
- **Mean Median Dependency Head Distances** – Average median dependency-head distance per generation.
- **Mean Maximum Dependency Root Distances** – Average maximum distance from any word to the sentence root per generation.
- **Mean Mean Dependency Root Distances** – Average of the mean distances from words to the sentence root per generation.
- **Mean Median Dependency Root Distances** – Average of the median distances from words to the sentence root per generation.

Domain Classification Features

- **% Generated Academic-like Text** – Percentage of generations classified as academic-like.
- **% Generated Books-like Text** – Percentage of generations classified as books-like.
- **% Generated Code-like Text** – Percentage of generations classified as code-like.
- **% Generated Reference-like Text** – Percentage of generations classified as reference-like.
- **% Generated Specialized Text** – Percentage of generations classified as specialized (e.g., music scores, chess PGNs, biomedical data).
- **% Generated Web-like Text** – Percentage of generations classified as web-like.

Classifier and Language Metrics

- **Mean Educational Classifier Score** – Average score assigned by the educational classifier.
- **% Generated English Text** – Average percentage of text generated in English.

Lexical Diversity and Entropy Metrics

- **Mean Bigram Entropy** – Average entropy computed on bigrams across generations.
- **Type-Token Ratio** – Average ratio of unique tokens to total tokens.
- **Unique Tokens** – Average number of unique tokens per generation.

Lexical and Stylistic Features

- **Content-Function Ratio** – Ratio of content words (nouns, verbs, adjectives, adverbs) to function words.
- **Question Words Ratio** – Ratio of question-related words (e.g. how, what, why, when, where, who, which, whose) per 100k words.
- **Imperative Words Ratio** – Ratio of imperative words (e.g. do, make, consider, take, use, ensure, check, build, apply, run, create, find, go, try, turn, start, stop, put, keep, leave, get, move) per 100k words.
- **Conjunctions Ratio** – Ratio of conjunction words (e.g. and, but, or, so, because, although, however, therefore, yet) per 100k words.
- **Instruction Words Ratio** – Ratio of instruction-oriented phrases (e.g. “Question:”, “Answer:”, “Instruction:”, “User:”, “Assistant:”, “Q:”, “A:”) per 100k words.
- **Numbers Ratio** – Ratio of numerical tokens in the generated text.

D.3 List of all evaluations and settings

Although we ideally would evaluate the full cross-product of models and tasks, we found that due to some models being incompatible with LM Evaluation Harness and compute constraints we could not evaluate all 92 models on every dataset. We list in [Table D.2](#) the number of evaluations we currently have for each benchmark and will continue to fill out evaluations in the database.

D.4 Task Deviations from Kaplan-style Scaling Laws

In [Table D.3](#), we document the R^2 value for a fitted power law on the performance of each model.

D.5 Free-generation Domain Classification

We classify model generations into top-level domains with GPT-4o-mini. We found that this multi-stage prompt [Listing D.1](#), [Listing D.2](#) had reasonable precision on a sample of Dolma by domain [\[254\]](#), so use it to classify free-generations.

D.6 Domain Classifier Validation

To validate the reliability of the 4o-mini based classifier, we had an author of this paper annotate 300 selected samples from three pretraining datasets (the Pile, the SmolLM corpus, and RefinedWeb) according to the same annotation standards used in [§ D.5](#). Samples annotated as "unknown" or "incoherent" by either the model or the human annotator were excluded, as these samples are not included in computing the domain mix.

After filtering, we analyzed 258 text samples and found that the human annotator and the model had an 85.8% absolute agreement, and a Cohen’s κ of 0.746, indicating high agreement between human classifications and the model’s classifications.

D.7 Free-generation Validation

To validate our free-generation approach as a proxy for pretraining data composition, we analyzed correlations between the free-generation features of models and their pretraining data.

Namely, for models trained on three open pretraining datasets (the Pile, the SmolLM corpus, and Refinedweb), we compared the features of their free-generations to features produced by the same taggers and the LM-based classifier ([§ D.5](#)) on a randomly sampled 1M document subset of the pretraining corpora. Due to costs, for the domain classification 5k examples from the 1M were used per corpus. The 1M documents were uniformly sampled with reservoir sampling.

Additionally, we calculated two holistic model-wise correlations, which measure how well each model’s complete generation profile matches its training data:

1. **Domain-level correlations:** For each domain category (web, code, academic, books, reference), we computed the correlation between the percentage of that domain in the model’s documented pretraining data and the percentage of free-generations classified into that category.
2. **Feature-level correlations:** For linguistic features (conjunctions ratio, imperative verbs ratio, question words ratio, numbers ratio, character length), we assessed how well each feature in free-generations correlates with corresponding features in pretraining data.

These model-wise correlations are notably higher than individual feature correlations, indicating that models preserve the overall "fingerprint" of their training data better than any single dimension alone.

D.8 Free-generation Examples

Llama-2-70B, web-like: Gwen Stefani Couldn't Recognize Her First Song at an Event with Blake Shelton What's the first song you ever wrote and or recorded? And how would you feel if you never heard it again until you were Facetiming with hundreds of thousands of people while attending a big country music event? That's exactly what happened to Gwen Stefani, when she was honored at the Country Radio Seminar event this week. In Nashville, . One can only imagine that when you've had as crazy a career as she's had, it may be a chore to actually remember all of the milestones over the years, but just imagine also doing it on the spot, for everyone to see. Well, as it turns out, the doubly talented Dolly Parton was set to introduce Gwen as "an artist who broke barriers" when she cited the lyrics from one of Stefani's first-ever "country" songs..."year 3,000." The diplomatic "The Voice" coach Blake Shelton interjected at the video she was showing at that moment, asking "was there other work done

phi-2, books-like: The sun glistened brightly in the evening sky, casting warm rays across the quaint little kitchen where a pair of diligent parents were preparing a scrumptious meal. The air was filled with anticipation as they embarked on a culinary adventure, their hearts filled with love and joy.

With their sleeves rolled up, the parents deftly wielded an assortment of sharp knives, each one an extension of their skill and precision. Like artists with their brushes, they delicately maneuvered through the ingredients, dancing with knives as their partners. The blades moved with a swiftness that defied the physics of everyday movements, gliding through the air with grace and purpose.

...

Qwen2-72B, reference (instruct)-like: Children: Family ticket (at most 2 adults and 2 children) \$23.00 (1) Mr Black wants to see the film with his wife and two sons (they're both over 9), how much should he pay? ____ A. \$23. B. \$46. C. \$48. D. \$49. (2) Mr Brown wants to see the film with his 5-year-old twin sons, how much should he pay _____. A. \$18. B. \$23. C. \$28.5 D. \$32 (3) Lisa can't see snow in this film. ____ A. True B. False (4). Jane is a maths teacher and she has two daughters (she's 12 and her sister is 8). They want to see the film. How much should she pay? ____ A. \$28.5 B. \$27 C. \$33 D. \$26.5 (5) Mike is sixteen. Today is Saturday and he wants to see the film with his 40-year-old parents. They will pay ____ for the tickets. A. \$22.

OPT-6.7B, web-like: I don't think this is a question. It's a statement. I say get a dog, the world is better with dogs.

I said to myself the exact same thing. I don't know why you were down-voted.
because reddit is a bunch of people who hate dogs.

Codegen-16B-mono, code-like:

```
1 from typing import Optional
2
3 from fastapi import FastAPI
4 from pydantic import BaseModel
5
6 app = FastAPI()
7
8
9 class Item(BaseModel):
10     name: str
11     price: float
12     is_offer: Optional[bool] = None
13
14
15 @app.get("/")
16 def read_root():
17     return {"Hello": "World from FastAPI"}
18
19
20 @app.get("/items/{item_id}")
21 def read_item(item_id: int, q: Optional[str] = None):
22     return {"item_id": item_id, "q": q}
23
24
25 @app.put("/items/{item_id}")
26 def update_item(item_id: int, item: Item):
27     return {"item_name": item.name, "item_id": item_id}
```

Pythia-410M, academic-like (patent): This invention relates to new and improved methods and apparatus for removing solid waste material from the waste stream of a power station where the solid waste material is intended for disposal after completion of the power plant.

Various attempts have previously been made to remove solid waste from the waste stream of plants. This is true, for example, to the discharge of sludge which is generally collected into a sludge tank and washed out of the plant in a washing tank which is generally connected to an open drain outlet of the plant. This prior art is discussed by U.S. Pat. No. 3,623,579 which issued to G. R. Clark and described a method for treating the waste stream to remove solid waste by flocculating and flocculating and agitating the solids in a tank to break bonds between the solid particles.

Furthermore, an apparatus was described by U.S. Pat. No. 4,016,823 which describes a method in which liquid sewage is removed from the waste stream and from the sewage treatment plant where the solid waste being removed is to be treated to produce ammonia-purified water for use in bathing baths or soaps and where the sewage from the wastewater treatment plant is removed to the sewage processing plant where this sewage is

mixed with water or treated as a fertilizer.

...

D.9 XGBoost Settings

For the inner grid search, the maximum depth of trees was in [2, 3, 5], while the learning rate was in [0.01, 0.1, 0.3] and the number of trees was in [50, 100].

D.10 Selected Features by Task

In [Table D.6](#), we show the selected features per benchmark.

D.11 LightGBM Results

The LightGBM version of [Table 4.2](#) can be found in [Table D.7](#).

D.12 SHAP Plots for remaining benchmarks

SHAP plots for the remaining benchmarks can be found in [Figure D.1](#) – [Figure D.9](#). Please note that lower scores are better for Brier score tasks (ANLI, XNLI, MathQA, LogiQA2)

D.13 Details on confirmatory pretraining runs

D.13.1 Training

For our confirmatory experiments, we trained 460M parameter Llama-2 architecture models from scratch using the Megatron-Deepspeed library. We capped training tokens to 10B, while using a cosine learning rate schedule set to a length of 100B tokens (meaning that each checkpoint is approximately 10% through a “full” pretraining run). Training took place on one node per checkpoint, with 8 H100 GPUs. Each checkpoint took roughly 6 hours to train.

For our data mixes, we constructed various mixes by using the subsets of the Dolma v1 dataset. In the web vs. other experiments, we fixed the relative percentages of all other data sources while varying the web percentage.

The training configuration is as follows:

training:

```
num_layers: 14
num_attention_heads: 12
seq_length: 2048
```

```
num_kv_heads: 12
hidden_size: 1536
ffn_hidden_size: 4128
tune_steps: 1000
lr: 0.00015
min_lr: 1.0e-5
weight_decay: 1e-2
grad_clip: 1.0
lr_warmup_steps: 100
save_interval: 2000
eval_interval: 2000
train_epochs: 1
tp: 1
micro_batch_size: 16
global_batch_size: 512
seed: 42
```

Aside from data mix, all experiments used identical hyperparameters to ensure fair comparison.

D.13.2 Evaluation

To assess the impact of different data mixes on model performance, we evaluated our models on the following tasks:

1. **Natural language inference:** Lambada, winogrande, arc challenge
2. **Code generation:** Humaneval
3. **Math:** GSM8K
4. **Factuality:** TruthfulQA

Note that we do not select the full evaluation set due to time constraints. As LM eval harness does not implement perplexity/loss based evaluations for all tasks, we manually convert multiple-choice tasks to loss-based metrics, and mask out the prompt or question when calculating loss for all tasks.

D.13.3 Conversion to Loss-Based Metrics

To ensure consistent evaluation across different tasks and models, we converted various benchmark datasets to loss-based metrics. This approach allows for more direct comparison between models and clearer interpretation of improvements. Here’s how we implemented loss calculations for each dataset type:

Multiple Choice Tasks (ARC Challenge, Winogrande, HellaSwag, TruthfulQA): For these datasets, we calculated two primary loss-based metrics:

- **Average Loss:** We computed the negative normalized log probability of the correct answer. For each question, we formatted the input as "Question + Answer Choice", then calculated the sequence log probability normalized by token length for each choice. The negative log probability of the correct answer was used as the loss.
- **Margin-based Loss:** For TruthfulQA specifically, we calculated a margin between truthful and non-truthful answers. This was computed as the negative of the difference between the best truthful answer's log probability and the best non-truthful answer's log probability. A lower loss indicates better differentiation between truthful and non-truthful information.

Generation Tasks (GSM8K, HumanEval, Lambada): For generation tasks, we calculated:

- **Answer Loss:** We computed the cross-entropy loss on the solution tokens only. Note that for Lambada, this is only the last word.

All log probabilities were normalized by sequence length as well.

D.13.4 Full Results

Exact loss values for the code vs. natural language mixes and the web vs. other mixes can be found respectively in [Table D.8](#) and [Table D.9](#).

Task Name	# Models Evaluated
Commonsense Reasoning / NLI	
ANLI	82
HellaSwag	92
Winogrande	92
XNLI	82
Math / Logic	
GSM8K	92
LogiQA2	82
MathQA	82
General Knowledge	
ARC Challenge	92
Lambada	92
MMLU	92
Other	
TruthfulQA	92
HumanEval	91

Table D.2: Number of models evaluated for each benchmark task. Note that some models encountered technical errors when being loaded or in lm-eval-harness. The number of models will continue to be updated.

Benchmark	R^2
gsm8k	0.85
arc challenge	0.82
hellaswag	0.80
winogrande	0.80
mmlu 5-shot	0.80
mmlu 0-shot	0.74
mathqa	0.70
ANLI	0.61
humaneval	0.61
lambada	0.55
LogiQA2	0.50
XNLI	0.41
truthfulqa	0.29

Table D.3: Overview of R^2 values by benchmark.

Domain	Pearson r	p-value
web	0.917	7.55e-12
reference	0.832	3.99e-08
academic	0.824	7.21e-08
code	0.679	7.14e-05
books	0.374	5.02e-02

Table D.4: Domain correlations between pretraining data and free-generations. Web content shows the strongest Pearson correlation, suggesting models most faithfully reproduce web distribution patterns, while books content shows the weakest relationship.

Feature	Pearson r	p-value
conjunctions_ratio	0.690	4.80e-05
question_words_ratio	0.554	2.21e-03
numbers_ratio	0.449	1.66e-02
imperative_verbs_ratio	0.443	1.83e-02
char_len	0.214	2.75e-01

Table D.5: Linguistic feature correlations between pretraining data and free-generations. Connective elements like conjunctions show stronger correlations, while structural features like character length are less preserved in model generations.

Listing D.1: Multistage classification prompt.

<PROMPT 1>. [SYSTEM] You are a system tasked with classifying documents. First, determine if this document is relatively coherent. These documents are generated by language models, so they may not make sense. Classify a document as incoherent ONLY if it shows extreme repetition, code mixes in a way that does not make sense (such as different languages referencing entirely different subjects), or if it is mostly gibberish. Don't worry about logic errors or factual inconsistencies. If multiple documents are mixed into one, classify it as incoherent. Respond ONLY with "incoherent" if the document is incoherent, otherwise respond with "not_incoherent"

[USER] Please classify the document as incoherent or not_incoherent.\nDocument: {document}

If not incoherent...

<PROMPT 2>. [SYSTEM] Determine if this document contains programming code. Look for:

1. Programming language keywords (def, class, import, etc)
2. Code blocks (marked with backticks, indentation patterns)
3. Stack Overflow-style Q&A about programming
4. File extensions (.py, .js, etc)
5. Documentation about code/config files

Respond ONLY with:

- "code" if ANY of these are present
- "not_code" otherwise

[USER] Please classify the document as code or not_code.\nDocument: {document}

If not code...

<PROMPT 3> [SYSTEM] For documents WITHOUT programming code, determine if this is web content. Web content includes news articles, social media and online forums, blog posts, shopping websites, and other general websites. This includes a wide variety of content, and anything that looks like it may be a web article at all should be included. Look for:

1. URLs or hyperlinks
2. Social media formatting (@mentions, #hashtags)
3. "Click here" or UI elements
4. Comment threads or forum posts
5. Shopping/e-commerce language
6. Bylines or author names
7. Descriptions of products or product features

Respond ONLY with:

- "web" if ANY of these are present
- "not_web" otherwise

[USER] Please classify the document as web or not_web.\nDocument: {document}

If not web...

<PROMPT 4> [SYSTEM] For documents WITHOUT programming code, determine if this is academic or patent-related content. Academic content consists of research papers and snippets of research in both sciences and humanities, as well as patent applications. Student essays or assignments should also be included

Listing D.2: Multistage classification prompt (contd).

If not academic...

<PROMPT 5> [SYSTEM] For documents WITHOUT programming code, determine if this is a book, reference material (including media content), or a specific dataset. Books include literary works, fiction, and narrative nonfiction. Reference material includes wikipedia, dictionaries, textbooks and textbook like content, and encyclopedias. Please note that reference should also include instruction or human preference datasets for language model training. Media content includes podcasts, subtitles, and other media-related text. Specific datasets are unique and not covered by the other categories, such as biomedical datasets or molecules, chess PGNs or specific data formats not covered by any other category. Look for:

For the books category:

1. Chapter headings or book titles
2. Fictional character names or dialogue
3. For literary nonfiction, look for a more narrative and less didactic tone
4. Extended narrative prose or dialogue

Do NOT classify as books if the document:

- Only has a single dialogue snippet
- Could be a web article
- Is primarily informational or educational (use reference instead)

For the reference category:

1. Definitions or explanations of terms
2. Encyclopedic formatting
3. Textbook-like language
4. Explanations or examples meant to educate a reader
5. Chat formatting like 'User:/Assistant:' or similar tokens
6. Court documents or legal language (NOT patents)
7. Wikipedia headers such as 'references' or 'external links'

For the media category (should be classified as reference):

1. Audio or video timestamps
2. Subtitles or captions

For the specific datasets category:

1. Unique names or identifiers
2. Dataset-specific formatting
3. Data or metadata descriptions

If this seems to be a web document (social media, news, blogs, forums, shopping), you can also back off to the 'web' category.

Respond ONLY with:

- "books" if the document is a book
- "reference" if the document is reference material
- "specific_datasets" if the document is a specific dataset
- "web" if the document is web content
- "unknown" if none of these are present

[USER] Please classify the document as books, reference, media, specific datasets, or unknown.\nDocument: {document}"

Table D.6: Greedily-selected features per benchmark.

Benchmark	Selected Features
arc challenge (25-shot)	total params, pretraining summary total tokens billions, question words ratio, layer norm type, dimension, pretraining summary percentage code
gsm8k (5-shot)	total params, pretraining summary total tokens billions, pretraining summary percentage reference, edu classifier std, pretraining summary percentage books
hellaswag (10-shot)	total params, pretraining summary total tokens billions, pretraining summary percentage code, pretraining summary percentage reference, positional embeddings, pretraining summary percentage academic
mmlu 0-shot (0-shot)	total params, pretraining summary total tokens billions, layer norm type, activation, pretraining summary percentage code
truthfulqa (0-shot)	total params, pretraining summary total tokens billions, domain web pct mean, dep parse dep root dist max mean, pretraining summary percentage english, entropy mean, layer norm type
winogrande (5-shot)	total params, pretraining summary total tokens billions, question words ratio, layer norm type, pct english mean, positional embeddings, pretraining summary percentage books, pretraining summary percentage code, block type
anli (0-shot)	total params, pretraining summary total tokens billions, pretraining summary percentage code, pretraining summary percentage web, pretraining summary percentage books, positional embeddings
logiqa2 (0-shot)	total params, pretraining summary total tokens billions, pretraining summary percentage web, domain reference pct mean, dep parse dep root dist mean std, dep parse dep root dist median std
mathqa (5-shot)	total params, pretraining summary total tokens billions, pretraining summary percentage books, num heads
xnli (0-shot)	total params, pretraining summary total tokens billions, pretraining summary percentage web
lambada (0-shot)	total params, pretraining summary total tokens billions, pretraining summary percentage code, block type
mmlu 5-shot (5-shot)	total params, pretraining summary total tokens billions, sequence length, biases, num heads, dimension, edu classifier mean, pretraining summary percentage academic
humaneval (0-shot)	total params, pretraining summary total tokens billions, pretraining summary percentage code, layer norm type, pretraining summary percentage english, biases

Benchmark	Setting	Baseline MAE	Scaling Laws MAE	All Features MAE
Accuracy				
Arc Challenge	25-shot	13.23%	4.91%	3.61%
GSM8k	5-shot	15.65%	11.03%	5.78%
Hellaswag	10-shot	12.26%	4.29%	3.14%
Humaneval	0-shot	11.79%	8.61%	6.80%
Lambada	0-shot	16.89%	9.60%	6.39%
MMLU (0-shot)	0-shot	11.98%	9.12%	4.21%
MMLU (5-shot)	5-shot	12.25%	8.39%	3.05%
TruthfulQA	0-shot	3.72%	3.40%	2.61%
Winogrande	5-shot	10.14%	3.99%	3.09%
Brier score ($\times 100$)				
XNLI	0-shot	7.22	4.70	4.32
ANLI	0-shot	9.48	6.14	6.05
MathQA	0-shot	7.57	3.89	2.95
LogiQA2	0-shot	12.74	8.77	8.29

Table D.7: MAE comparison of Scaling Laws and All Features predictors versus Baseline. Note that a significance test was not carried out for LGBM, so this reflects results for one run, though a hyperparameter search is still carried out over the same values as in § D.9 for both predictors. Brier scores are scaled $\times 100$ for comparability. Both predictors here use LGBM.

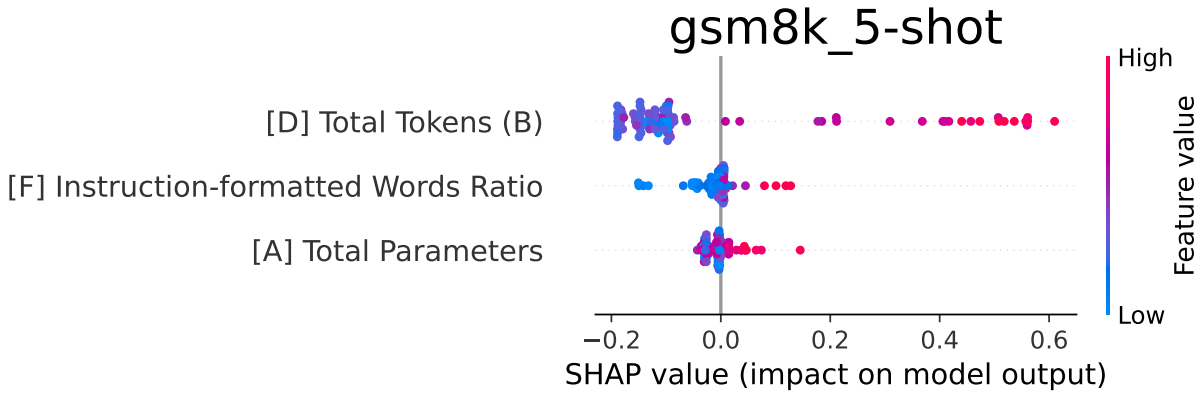


Figure D.1: SHAP values for GSM8k

Data Mix	Lambada	Humaneval	Winogrande	Arc challenge	GSM8k
NL 70/Code 30	3.426	1.331	3.967	3.967	2.295
NL 75/Code 25	3.426	1.331	3.966	3.651	2.295
NL 80/Code 20	3.419	1.350	3.989	3.685	2.293
NL 90/Code 10	3.406	1.533	3.944	3.620	2.300

Table D.8: Loss-based evaluation metrics across different data mix ratios

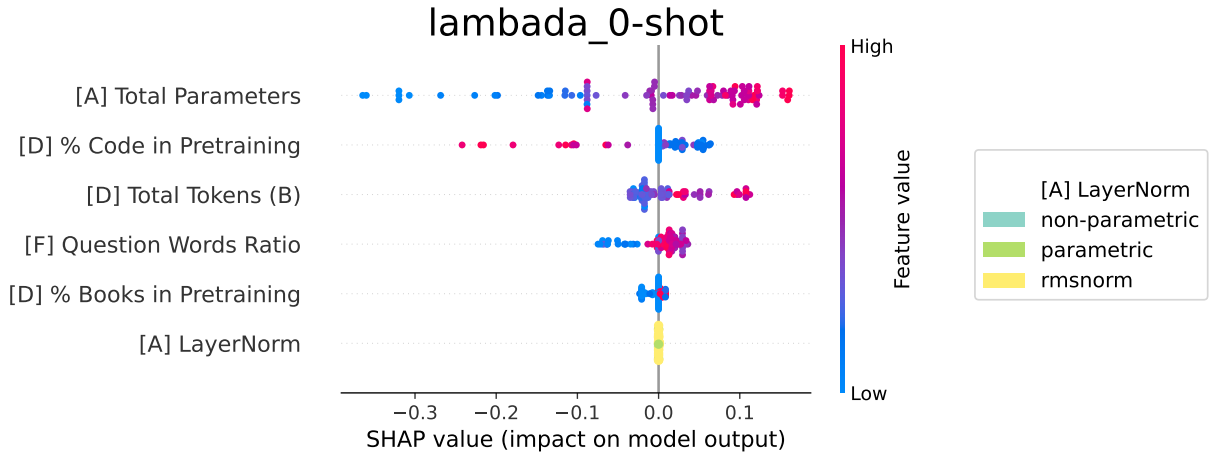


Figure D.2: SHAP values for Lambada

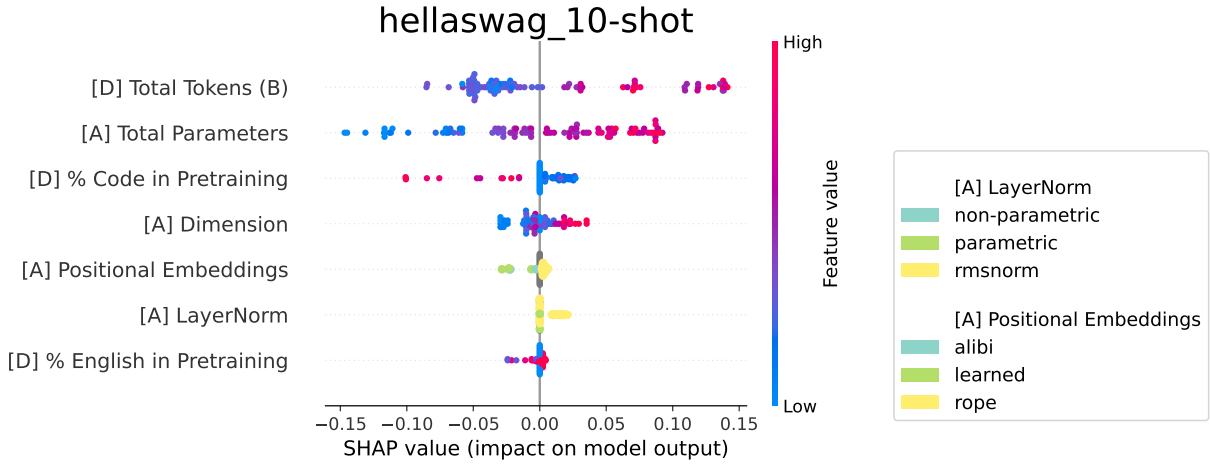


Figure D.3: SHAP values for Hellaswag

Data Mix	Margin Loss	Accuracy (%)
Web 30/Other 70	0.2363	28.86
Web 50/Other 50	0.2342	28.40
Web 90/Other 10	0.2462	28.03

Table D.9: TruthfulQA evaluation metrics across different web/other ratios

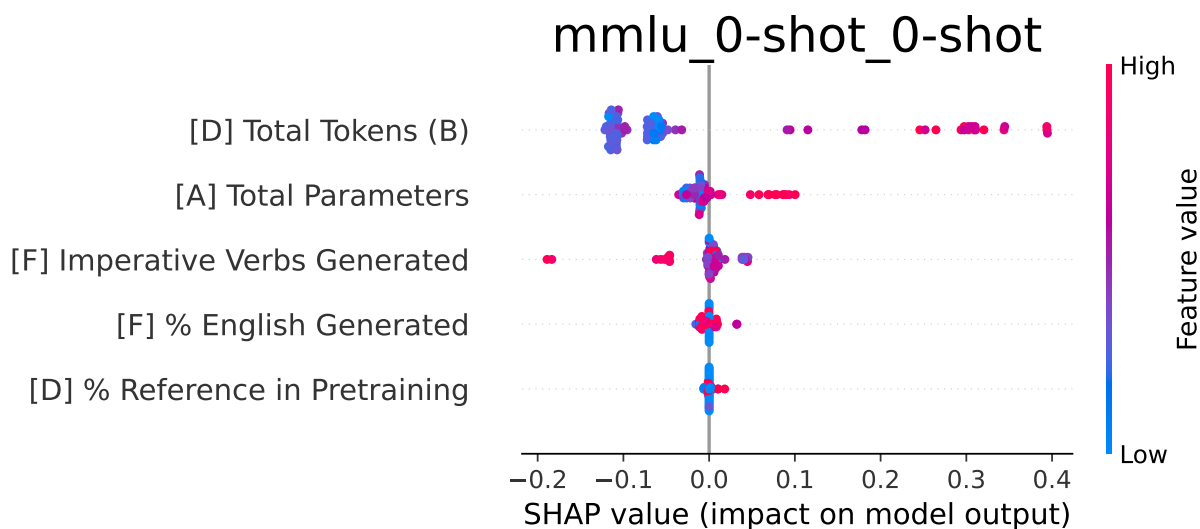


Figure D.4: SHAP values for MMLU 0-shot

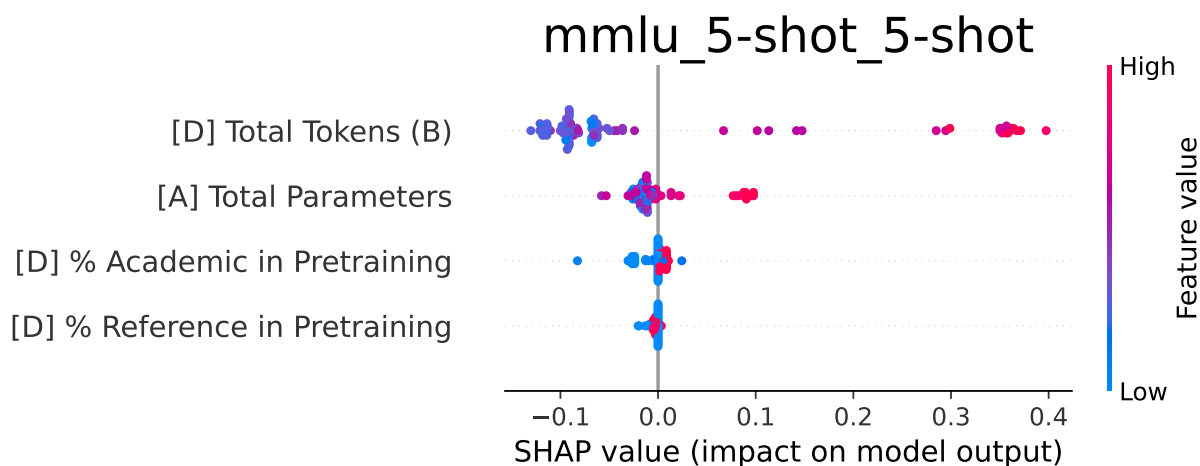


Figure D.5: SHAP values for MMLU 5-shot

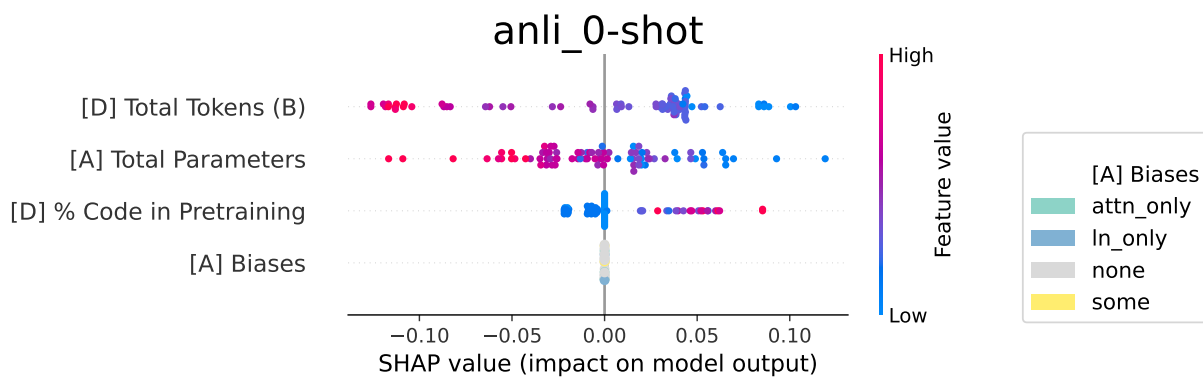


Figure D.6: SHAP values for ANLI

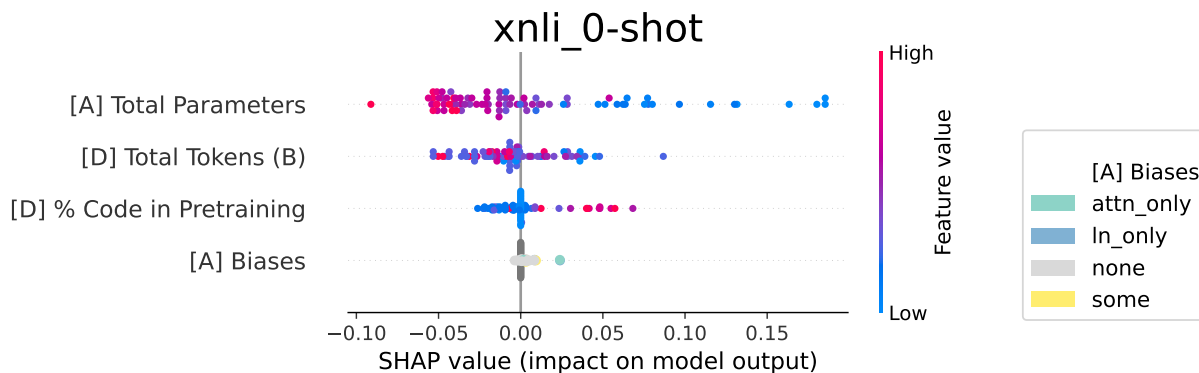


Figure D.7: SHAP values for XNLI

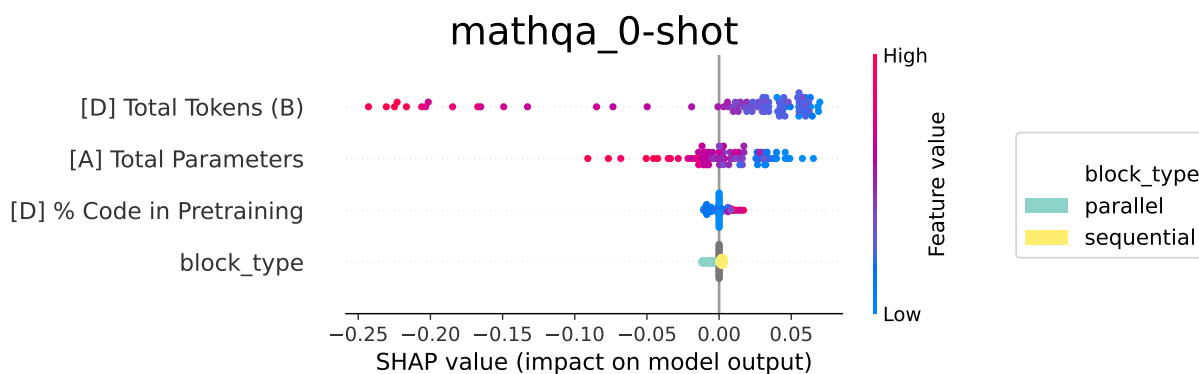


Figure D.8: SHAP values for MathQA

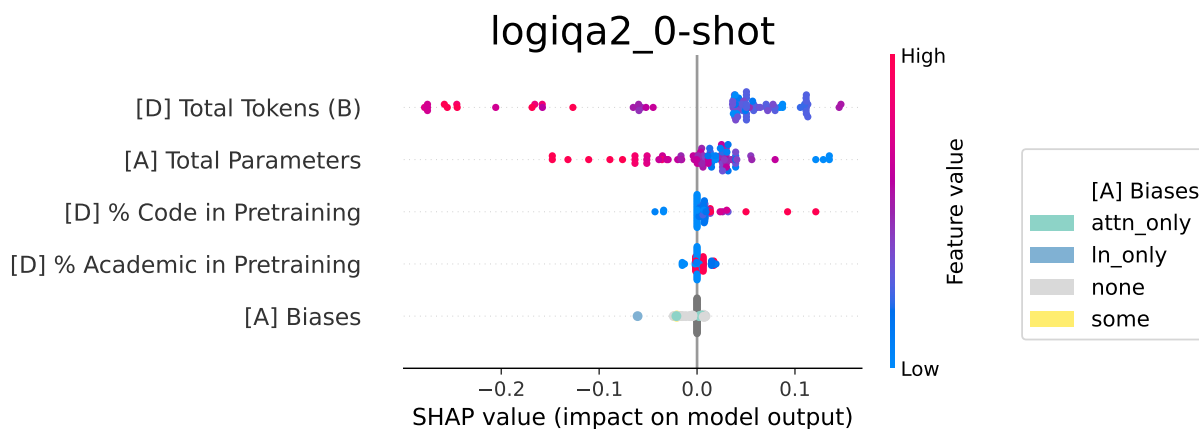


Figure D.9: SHAP values for LogiQA2

APPENDIX E

APPENDIX FOR CHAPTER 4.3

E.1 Pretraining Settings

We pretrained Pythia 70M, 160M, and 410M from scratch on the C4 dataset. All three models were trained for 128B tokens or approximately 61k steps, with very similar settings (documented in Table E.1). L40S GPUs were used for all pretraining and midtraining runs. Models were trained with the LitGPT library [161].

Table E.1: Core pretraining hyperparameters for Pythia-70M, 160M, and 410M.

Hyperparameter	70M	160M	410M
Global batch size	1024	1024	1024
Micro batch size	16	16	8
LR schedule	Cosine w/ 10% warmup	Cosine w/ 10% warmup	Cosine w/ 10% warmup
Max LR	3×10^{-4}	3×10^{-4}	3×10^{-4}
Min LR	1×10^{-6}	1×10^{-6}	1×10^{-6}
Optimizer	AdamW	AdamW	AdamW
Betas	(0.9, 0.95)	(0.9, 0.95)	(0.9, 0.95)
Weight decay	0.1	0.1	0.1
Precision	BF16	BF16	BF16
Num ranks	4	4	8

E.2 Posttraining Settings

We fine-tuned all models on four downstream datasets: PyCode (our 5K-sample subset of CodeSearchNet-Python), GSM8K (7.5K math problems), LIMA (1K instruction examples), and SciQ (13.7K science questions). For GSM8K only, the prompt/question

portion was masked during loss; for the others the loss was computed over the full sequence. A summary of the datasets is given in Table E.2. All runs used a cosine learning rate schedule with 10% linear warmup, trained for 4 epochs, global batch size 64, and micro-batch size 16 for 70M/160M (8 for 410M). Peak learning rates were selected by grid search on the base pretrained checkpoint before midtraining, and the LR grid is given in Table E.3. Selected LRs for the final checkpoint of each model size are given in Table E.4.

Table E.2: Finetuning datasets.

Dataset	# Train Samples	Prompt masked
PyCode (CodeSearchNet-Python subset)	5,000	No
GSM8K	7,500	Yes
LIMA	1,000	No
SciQ	13,679	No

Table E.3: Grid of candidate peak learning rates swept during tuning.

LR grid
4e-6, 8e-6, 1e-5, 2e-5, 4e-5, 5e-5, 6e-5, 7e-5, 8e-5, 9e-5, 1e-4, 1.2e-4, 1.4e-4, 1.6e-4, 1.8e-4, 2e-4, 2.4e-4, 4e-4, 5e-4, 6e-4, 8e-4, 1e-3, 2e-3, 3e-3, 4e-3, 6e-3

Table E.4: Selected peak learning rates for fine-tuning (cosine schedule with 10% warmup).

Dataset	70M	160M	410M
GSM8K	8e-4	4e-4	4e-4
LIMA	1.2e-4	5e-5	5e-5
PyCode	1e-3	5e-4	4e-4
SciQ	8e-4	2.4e-4	6e-4

E.3 Dataset Similarity Matrix

We compute dataset similarity using surface-level token statistics after initial experimentation with embedding models gave implausible results for code datasets’ similarities to other natural language datasets. For each pair of pretrain/midtrain and downstream datasets, we sample $\max(\text{dataset_size}, 10,000)$ examples. Midtrain mixes are simulated by their actual compositions (e.g., StarCoder is treated as 20% StarCoder + 80% C4).

From the (possibly mixed) texts we build unigram frequency vectors at a token level, normalize to probabilities, and compute: vocabulary Jaccard, overlap ratio, token-frequency cosine similarity, and a Jensen–Shannon-based similarity. These are combined as

$$\text{Combined} = 0.4 \cdot \text{cosine} + 0.3 \cdot \text{Jaccard} + 0.3 \cdot \text{JS_similarity},$$

and used to fill the similarity matrix (diagonal entries are 1). This mixture-aware score reflects both specialty content and dilution by C4. Figure E.1 shows the resulting similarity matrix between pre/midtrain datasets and SFT datasets.

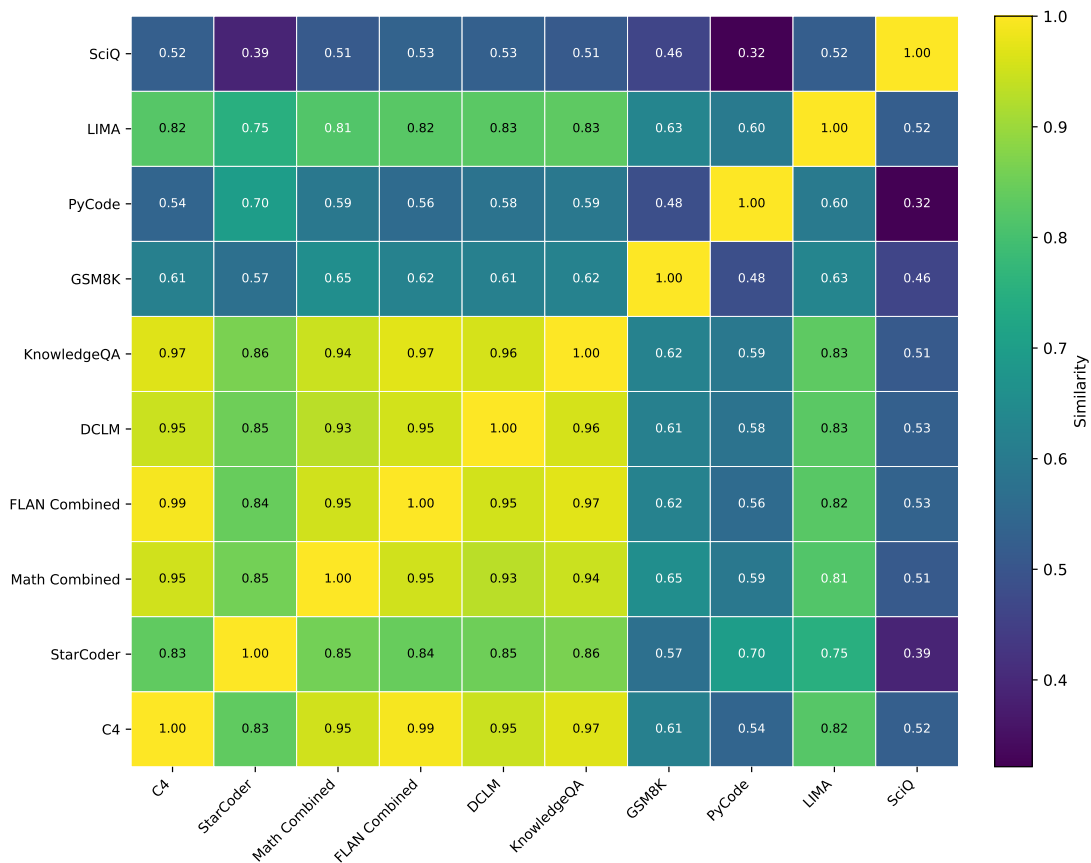


Figure E.1: Token-based similarity matrix for pre/midtrain and SFT datasets. Note that these midtrain datasets are corrected for mix weight in this matrix.

E.4 SFT in-domain loss and C4 Losses after Finetuning for 70m and 160m models

Table E.5 depicts validation losses as well as C4 validation losses after finetuning on each SFT dataset, for the 70m and 160m models.

Table E.5: SFT and C4 validation losses for 70m and 160m models across downstream datasets and midtraining mixtures, averaged across 5 seeds for each SFT dataset. Bold values indicate best performance within each dataset and model size combination.

Model Size	Downstream Dataset	Midtrain Mix	SFT Val Loss	C4 Val Loss
70m	Pycode	C4	2.633	6.067
		Starcoder (20%)	2.530	5.996
		Math (12%)	2.605	6.049
		FLAN (5%)	2.648	6.112
		KnowledgeQA (20%)	2.602	6.046
		DCLM (20%)	2.592	6.007
	GSM8k	C4	1.405	6.409
		Starcoder (20%)	1.363	6.367
		Math (12%)	1.359	6.391
		FLAN (5%)	1.401	6.407
		KnowledgeQA (20%)	1.373	6.434
		DCLM (20%)	1.397	6.421
	LIMA	C4	4.458	4.159
		Starcoder (20%)	4.382	4.168
		Math (12%)	4.418	4.180
		FLAN (5%)	4.460	4.158
		KnowledgeQA (20%)	4.371	4.152
		DCLM (20%)	4.395	4.132
	SciQ	C4	3.573	7.334
		Starcoder (20%)	3.594	7.245
		Math (12%)	3.599	7.289
		FLAN (5%)	3.550	7.236
		KnowledgeQA (20%)	3.537	7.268
		DCLM (20%)	3.529	7.269
160m	Pycode	C4	2.382	5.247
		Starcoder (20%)	2.205	5.104
		Math (12%)	2.382	5.228
		FLAN (5%)	2.389	5.253
		KnowledgeQA (20%)	2.420	5.342
		DCLM (20%)	2.405	5.268
	GSM8k	C4	1.188	5.339
		Starcoder (20%)	1.186	5.362
		Math (12%)	1.139	5.198
		FLAN (5%)	1.071	5.320
		KnowledgeQA (20%)	1.190	5.346
		DCLM (20%)	1.186	5.336
	LIMA	C4	4.184	3.606
		Starcoder (20%)	4.276	3.556
		Math (12%)	4.421	3.599
		FLAN (5%)	4.472	3.580
		KnowledgeQA (20%)	4.498	3.566
		DCLM (20%)	4.446	3.572
	SciQ	C4	3.371	5.537
		Starcoder (20%)	3.362	5.471
		Math (12%)	3.316	5.437
		FLAN (5%)	3.362	5.444
		KnowledgeQA (20%)	3.339	5.373
		DCLM (20%)	3.341	5.383

E.5 Midtraining Leg Length vs. Benefit

Figure E.2 shows the relationship between the two “legs” $\text{dist}(\text{C4}, \text{midtrain})$ and $\text{dist}(\text{midtrain}, \text{SFT})$ and the benefit of midtraining. As in Figure 4.5 this is computed based on the similarity matrix in Appendix E.3.

E.6 Representative training loss curves for midtrained vs. base models

Figure E.3 shows a representative training loss curve for a midtrained model when its domain is aligned to SFT data.

E.7 Additional CKA results on APPS

Figures E.4 and E.5 display the CKA layer similarity for 160m and 410m models.

E.8 CKA results on C4

Figures E.6, E.7, and E.8 show the CKA layer similarity for all model sizes with C4 as a probe.

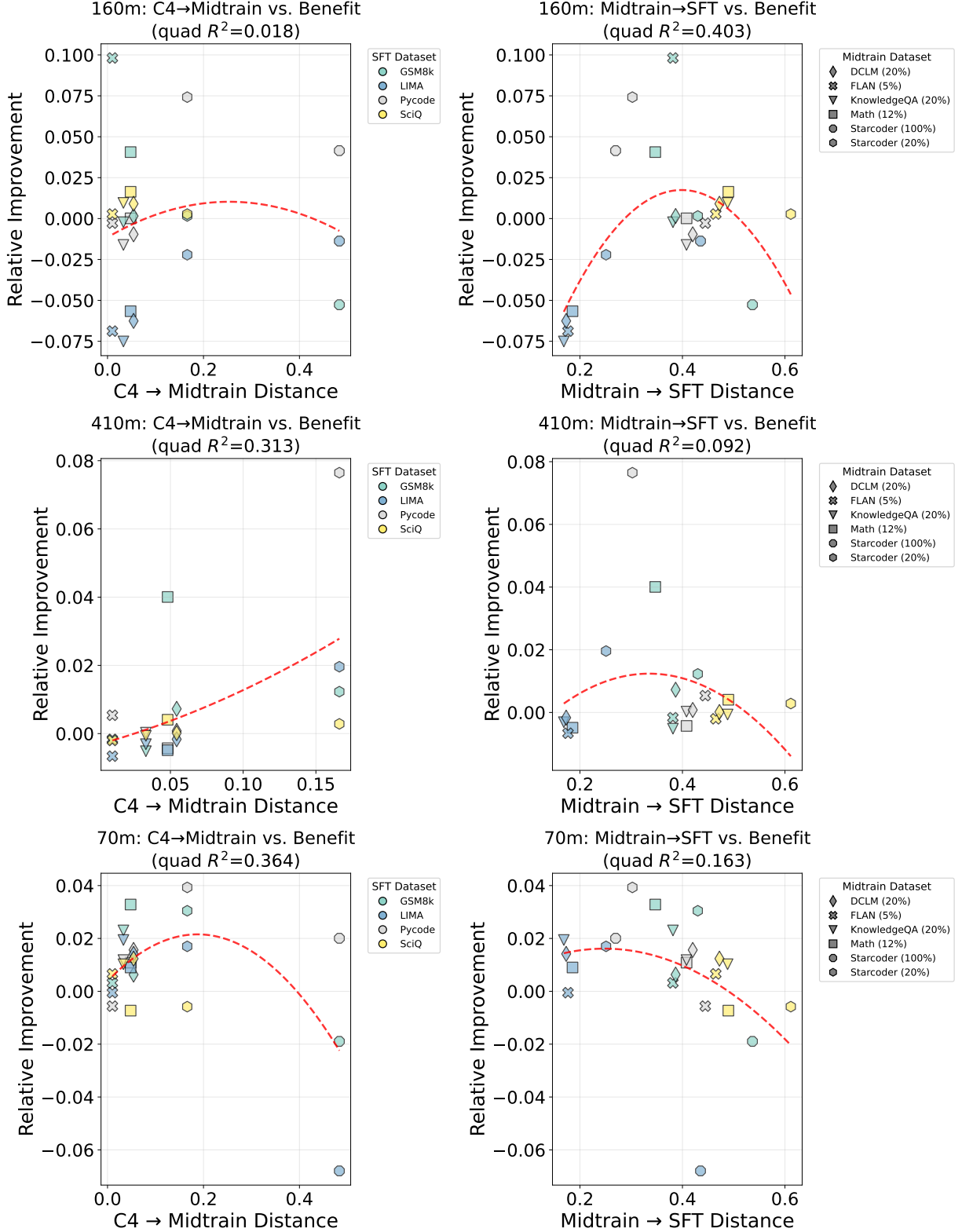


Figure E.2: Relationship between the two “legs” $\text{dist}(\text{C4}, \text{midtrain})$ and $\text{dist}(\text{midtrain}, \text{SFT})$ and the benefit of midtraining.

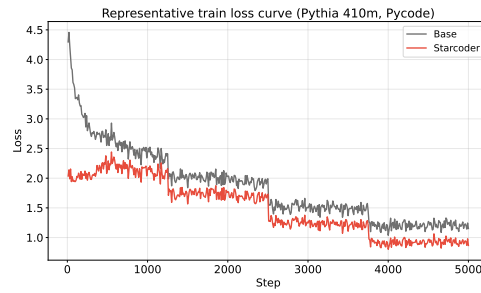


Figure E.3: Representative training loss curve for a midtrained model and base model on Pycode, for Pythia-410m. The midtrained model starts with a lower training loss, and maintains a slight gap throughout training.

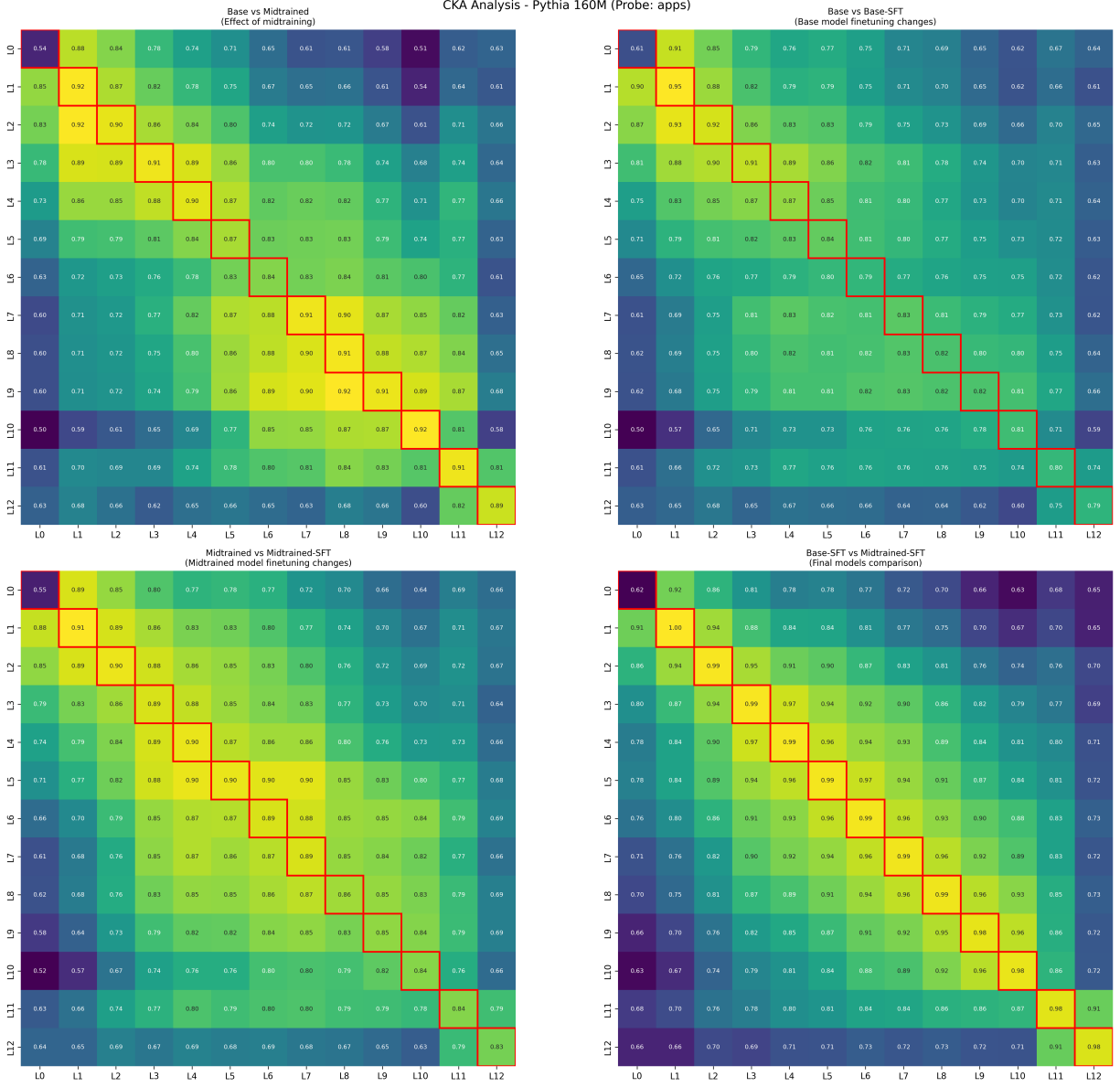


Figure E.4: CKA layer analysis for Pythia-160M with APPS as a probe.

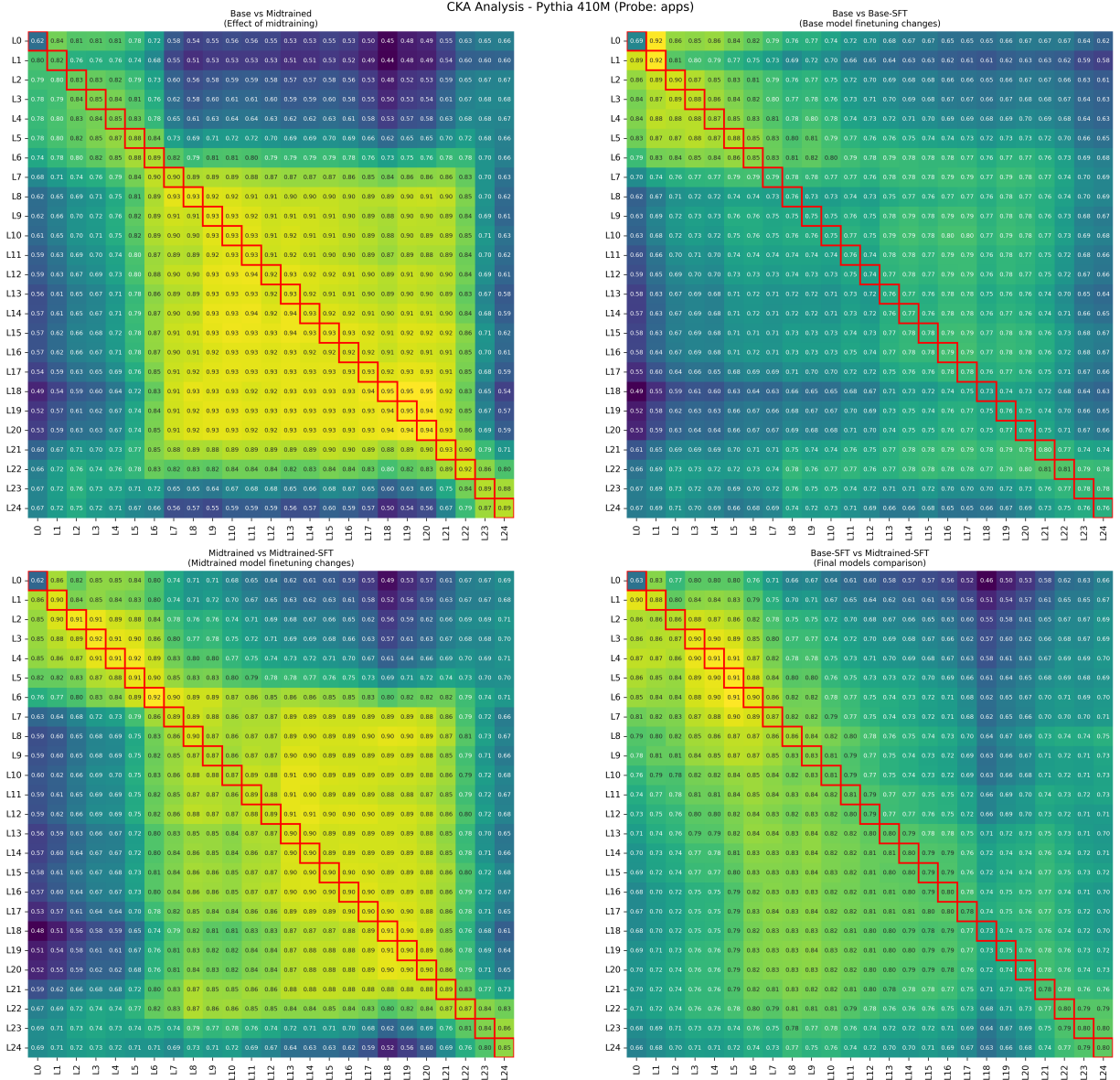


Figure E.5: CKA layer analysis for Pythia-410M with APPS as a probe.

CKA Analysis - Pythia 70M (Probe: c4)



Figure E.6: CKA layer analysis for Pythia-70M with C4 as a probe.

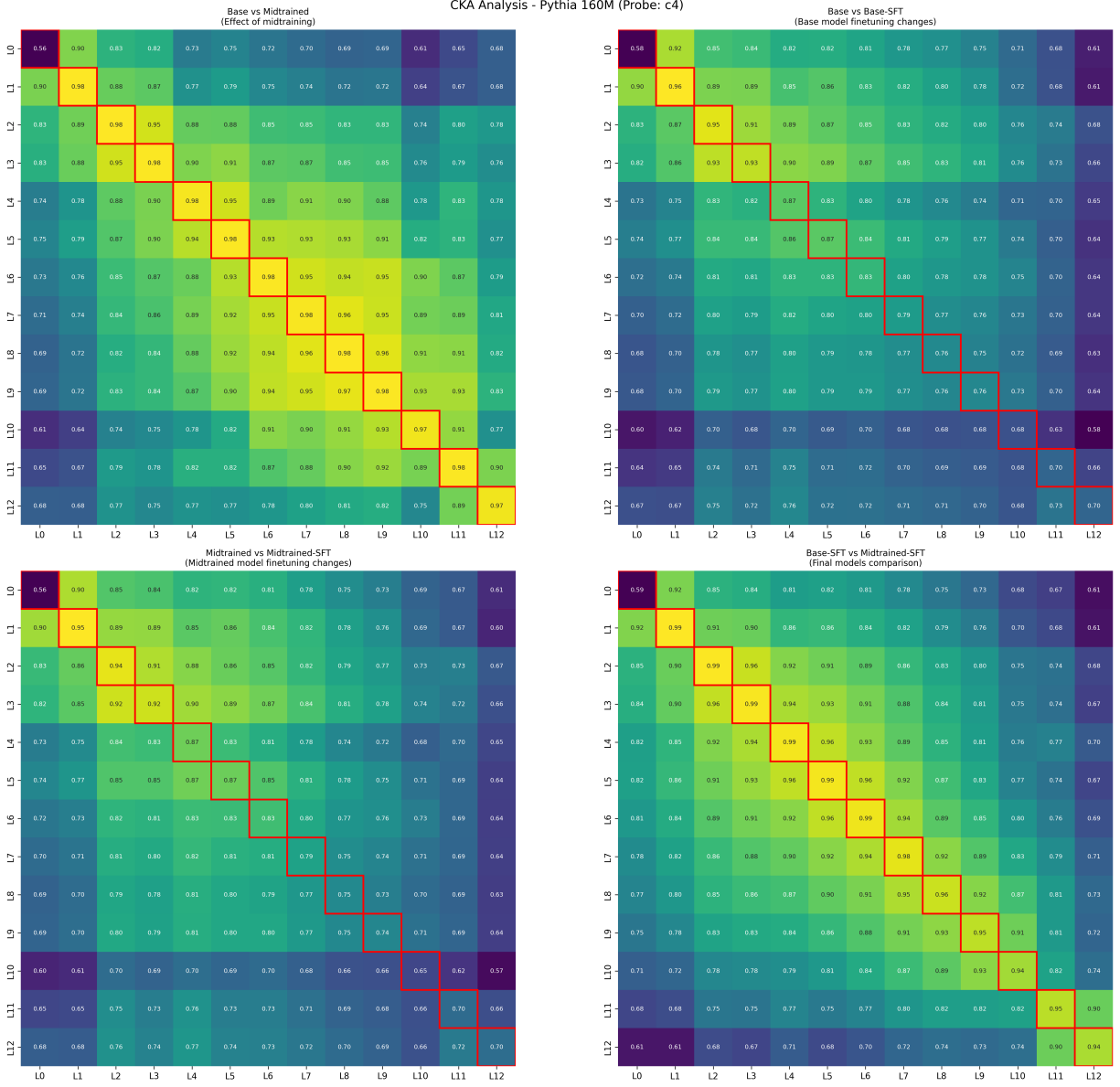


Figure E.7: CKA layer analysis for Pythia-160M with C4 as a probe.

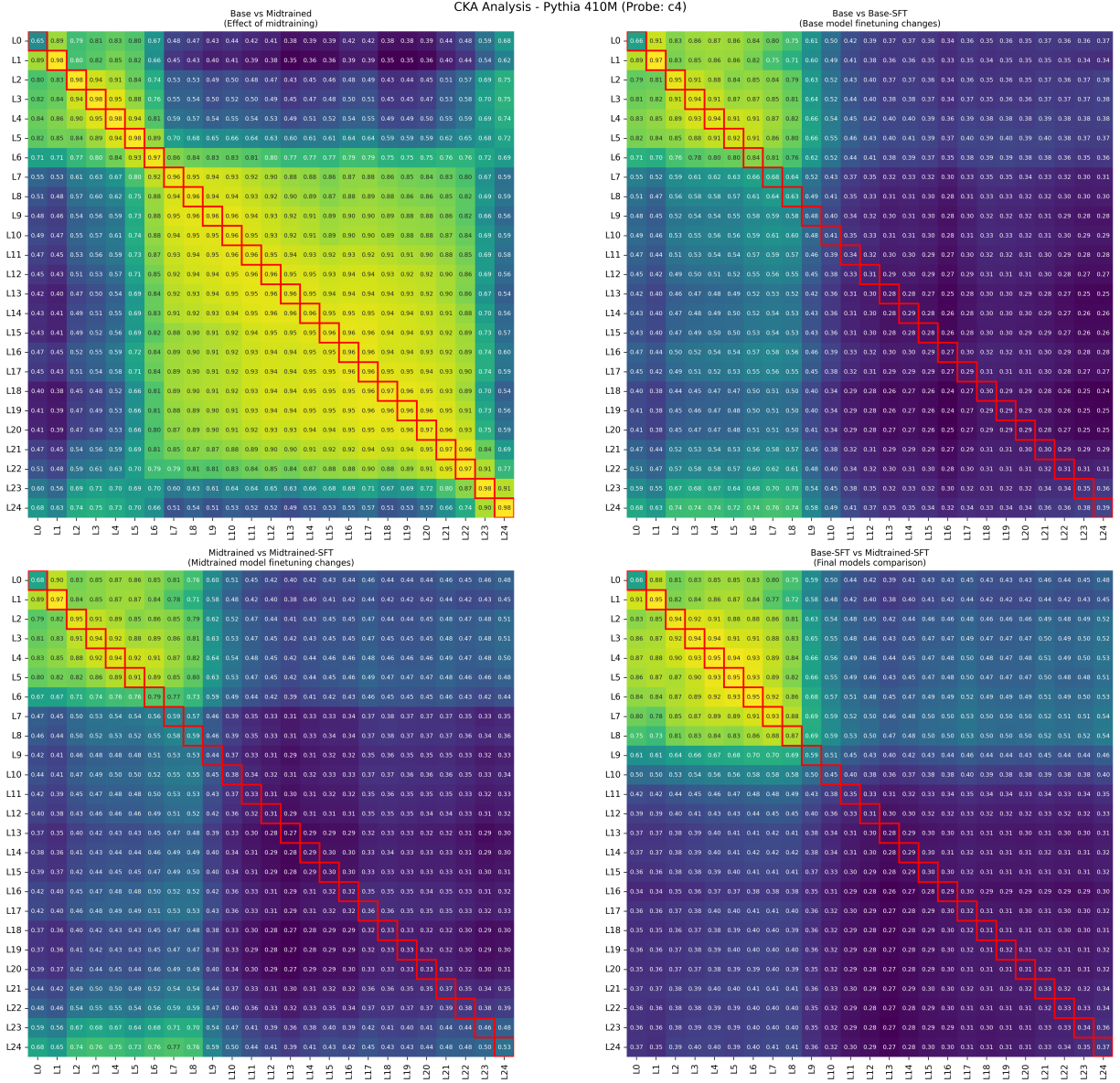


Figure E.8: CKA layer analysis for Pythia-410M with C4 as a probe.

APPENDIX F

APPENDIX FOR CHAPTER 5.2

F.1 Synthetic Dataset Training Details

Three encoder/decoder transformer sizes were trained, differing only in number of encoder/decoder layers. The small size had 3 encoder/decoder layers, the medium size had 8 encoder/decoder layers, and the large size had 16 encoder/decoder layers. For all models, the hidden dimension was 512, the embedding dimension was 512, and there were 16 attention heads.

In the experiments without informative context, the small transformer was trained for 10 epochs, the medium transformer for 20, and the large for 30. In experiments with context, this was changed to 15, 15, and 25 respectively. These values were based on early experimentation with loss plateaus on the validation set.

Sentences in the synthetic dataset were composed of tokens as described in § 5.2.4. Sentences were constrained to be 1-6 tokens in length.

Experiments with the synthetic dataset were implemented in PyTorch.

F.2 Idiom Sources

Sources of idioms were pulled from language-learning websites below:

fi:

- https://en.wiktionary.org/wiki/Appendix:Finnish_idioms

fr:

- <https://speechling.com/blog/20-most-common-french-idioms-to-get-you-talking-like-a-native/>
- <https://frenchtogether.com/french-idioms/>
- <https://vidalingua.com/blog/funny-french-idioms-explained-english>

ja:

- <https://www.theintrepidguide.com/japanese-expressions-and-idioms/>

- https://en.wikipedia.org/wiki/Japanese_proverbs
- <https://www.fluentu.com/blog/japanese/japanese-idioms-2/>
- <https://kotowaza.jitenon.jp/>

F.3 OpenSubtitles Training Details

A separate `deltalm`-base was trained from the pretrained model for 2M steps on each language. The Adam optimizer was used [140], with a learning rate of $1e-4$, betas of $(0.9, 0.98)$, and an inverse square root learning rate scheduler with 4000 warmup updates (with a warmup learning rate of $1e-7$, and a minimum learning rate of $1e-9$). The maximum number of tokens in a batch was set to 1024, with maximum source and target lengths of 512. Label smoothing of 0.1 was used, and the loss function used was cross-entropy.

For upweighting experiments, the upweight coefficient for each language was found through a hyperparameter search on the validation set over $\alpha \in \{3, 5, 10, 50\}$. The values for each language were $\alpha_{fi} = 3, \alpha_{fr} = 10, \alpha_{ja} = 3$.

For kNN-MT, three datastores were built for approximate kNN search using the training set from OpenSubtitles. These datastores were built with the `faiss` library [129]. The Finnish datastore contained 248M vectors with 507k centroids, while the French datastore contained 348M with 713k centroids, and the Japanese datastore 17.8M with 73k centroids. All vectors were stored in fp16 with a code size of 32. The vectors used as keys in the datastore correspond to the *input* to the last feedforward layer. Additionally, a hyperparameter search for each language was carried out on the validation set, over values of $\lambda \in \{0.2, 0.4, 0.6, 0.8\}$, temperature $\in \{0.1, 1, 10\}$, and number of retrieved neighbours $\in \{5, 10, 15, 20\}$. Hyperparameters were selected based on BLEU score on the validation set.

The best hyperparameters were as follows: fi: ($\lambda = 0.4$, temp = 10, num neighbours = 20), fr: ($\lambda = 0.2$, temp = 10, num neighbours = 20), ja: ($\lambda = 0.4$, temp = 10, num neighbours = 20) During test time, the best hyperparameters for each language were used, with a probe size of 20 and beam size of 5.

Experiments with Δ LM and kNN-MT were implemented in fairseq (version from September 2021) [201].

F.4 Native Annotator Recruiting

Annotators were hired through Upwork, and consisted of professional translators in French and Finnish who were also fluent in English. Annotators were paid between \$10 – 25USD an hour, depending on their individual hourly rate. No personally identifying information was collected.

Annotators were shown the task description below, and also had access to a compiled list of literal and figurative translations for each idiom.

Description:

This job is to create a dataset that will allow us (researchers at Carnegie Mellon University, PI Graham Neubig) to study the ability of machine translation systems to translate idioms. We would like you to look at translations and decide whether they are literal or idiomatic translations. These will be English sentences containing the translation of a sentence in either French, Finnish, or Japanese. All the sentences contain a phrase that is an idiom in the foreign language, but sometimes the translations may be meant literally. As an example, in English “kick the bucket” can mean to die when used in an idiomatic way, or to literally kick a bucket when used in a literal way. You should mark the example with “idiomatic” if you think it’s an idiomatic example, or “literal” if you think it’s a literal example.

If you think the translation is not a good translation, we will ask you to mark this entry with the word “None”, but this should occur in relatively few cases.

We estimate that the job will take 2-3 hours overall.

Examples:

Original sentence: Jalkaväkeä oli pilvin pimein. English translation (reference): Arty and infantry just kept on coming. Contains idiom: pilvin pimein Idiom literal meaning: In dark clouds Idiom figurative meaning: A huge (often excessive) amount of something. Label: idiomatic

Original sentence: Harmi, ettei lehdistötilaisuudessaasi ollut lammasta. English translation (reference): Too bad they didn’t have lamb at your press conference, Francis. Contains idiom: olla lammas Idiom literal meaning: To be a lamb Idiom figurative meaning: a person who is like a lamb does nothing alone. the person does Label: literal

F.5 Standards for Human Evaluation

Evaluation standards were based on MQM standards, in particular the major and critical severity levels, which are defined below [172]:

- **Major severity error:** severity level of an error that seriously affects the understandability, reliability, or usability of the content for its intended purpose or hinders the proper use of the product or service due to a significant loss or change in meaning or because the error appears in a highly visible or important part of the content.
- **Critical severity level:** severity level of an error that renders the entire content unfit for purpose or poses the risk for serious physical, financial, or reputational harm.

We slightly adapted this for idioms, where if it was possible to infer the meaning of a sentence through the existence of a similar English idiom, but it would not be something generally said by English speakers, we assigned a major severity. Due to Δ LM often making errors with numbers and named entities, we assigned these errors a major rather than critical severity in most cases, although these would generally be critical severity errors in business documents. If part of a sentence was missing, we based the error severity on whether or not the missing portion was crucial to the intent of the sentence. Examples of sentences labelled with error severity are provided in Table F.1:

Language	Src	Ref	Hyp	Severity
fr	- Eh oui, l’habit ne fait pas le moine.	Yes, clothes don’t make the man.	Yes, clothes don’t make you a monk.	Major
	Il t’a fait une queue de poisson sur l’autoroute.	He cut you off on the freeway.	He made you a fish tail on the highway.	Severe
	Il pleut des cordes.	It’s raining cats and dogs.	It’s raining.	Major
fi	Ei pane tikkuakaan ristiin.	He does nothing.	Doesn’t even cross a match.	Severe
	Tämä meni yli hilseen.	This is just way over my head.	This is over the top.	Major
	Hieno aasinsilta itse-murhaan.	Nice speech, Mr. Hobson. Way to hit the suicide theme.	It’s a nice morning for suicide.	Severe
ja	□□□□□□□□□□ □□□□□□□□	You need to learn the element of surprise.	You need to learn to play hard-to-get.	Severe
	□□□□□□□□□□	I’m not falling for this shit.	I won’t have to do that again.	Major
	□□□□□□□□□□	You swindled me!	You’ve given me enough to eat!	Severe

Table F.1: Examples of categorized errors made by Δ LM.

F.6 Statistical Significance Testing

For each language and test set, we examine the null hypothesis that the BLEU scores of the two systems actually have the same distribution, and the difference occurred by chance.

For each source sentence, we shuffled the translations produced by the two systems with probability 0.5, and then recalculated BLEU scores. This was repeated 1000 times, and the number of times the shuffled difference was greater or equal to the observed difference was recorded to find the p-value.

F.7 Statistical Significance Results

Statistical significance results are shown in [Table F.2](#).

F.8 Effect of Frequency on Commercial Translations

The effect of idiom frequency on commercial models’ translations is in [Figure F.1](#) through [Figure F.5](#).

Language	Test set	p-val
fi	idioms	0.0*
	literal	0.003*
	rand	0.018*
fr	idioms	0.0*
	literal	0.234
	rand	0.0*
ja	idioms	0.058
	literal	0.083
	rand	0.088

Table F.2: p-values obtained using approximate randomization on translations produced by the base model and the upweight+knn model.

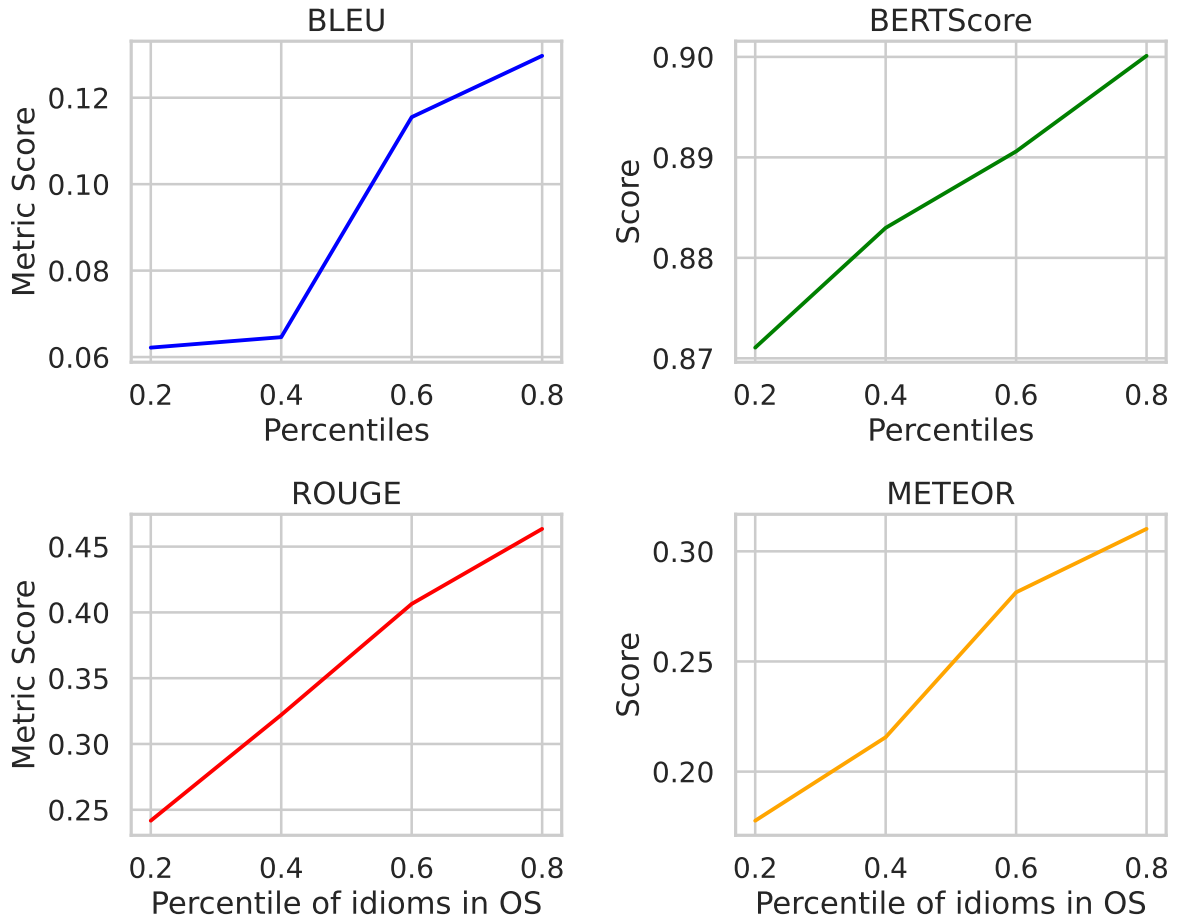


Figure F.1: Effect of idiom frequency on translations by DeepL in Finnish

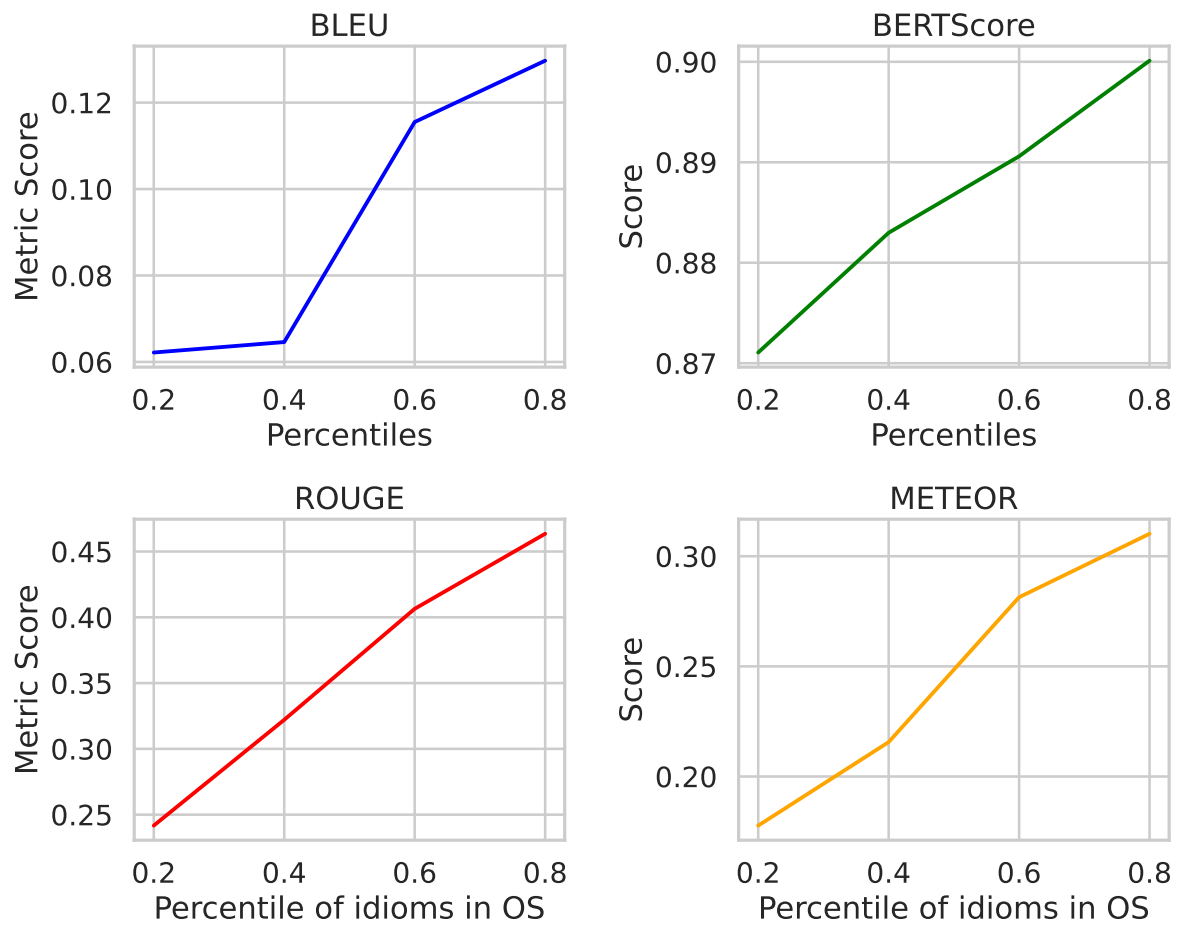


Figure F.2: Effect of idiom frequency on translations by Google in Finnish

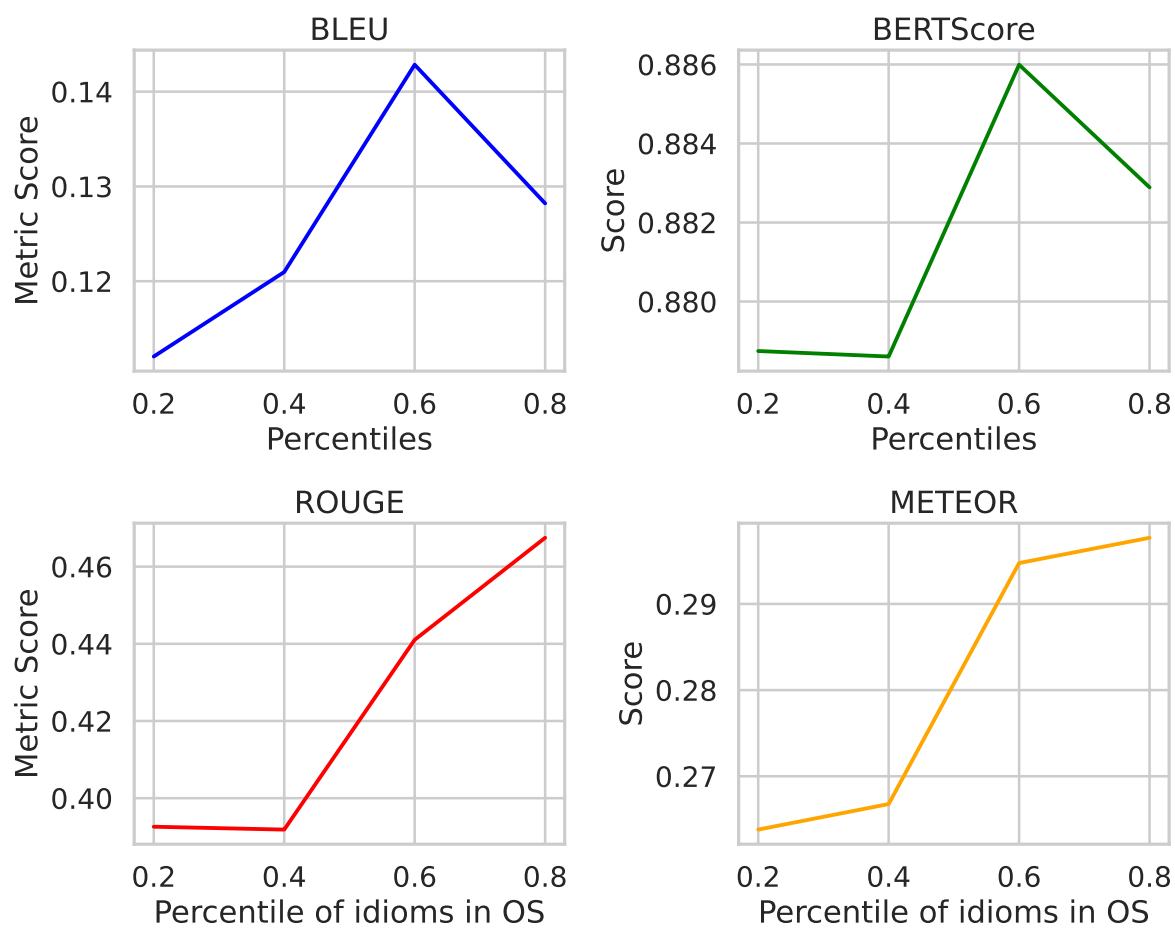


Figure F.3: Effect of idiom frequency on translations by Google in French

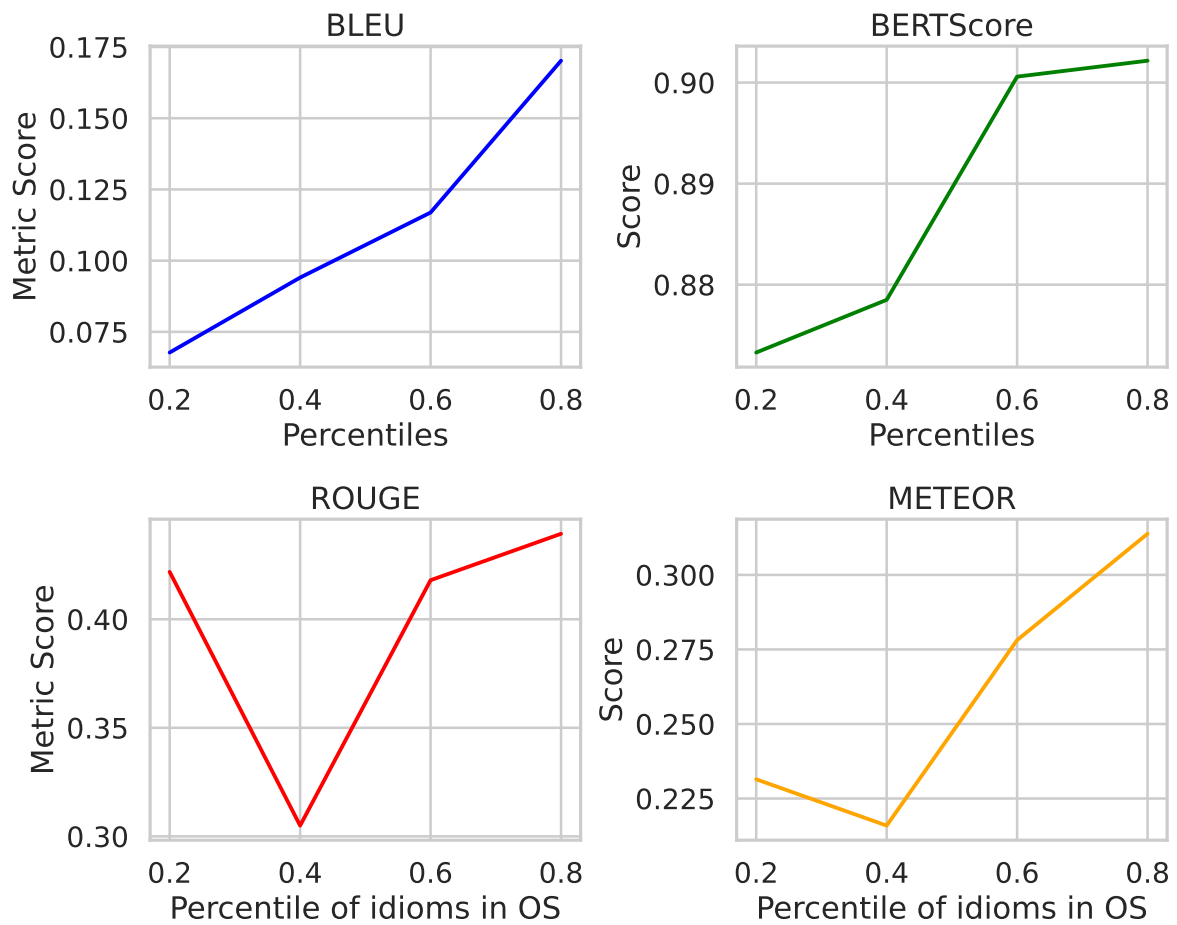


Figure F.4: Effect of idiom frequency on translations by DeepL in Japanese

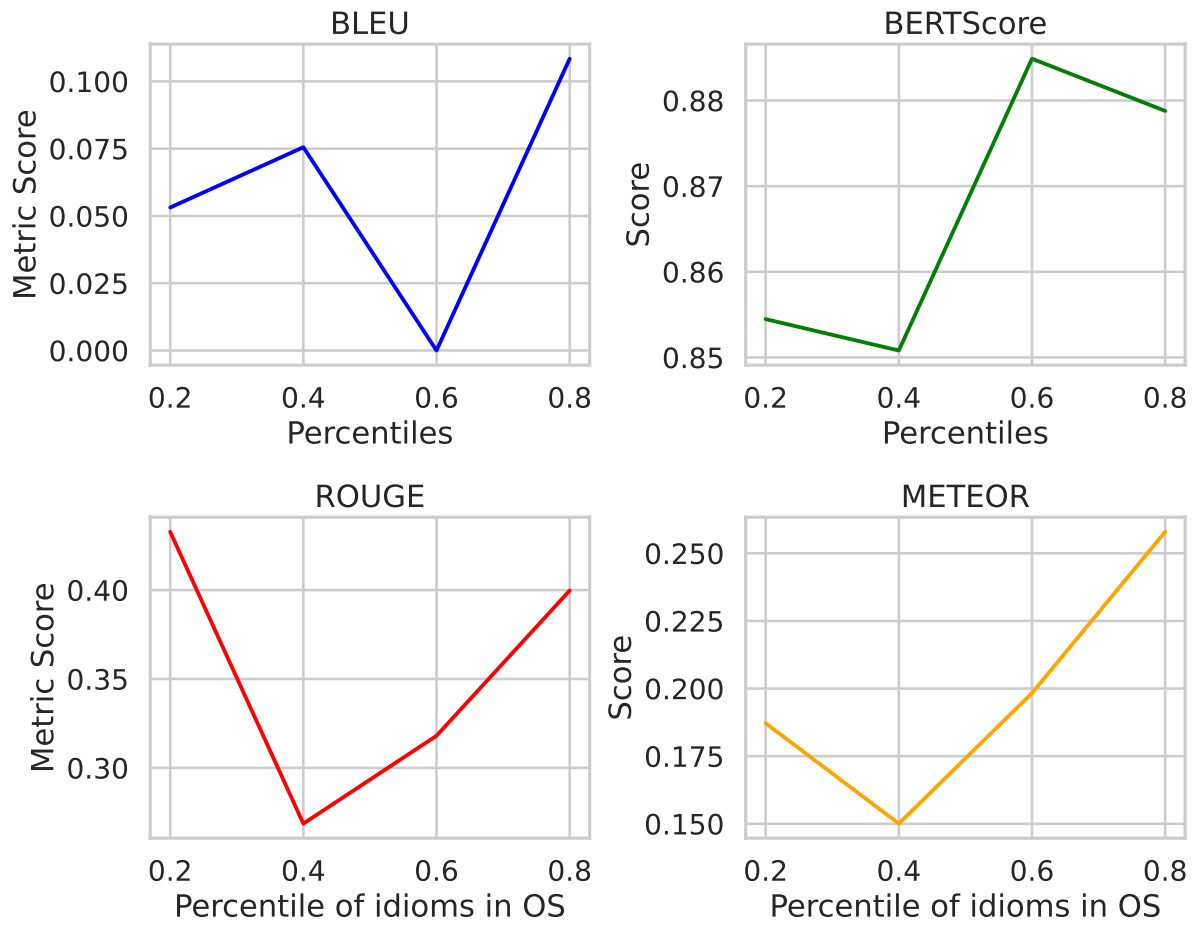


Figure F.5: Effect of idiom frequency on translations by Google in Japanese

F.9 Effect of Frequency on DeltaLM

The effect of frequency on Δ LM for Finnish and Japanese is shown in [Figure F.6](#) and [Figure F.7](#)

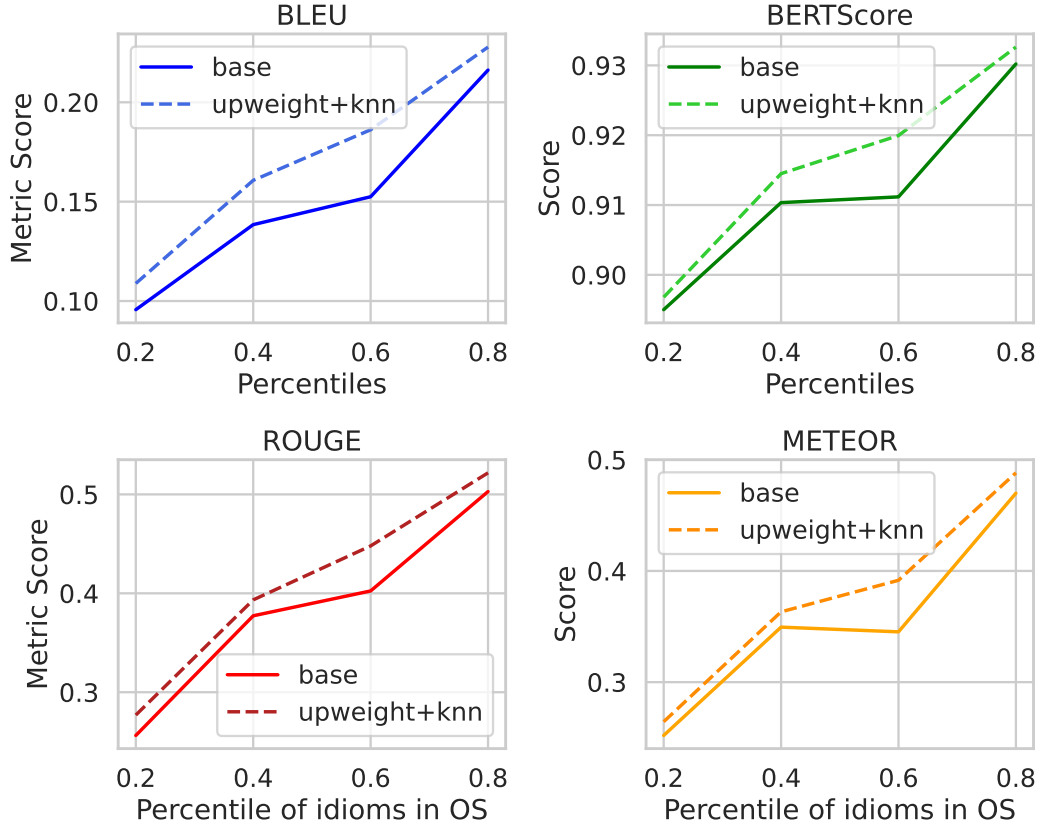


Figure F.6: Translation metrics for Finnish idiomatic sentences, for base and upweight+knn models.

F.10 Automatic Metrics in Detail

Exact results for automatic metrics are shown in [Table F.3](#).

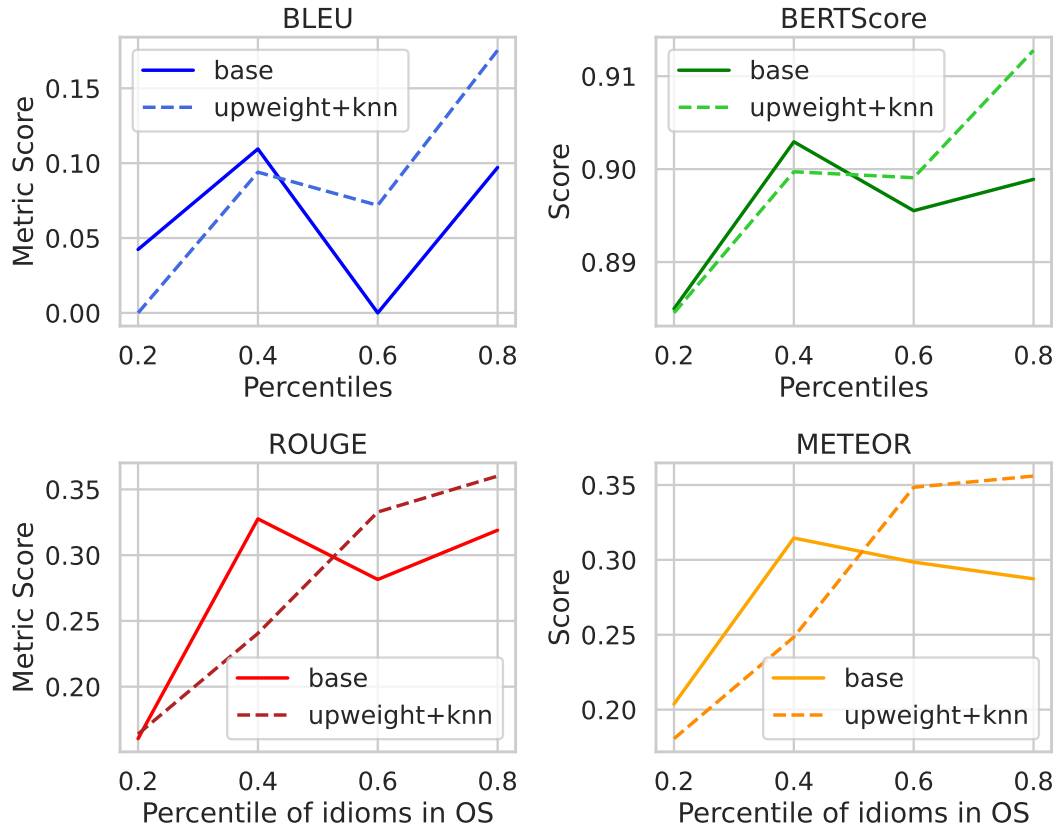


Figure F.7: Translation metrics for Japanese idiomatic sentences, for base and upweight+knn models.

Language	Test Set	Model	BLEU	RougeL-sum	BertScore	Meteor
fi	Idioms	Normal	0.1608	0.3737	0.9126	0.3592
		Knn	0.179	0.3904	0.9152	0.3745
		Upweight	0.1773	0.3895	0.9154	0.3709
		Knn + upweight	0.182	0.4027	0.9172	0.3833
	Literal	Normal	0.2093	0.5053	0.9350	0.5050
		Knn	0.2257	0.535	0.937	0.5161
		Upweight	0.2314	0.5281	0.9364	0.5258
		Knn + upweight	0.2362	0.5335	0.9387	0.5217
	Random-in	Normal	0.2056	0.5044	0.9321	0.4739
		Knn	0.1991	0.5067	0.9316	0.4732
		Upweight	0.2044	0.5048	0.9324	0.4728
		Knn + upweight	0.2120	0.5224	0.9351	0.488
	Random-out	Normal	0.2365	0.535	0.9145	0.4971
		Knn	0.2557	0.5602	0.9183	0.5178
		Upweight	0.2374	0.5458	0.9163	0.5102
		Knn + upweight	0.2498	0.564	0.9193	0.5261
fr	Idioms	Normal	0.2001	0.4329	0.9211	0.4393
		Knn	0.2154	0.4547	0.9235	0.4493
		Upweight	0.2174	0.4398	0.9235	0.4419
		Knn + upweight	0.2309	0.4568	0.926	0.4581
	Literal	Normal	0.2778	0.5261	0.9377	0.5504
		Knn	0.2824	0.5318	0.9387	0.5544
		Upweight	0.2923	0.5332	0.9384	0.5564
		Knn + upweight	0.2883	0.5325	0.9399	0.5549
	Random-in	Normal	0.3197	0.5942	0.9443	0.5803
		Knn	0.3142	0.5949	0.9438	0.5774
		Upweight	0.3079	0.5884	0.9427	0.5715
		Knn + upweight	0.3258	0.6036	0.9461	0.5865
	Random-out	Normal	0.2778	0.528	0.9377	0.5504
		Knn	0.3396	0.6573	0.928	0.6146
		Upweight	0.3078	0.5458	0.9223	0.5856
		Knn + upweight	0.3353	0.652	0.9272	0.6075
ja	Idioms	Normal	0.09048	0.2826	0.9234	0.2998
		Knn	0.09376	0.2767	0.9034	0.2905
		Upweight	0.09505	0.285	0.9052	0.3022
		Knn + upweight	0.09589	0.2841	0.905	0.2982
	Literal	Normal	0.1416	0.3951	0.9222	0.4222
		Knn	0.1418	0.3947	0.9223	0.4183
		Upweight	0.1427	0.3923	0.9221	0.4222
		Knn + upweight	0.1509	0.3962	0.9226	0.4228
	Random-in	Normal	0.1443	0.4008	0.9212	0.3927
		Knn	0.1535	0.4036	0.9218	0.3960
		Upweight	0.1430	0.3981	0.9212	0.3925
		Knn + upweight	0.1543	0.407	0.9224	0.4009
	Random-out	Normal	0.0948	0.3539	0.8946	0.3436
		Knn	0.09051	0.3389	0.8934	0.3284
		Upweight	0.09228	0.3522	0.8947	0.3392

BIBLIOGRAPHY

- [1] 01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2025. URL <https://arxiv.org/abs/2403.04652>.
- [2] Marah Abdin, Jyoti Aneja, Sébastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Mojan Javaheripi, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacrose, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. Phi-2: The surprising power of small language models. Microsoft Research Blog, 2023. Published December 12, 2023. Available at: <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/>.
- [3] Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. Metaphors in pre-trained language models: Probing and generalization across datasets and languages, 2022. URL <https://arxiv.org/abs/2203.14139>.
- [4] Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Sharma. No more beating about the bush : A step towards idiom handling for Indian language NLP. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL <https://aclanthology.org/L18-1048/>.
- [5] Meta AI. Llama 3: Advancing foundation models. <https://ai.facebook.com/blog/llama-3>, 2023. Accessed: 2025-02-21.
- [6] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou.

- What learning algorithm is in-context learning? investigations with linear models, 2023.
- [7] Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. Efficient online data mixing for language model pre-training, 2023. URL <https://arxiv.org/abs/2312.02406>.
 - [8] Nura Aljaafari, Danilo S. Carvalho, and André Freitas. Trace: Training and inference-time interpretability analysis for language models. *arXiv preprint arXiv:2507.03668*, 2025. URL: <https://arxiv.org/abs/2507.03668>.
 - [9] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. Smollm - blazingly fast and remarkably powerful, 2024. URL <https://huggingface.co/blog/smollm>.
 - [10] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*, 2023. doi: 10.48550/arXiv.2309.14316. URL <https://arxiv.org/abs/2309.14316>.
 - [11] Jacob Andreas. Measuring compositionality in representation learning. In *7th International Conference on Learning Representations*. ICLR, 2019.
 - [12] Anthropic Fellows and Decode Research. Open-sourcing circuit-tracing tools for large language model interpretability. Anthropic research blog post (with accompanying open-source library and Neuronpedia frontend), May 2025. Release includes Python ‘circuit-tracer’ library for attribution graph generation and the Neuronpedia visualizer.
 - [13] Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code, or not to code? exploring impact of code in pre-training, 2024. URL <https://arxiv.org/abs/2408.10914>.
 - [14] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL <https://arxiv.org/abs/2309.16609>.
 - [15] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Das-Sarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
 - [16] Christos Baziotis, Prashant Mathur, and Eva Hasler. Automatic evaluation and analysis of idioms in neural machine translation, 2022. URL <https://arxiv.org/abs/2210.04545>.

- [17] Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. A corpus of non-native written English annotated for metaphor. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2014. URL <https://aclanthology.org/N18-2014/>.
- [18] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3615–3620, 2020. doi: 10.18653/v1/2020.emnlp-demos.82. URL <https://aclanthology.org/2020.emnlp-demos.82>.
- [19] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL <https://aclanthology.org/2020.acl-main.463/>.
- [20] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 41–48, 2009. URL <https://doi.org/10.1145/1553374.1553380>.
- [21] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
- [22] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
- [23] Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. Establishing task scaling laws via compute-efficient model ladders, 2024. URL <https://arxiv.org/abs/2412.04403>.
- [24] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.
- [25] Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. Bracketing Guidelines for Treebank II Style Penn Treebank Project, 1995.
- [26] BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni,

François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klam, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Froberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Noumane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névél, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey

Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Cliniciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguiet, Thanh Le, Tobi Oyeade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Perinán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Ji Hyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sängler, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljevic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhan-shu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aaronsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL <https://arxiv.org/abs/2211.05100>.

- [27] Yuri Bizzoni and Mehdi Ghanimifard. Bigrams and BiLSTMs two neural networks for sequential metaphor detection. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee, editors, *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101, New

- Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0911. URL <https://aclanthology.org/W18-0911/>.
- [28] Yuri Bizzoni and Shalom Lappin. Predicting human metaphor paraphrase judgments with deep neural networks. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee, editors, *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0906. URL <https://aclanthology.org/W18-0906/>.
- [29] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. Gpt-neo: Large scale autoregressive language modeling with Mesh-TensorFlow. Zenodo, March 2021. URL <https://doi.org/10.5281/zenodo.5297715>. EleutherAI.
- [30] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. Gpt-neo: Large scale autoregressive language modeling with Mesh-TensorFlow. Zenodo, March 2021. URL <https://doi.org/10.5281/zenodo.5297715>. EleutherAI.
- [31] Cody Blakeney, Mansheej Paul, Brett W. Larsen, Sean Owen, and Jonathan Frankle. Does your data spark joy? Performance gains from domain upsampling at the end of training, June 2024. URL <http://arxiv.org/abs/2406.03476>. arXiv:2406.03476 [cs].
- [32] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
- [33] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- [34] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [35] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [36] Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws, 2023. URL <https://arxiv.org/abs/2210.14891>.

- [37] Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. Towards best practice for multiword expressions in computational lexicons. In Manuel González Rodríguez and Carmen Paz Suarez Araujo, editors, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA). URL <https://aclanthology.org/L02-1259/>.
- [38] Robyn Carston and Catherine Wearing. Metaphor, hyperbole and simile: A pragmatic approach. *Language and Cognition*, 3(2):283–312, 2011. doi: 10.1515/langcog.2011.010.
- [39] Santiago Castro. Fast Krippendorff: Fast computation of Krippendorff’s alpha agreement measure. <https://github.com/pln-fing-udelar/fast-krippendorff>, 2017.
- [40] Cerebras Systems. Cerebras-gpt: A family of open, compute-efficient large language models, 2023.
- [41] Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. It’s not rocket science : Interpreting figurative language in narratives. *ArXiv*, abs/2109.00087, 2021.
- [42] Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. Figurative language in recognizing textual entailment. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.297. URL <https://aclanthology.org/2021.findings-acl.297/>.
- [43] Chameleon Team. Chameleon: Mixed-Modal Early-Fusion Foundation Models, May 2024. URL <http://arxiv.org/abs/2405.09818>. arXiv:2405.09818 [cs].
- [44] Lili Chen, Alex Ippolito, Kyunghyun Chen, Katerina Fragkiadaki, and Hao Liu. Self-questioning language models. *arXiv preprint arXiv:2508.03682*, 2024.
- [45] Mark Chen, Jacob Tworek, et al. Evaluating large language models trained on code, 2021.
- [46] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, page 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL <http://dx.doi.org/10.1145/2939672.2939785>.
- [47] Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. Parallel structures in pre-training data yield in-context learning, 2024. URL <https://arxiv.org/abs/2402.12530>.
- [48] Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. Instruction pre-training: Language models are supervised multitask learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2550, 2024. doi: 10.18653/v1/2024.emnlp-main.148. URL

<https://aclanthology.org/2024.emnlp-main.148>.

- [49] Noam Chomsky. On certain formal properties of grammars. *Information and Control*, 2(2):137–167, 1959. ISSN 0019-9958. doi: [https://doi.org/10.1016/S0019-9958\(59\)90362-6](https://doi.org/10.1016/S0019-9958(59)90362-6). URL <https://www.sciencedirect.com/science/article/pii/S0019995859903626>.
- [50] Peter Clark et al. Arc: A large-scale challenge dataset for ai2 reasoning. In *Proceedings of EMNLP*, 2018.
- [51] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [52] Karl Cobbe, Vineet Kosaraju, et al. Training verifiers to solve math word problems, 2021.
- [53] Alexis Conneau et al. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, 2018.
- [54] Thomas C. Cooper. Processing of idioms by l2 learners of english. *TESOL Quarterly*, 33:233–262, 1999.
- [55] Thomas C. Cooper. Processing of Idioms by L2 Learners of English. *TESOL Quarterly*, 33(2):233–262, 1999. doi: <https://doi.org/10.2307/3587719>. URL <https://onlinelibrary.wiley.com/doi/abs/10.2307/3587719>.
- [56] MLJ Crawford, RS Harwerth, EL Smith III, and GK von Noorden. The effects of short-term experimental strabismus on the visual system in macaca mulatta. *Investigative Ophthalmology & Visual Science*, 37(3):S664, 1996.
- [57] Verna Dankers, Elia Bruni, and Dieuwke Hupkes. The paradox of the compositionality of natural language: A neural machine translation case study. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.286. URL <https://aclanthology.org/2022.acl-long.286/>.
- [58] Verna Dankers, Christopher Lucas, and Ivan Titov. Can transformer be too compositional? analysing idiom processing in neural machine translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.252. URL <https://aclanthology.org/2022.acl-long.252/>.
- [59] DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiusi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo

Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024. URL <https://arxiv.org/abs/2401.02954>.

- [60] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojuan Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhi-

- hong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2024. URL <https://arxiv.org/abs/2412.19437>.
- [61] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- [62] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [63] Nolan Dey, Daria Soboleva, Faisal Al-Khateeb, Bowen Yang, Ribhu Pathria, Hemant Khachane, Shaheer Muhammad, Zhiming Chen, Robert Myers, Jacob Robert Steeves, Natalia Vassilieva, Marvin Tom, and Joel Hestness. Btlm-3b-8k: 7b parameter performance in a 3b parameter model, 2023. URL <https://arxiv.org/abs/2309.11568>.
- [64] Fernando Diaz and Michael Madaio. Scaling laws do not scale, 2024. URL <https://arxiv.org/abs/2307.03201>.
- [65] Qingxiu Dong, Li Dong, Yao Tang, Tianzhu Ye, Yutao Sun, Zhifang Sui, and Furu Wei. Reinforcement pre-training. *arXiv preprint*, arXiv:2506.08007, June 2025. Submitted on 9 June 2025.
- [66] Igor Douven. Abduction. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.
- [67] Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online (v2020.3)*. Zenodo, 2013. doi: 10.5281/zenodo.7385533. URL <https://doi.org/10.5281/zenodo.7385533>.
- [68] Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. Back to square one: Artifact detection, training and commonsense disentanglement in the winograd schema, 2021. URL <https://arxiv.org/abs/2104.08161>.
- [69] Jeffrey L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993. doi: 10.1016/0010-0277(93)90058-4.
- [70] Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL <https://aclanthology.org/D19-1006/>.

- [71] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Examining the tip of the iceberg: A data set for idiom translation. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL <https://aclanthology.org/L18-1148/>.
- [72] Gilles Fauconnier and Mark Turner. Conceptual blending, form and meaning. 2003.
- [73] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.
- [74] Steven Feng, Shrimai Prabhumoye, Kezhi Kong, Dan Su, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Maximize Your Data’s Potential: Enhancing LLM Accuracy with Two-Phase Pretraining, December 2024. URL <http://arxiv.org/abs/2412.15285>. arXiv:2412.15285 [cs].
- [75] Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. Scaling laws for multilingual neural machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10053–10071. PMLR, 2023.
- [76] Allen Institute for AI. Olmo: Open language model. <https://huggingface.co/allenai/OLMo-7B-hf>, 2024.
- [77] George Foster, Cyril Goutte, and Roland Kuhn. Discriminative instance weighting for domain adaptation in statistical machine translation. In Hang Li and Lluís Màrquez, editors, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA, October 2010. Association for Computational Linguistics. URL <https://aclanthology.org/D10-1044/>.
- [78] H. Frankfurt. Peirce’s notion of abduction. *Journal of Philosophy*, 55:593–596, 1958.
- [79] Elias Frantar, Carlos Riquelme, Neil Houlsby, Dan Alistarh, and Utku Evci. Scaling Laws for Sparsely-Connected Foundation Models. *International Conference on Learning Representations (ICLR)*, 2023. URL <https://openreview.net/forum?id=i9K2ZWkYIP>.
- [80] Susan Fussell and Mallie Moss. Figurative language in emotional communication. 08 2008.
- [81] Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. Neural metaphor detection in context. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1060. URL <https://aclanthology.org/D18-1060/>.

- [82] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020.
- [83] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL <https://zenodo.org/records/10256836>.
- [84] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes, 2023.
- [85] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. URL <https://arxiv.org/abs/2403.08295>.
- [86] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier

Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshv, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL <https://arxiv.org/abs/2408.00118>.

- [87] Xinyang Geng and Hao Liu. Openllama: An open reproduction of LLaMA. https://github.com/openlm-research/open_llama, May 2023. URL https://github.com/openlm-research/open_llama. GitHub repository; accessed 2025-08-12.
- [88] D Gentner and B Bowdle. Metaphor as structure-mapping. In Jr R.W. Gibbs, editor, *The Cambridge handbook of metaphor and thought*, pages 109–128. Cambridge University Press, 2008.
- [89] Dedre Gentner. Structure-mapping: A theoretical framework for analogy*. *Cognitive Science*, 7(2):155–170, 1983. doi: <https://doi.org/10.1207/s15516709cog0702>

- _3. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0702_3.
- [90] Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation, 2021. URL <https://arxiv.org/abs/2109.07740>.
- [91] Kahlil Gibran. *Sand and Foam; a book of aphorisms*. A.A Knopf, 1926.
- [92] Sam Glucksberg. The psycholinguistics of metaphor. *Trends in cognitive sciences*, 7: 92–96, 03 2003. doi: 10.1016/S1364-6613(02)00040-2.
- [93] Yoav Goldberg and Jon Orwant. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL <https://aclanthology.org/S13-1035>.
- [94] Alison Gopnik and Andrew N Meltzoff. *Words, thoughts, and theories*. MIT Press, Cambridge, MA, 1997.
- [95] Sachin Goyal, Pratyush Maini, Zachary C. Lipton, Aditi Raghunathan, and J. Zico Kolter. Scaling laws for data filtering – data curation cannot be compute agnostic, 2024. URL <https://arxiv.org/abs/2404.07177>.
- [96] Yvette Graham, Nikita Mathur, and Timothy Baldwin. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274. Association for Computational Linguistics, 2014.
- [97] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. URL <https://arxiv.org/abs/2312.00752>.
- [98] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023. URL <https://arxiv.org/abs/2306.11644>.
- [99] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [100] Yiduo Guo, Jie Fu, Huishuai Zhang, Dongyan Zhao, and Yikang Shen. Efficient continual pre-training by mitigating the stability gap. *arXiv preprint arXiv:2406.14833*, 2024.
- [101] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

- [102] Hessel Haagsma, Johan Bos, and Malvina Nissim. MAGPIE: A large corpus of potentially idiomatic expressions. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.35/>.
- [103] Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. Understanding transformer memorization recall through idioms, 2022. URL <https://arxiv.org/abs/2210.03588>.
- [104] James Hawthorne. Inductive Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.
- [105] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=XPZlaotutsD>.
- [106] Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with apps. In *Advances in Neural Information Processing Systems (NeurIPS) 2021*, 2021. URL <https://proceedings.neurips.cc/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper.pdf>.
- [107] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.
- [108] Dan Hendrycks et al. Measuring massive multitask language understanding, 2020.
- [109] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling Laws for Transfer. *arXiv preprint arXiv:2102.01293*, 2021. URL <https://arxiv.org/pdf/2102.01293>.
- [110] Wan Yu Ho, Christine Kng, Shan Wang, and Francis Bond. Identifying idioms in Chinese translations. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 716–721, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL <https://aclanthology.org/L14-1391/>.
- [111] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den

- Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL <https://arxiv.org/abs/2203.15556>.
- [112] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *NeurIPS*, 2022. arXiv preprint arXiv:2203.15556.
 - [113] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.
 - [114] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.
 - [115] Shengding Hu, Zepeng Bai, Tiannan Chen, Siran Jiang, Xin Jiang, Chang Liu, Wenhao Li, Guanqun Lai, Zhiyang Cui, Chengyi Zhang, Ruibin Zhao, Junhao Han, Shaokun Dong, Junlong Fang, Boyu Shi, Ting Wang, Zhiyuan Liu, Zhixing Tang, Lei Li, Benyou Wu, Tao Wang, Chenghu Zheng, Tianxiang Zhao, Zhiyi Zhang, Wanli Fu, Bowen Dai, Jiaqi Zhou, Renze Li, Zijun Zheng, Haoran Xu, Xiao Sun, Bailing Zhou, Shuai Jiao, Jian Li, Bowen Cao, Xuming Zhao, Yikang Lu, Zixiang Qi, Jun Shi, Haonan Xiang, Jinhua Wu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.
 - [116] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies, April 2024. URL <http://arxiv.org/abs/2404.06395>. arXiv:2404.06395 [cs].
 - [117] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? (extended abstract). In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 5065–5069. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/708. URL <https://doi.org/10.24963/ijcai.2020/708>. Journal track.
 - [118] Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019.
 - [119] Technology Innovation Institute. The falcon series of open language models, 2023.
 - [120] Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vas-

- silvitskii, and Sanmi Koyejo. Scaling laws for downstream task performance of large language models. *arXiv preprint arXiv:2402.04177*, 2024.
- [121] Arthur Jampolsky. Unequal visual inputs and strabismus management: A comparison of human and animal strabismus. In *Symposium on Strabismus: Transactions of the New Orleans Academy of Ophthalmology*, chapter 26, pages 358–492. Mosby, St. Louis, 1978.
- [122] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. Investigating robustness of dialog models to popular figurative language constructs. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.592. URL <https://aclanthology.org/2021.emnlp-main.592/>.
- [123] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mixtral of experts, 2024. URL <https://arxiv.org/abs/2401.04088>.
- [124] Yiding Jiang, Allan Zhou, Zhili Feng, Sadhika Malladi, and J. Zico Kolter. Adaptive data optimization: Dynamic sample selection with scaling laws, 2024. URL <https://arxiv.org/abs/2410.11820>.
- [125] Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Lin, Wen-tau Yih, and Srinivasan Iyer. Instruction-tuned language models are better knowledge learners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5421–5434, 2024. doi: 10.18653/v1/2024.acl-long.296. URL <https://aclanthology.org/2024.acl-long.296>.
- [126] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *arXiv preprint arXiv:2310.06770*, 2023. URL <https://arxiv.org/abs/2310.06770>. Introduces SWE-bench: 2,294 real GitHub issue-resolution tasks.
- [127] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=VTF8yNQM66>.
- [128] Jacqueline S Johnson and Elissa L Newport. Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive Psychology*, 21(1):60–99, 1989.

- [129] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.
- [130] Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Indra Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.
- [131] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge, 2022.
- [132] Sara Kangaslahti, Elan Rosenfeld, and Naomi Saphra. Hidden breakthroughs in language model training. arXiv preprint arXiv:2506.15872, 2025. URL <https://arxiv.org/abs/2506.15872>. Submitted June 18, 2025.
- [133] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL <https://arxiv.org/abs/2001.08361>.
- [134] Annette Karmiloff-Smith. *Beyond Modularity: A Developmental Perspective on Cognitive Science*. MIT Press, Cambridge, MA, 1992.
- [135] Annette Karmiloff-Smith. *Beyond Modularity: A Developmental Perspective on Cognitive Science*. MIT Press, Cambridge, MA, 1992.
- [136] Nora Kassner, Benno Krojer, and Hinrich Schütze. Are Pretrained Language Models Symbolic Reasoners over Knowledge? *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 2020. URL <https://aclanthology.org/2020.conll-1.45/>.
- [137] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.
- [138] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation, 2021.
- [139] Najoung Kim and Tal Linzen. Cogs: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.731.
- [140] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- [141] Eric I Knudsen. Sensitive periods in the development of the brain and behavior. *Journal of Cognitive Neuroscience*, 16(8):1412–1425, 2004.
- [142] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke

- Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [143] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019. URL <https://proceedings.mlr.press/v97/kornblith19a.html>.
- [144] Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30:61 – 70, 1970. URL <https://api.semanticscholar.org/CorpusID:144036366>.
- [145] Kundan Krishna, Saurabh Garg, Jeffrey Bigham, and Zachary Lipton. Downstream datasets make surprisingly good pretraining corpora. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12207–12222, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.682. URL <https://aclanthology.org/2023.acl-long.682/>.
- [146] Patricia K Kuhl. Learning and representation in speech and language. *Current Opinion in Neurobiology*, 4(6):812–822, 1994.
- [147] Tanishq Kumar, Zachary Ankner, Benjamin F. Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision, 2024. URL <https://arxiv.org/abs/2411.04330>.
- [148] Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, 2017. URL <https://api.semanticscholar.org/CorpusID:46761158>.
- [149] Brenden M. Lake and Marco Baroni. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *ArXiv*, abs/1711.00350, 2017.
- [150] Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2879–2888. PMLR, 2018. doi: 10.48550/arXiv.1711.00350.
- [151] Brenden M. Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. In *Annual Meeting of the Cognitive Science Society*, 2019. URL <https://api.semanticscholar.org/CorpusID:58006558>.
- [152] G. Lakoff and M. Johnson. *Metaphors we Live By*. University of Chicago Press, 1981.
- [153] Cristine H. Legare. The contributions of explanation and exploration to children’s scientific reasoning. *Child Development Perspectives*, 8(2):101–106, 2014. doi: 10.1111/cdep.12070.

- [154] Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. A report on the 2018 VUA metaphor detection shared task. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee, editors, *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0907. URL <https://aclanthology.org/W18-0907/>.
- [155] Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. A report on the 2020 VUA and TOEFL metaphor detection shared task. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna Feldman, and Debanjan Ghosh, editors, *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.figlang-1.3. URL <https://aclanthology.org/2020.figlang-1.3/>.
- [156] Lori Levin, Teruko Mitamura, Brian MacWhinney, Davida Fromm, Jaime Carbonell, Weston Feely, Robert Frederking, Anatole Gershman, and Carlos Ramirez. Resources for the detection of conventionalized metaphors in four languages. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 498–501, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL <https://aclanthology.org/L14-1387/>.
- [157] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.733. URL <https://aclanthology.org/2020.emnlp-main.733/>.
- [158] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruva Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Benjamin Recht, Luke Zettlemoyer, Sham Iyer, Tianyi Zhuang, Percy Liang, Alexander Rush, Naveen Jain, Vikas Raunak, Claire Cardie, and Ludwig Schmidt. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024.
- [159] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier,

- Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Naman Rao, Roza Stojnic, Miltiadis Allamanis, Pierre Laffitte, Pawan Kumar Rustamov, Kaspar Valter, Chenyu Mittal, Komila Khamidov Tesfamikael, Christopher Murati, Sacha Lee, Albert Q. Wan, Armel Suharyanto, Juliette Copet, David R. So, Leandro Kolubako, Gabriel Machado Pina, Soroush Behtash, Denis Moskovskiy, Deepan Siddarth, Nicolas Luce-Rainville, Mostafa Dehghani, Marcin Szafraniec, Piotr Cardozo, Jenia Jitsev, Ekaterina Kochmar, Antonio Torralba, Dragomir Radev, Alexander M. Rush, Preslav Nakov, Thomas Wang, Wolf Zuo, Hugh Echikson, Laura Schuelke, Jeffrey Carmichael, Kumar Shridhar Sadagopan, Zhuang Ling, Colin Kwiatkowski, Andy Lohn, Jonas Mueller, and Harm de Vries Floetenmeyer. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*, 2023.
- [160] Yanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*, 2023.
 - [161] Lightning AI. Litgpt. <https://github.com/Lightning-AI/litgpt>, 2023.
 - [162] Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. Riddle-sense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge, 2021. URL <https://arxiv.org/abs/2101.00376>.
 - [163] Stephanie Lin et al. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of EMNLP*, 2021.
 - [164] Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*, 2024.
 - [165] Peter Lipton. *Inference to the Best Explanation*. Routledge, 2001.
 - [166] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL <https://aclanthology.org/L18-1275>.
 - [167] Dayan Liu. Translation and culture: Translating idioms between english and chinese from a cultural perspective. *Theory and Practice in Language Studies*, 2:2357–2362, 2012.
 - [168] Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training, 2025. URL <https://arxiv.org/abs/2407.01492>.
 - [169] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
 - [170] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer

- Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL <http://arxiv.org/abs/1907.11692>.
- [171] Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. Llm360: Towards fully transparent open-source llms, 2023. URL <https://arxiv.org/abs/2312.06550>.
- [172] Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463, 2014.
- [173] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ip-polito. A Pretrainer’s Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2024. URL <https://aclanthology.org/2024.naacl-long.179/>.
- [174] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *CoRR*, cs.CL/0205028, 2002. URL <http://dblp.uni-trier.de/db/journals/corr/corr0205.html#cs-CL-0205028>.
- [175] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- [176] Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. Wangchanberta: Pretraining transformer-based thai language models, 2021.
- [177] Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders, 2021. URL <https://arxiv.org/abs/2106.13736>.
- [178] Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help llms reasoning?, 2023. URL <https://arxiv.org/abs/2309.16298>.
- [179] Rui Mao, Chenghua Lin, and Frank Guerin. Word embedding and WordNet based metaphor identification and interpretation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia, July

2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1113. URL <https://aclanthology.org/P18-1113/>.
- [180] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The Penn Treebank, 1993.
- [181] Evgeny Matusov. The challenges of using neural machine translation for literature. In James Hadley, Maja Popović, Haithem Afli, and Andy Way, editors, *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland, August 2019. European Association for Machine Translation. URL <https://aclanthology.org/W19-7302/>.
- [182] Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn’t better, 2024. URL <https://arxiv.org/abs/2306.09479>.
- [183] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *Journal of Machine Learning Research*, 24(214):1–50, 2023. URL <http://jmlr.org/papers/v24/22-0496.html>.
- [184] Meta AI. Opt: Open pre-trained transformer language models, 2022.
- [185] Meta AI. Xglm: Cross-lingual generative language models, 2022.
- [186] J. S Mio and A. N Katz. *Metaphor: Implications and Applications*. Psychology Press, 1996.
- [187] Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Öncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *ICLR 2025 Workshop*, 2024. arXiv preprint arXiv:2410.05229.
- [188] Saif Mohammad, Ekaterina Shutova, and Peter Turney. Metaphor as a medium for emotion: An empirical study. In Claire Gardent, Raffaella Bernardi, and Ivan Titov, editors, *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-2003. URL <https://aclanthology.org/S16-2003/>.
- [189] MosaicML. Mpt-7b: A new standard for open-source, commercially usable llms. <https://huggingface.co/mosaicml/mpt-7b>, 2023.
- [190] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. URL <https://openreview.net/pdf?id=j5BuTrEj35>.

- [191] Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024.
- [192] Makoto Nagao. A framework of a mechanical translation between japanese and english by analogy principle. In *Proc. of the International NATO Symposium on Artificial and Human Intelligence*, page 173–180, USA, 1984. Elsevier North-Holland, Inc. ISBN 0444865454.
- [193] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL <https://arxiv.org/abs/2301.05217>.
- [194] Elissa L Newport, Daphne Bavelier, and Helen J Neville. Critical thinking about critical periods: Perspectives on a critical period for language acquisition. In Emmanuel Dupoux, editor, *Language, brain and cognitive development: Essays in honor of Jacques Mehler*, pages 481–502. MIT Press, Cambridge, MA, 2001.
- [195] Yixin Nie, Jing Wang, Mohit Bansal, and Kai-Wei Chang. Adversarial natural language inference. In *Proceedings of ACL*, 2020.
- [196] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.
- [197] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis, 2023.
- [198] OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hananeh Hajishirzi. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025.
- [199] Abdulfattah Omar and Yasser Gomaa. The machine translation of literature: Implications for translation pedagogy. *International Journal of Emerging Technologies in Learning (ijET)*, 15:228, 06 2020. doi: 10.3991/ijet.v15i11.13275.
- [200] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks,

Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brit-tany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecofet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gib-son, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gor-don, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Kon-stantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McK-inney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vin-nie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rameev Nayak, Arvind Nee-lakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascan-dolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schul-man, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Kata-rina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Fe-lipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder,

- Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.
- [201] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling, 2019.
 - [202] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.
 - [203] Susan Oyama. A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, 5(3):261–283, 1976.
 - [204] Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. Training software engineering agents and verifiers with swe-gym. In *International Conference on Learning Representations (ICLR) 2025*, 2025. URL <https://arxiv.org/abs/2412.21139>. ICLR 2025, public release; SWE-Gym environment with 2,438 real-world tasks.
 - [205] Denis Paperno et al. Lambada: Word prediction requiring a broad discourse context. In *Proceedings of ACL (Long Papers)*, 2016.
 - [206] Tim Pearce and Jinyeop Song. Reconciling kaplan and chinchilla scaling laws, 2024. URL <https://arxiv.org/abs/2406.12907>.
 - [207] Karl Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.
 - [208] Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. A howling success or a working sea? testing what BERT knows about metaphors. In Jasmin Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors, *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.blackboxnlp-1.13. URL <https://aclanthology.org/2021.blackboxnlp-1.13/>.
 - [209] Charles Sanders Peirce. *Collected Papers of Charles Sanders Peirce, Volume 5*, volume 5. Harvard University Press, 1965. URL <http://www.hup.harvard.edu/catalog.php?isbn=9780674138001>. Accessed: [insert current date].
 - [210] Francis Jeffry Pelletier. The principle of semantic compositionality. *Topoi*, 13:11–24, 1994.
 - [211] Jean Piaget. *The Origins of Intelligence in Children*. International Universities Press, New York, 1952. Originally published as *La naissance de l’intelligence chez l’enfant*

- (1936).
- [212] Jean Piaget. *The Construction of Reality in the Child*. Basic Books, New York, 1954. English translation of *La construction du réel chez l'enfant* (1937).
 - [213] Thierry Poibeau. On "human parity" and "super human performance" in machine translation evaluation. Marseille, France, June 2022. Language Resource and Evaluation Conference.
 - [214] Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. Collecting diverse natural language inference problems for sentence representation evaluation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1007. URL <https://aclanthology.org/D18-1007/>.
 - [215] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL <https://aclanthology.org/W15-3049>.
 - [216] Paul H. Portner. *What is meaning?: Fundamentals of formal semantics*. Blackwell Publishing, Malden, MA, 2005.
 - [217] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.
 - [218] Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations (ICLR) 2024 Poster Track*, 2024. URL <https://arxiv.org/abs/2402.14811>. arXiv:2402.14811.
 - [219] Malay Pramanick, Ashim Gupta, and Pabitra Mitra. An LSTM-CRF based approach to token-level metaphor detection. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee, editors, *Proceedings of the Workshop on Figurative Language Processing*, pages 67–75, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0908. URL <https://aclanthology.org/W18-0908/>.
 - [220] Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=R8sQPpGCv0>.
 - [221] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Compu-*

- tational Linguistics: System Demonstrations*, 2020. URL <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>.
- [222] Tianyi Qi, Yichao Ouyang, Yu Liang, Lifu Huang, Xiao Dang, Kevin Gimpel, and Noah A. Smith. Stanza: A python nlp toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, 2020.
 - [223] Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement, 2023.
 - [224] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
 - [225] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
 - [226] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
 - [227] Andrew Radford. *English Syntax: An Introduction*. Cambridge University Press, Cambridge, UK, 2004.
 - [228] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <https://arxiv.org/abs/1910.10683>.
 - [229] Carlos Ramisch, Silvio Cordeiro, and Aline Villavicencio. Filtering and measuring the intrinsic quality of human compositionality judgments. pages 32–37, 01 2016. doi: 10.18653/v1/W16-1804.
 - [230] Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2026. URL <https://aclanthology.org/P16-2026/>.
 - [231] Siva Reddy, Diana McCarthy, and Suresh Manandhar. An empirical study on compositionality in compound nouns. In Haifeng Wang and David Yarowsky, editors, *Proceedings of 5th International Joint Conference on Natural Language Process-*

- ing, pages 210–218, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing. URL <https://aclanthology.org/I11-1024/>.
- [232] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Unbabel’s participation in the WMT20 metrics shared task. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online, November 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.wmt-1.101/>.
 - [233] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL <https://arxiv.org/abs/2004.09813>.
 - [234] Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance, 2024. URL <https://arxiv.org/abs/2405.10938>.
 - [235] Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwarak Talupuru, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural knowledge in pretraining drives reasoning in large language models, 2024. URL <https://arxiv.org/abs/2411.12580>.
 - [236] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for nlp. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-45715-2.
 - [237] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, 2020.
 - [238] Rei Sakaguchi et al. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of ACL*, 2019.
 - [239] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.
 - [240] Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. Release of pre-trained models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13898–13905, 5 2024. URL <https://aclanthology.org/2024.lrec-main.1213>. <https://arxiv.org/abs/2404.01657>.
 - [241] William Saxton, Edward Grefenstette, et al. Mathqa: A challenge dataset for solv-

- ing math word problems. In *Proceedings of EMNLP*, 2019.
- [242] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage?, 2023. URL <https://arxiv.org/abs/2304.15004>.
 - [243] Laura E Schulz. The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, 16(7):382–389, 2012.
 - [244] Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. Evaluating machine translation performance on chinese idioms with a blacklist method, 2017. URL <https://arxiv.org/abs/1711.07646>.
 - [245] Stewart Shapiro and Teresa Kouri Kissel. Classical Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2024 edition, 2024.
 - [246] Lloyd S Shapley. A value for n-person games. In *Contributions to the Theory of Games*, pages 307–317. Princeton University Press, 1953.
 - [247] Ekaterina Shutova. Automatic metaphor interpretation as a paraphrasing task. In Ron Kaplan, Jill Burstein, Mary Harper, and Gerald Penn, editors, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037, Los Angeles, California, June 2010. Association for Computational Linguistics. URL <https://aclanthology.org/N10-1147/>.
 - [248] Ekaterina V. Shutova. Computational approaches to figurative language. Technical Report UCAM-CL-TR-803, University of Cambridge, Computer Laboratory, Cambridge, UK, August 2011. URL <https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-803.pdf>.
 - [249] Robert S. Siegler. *Emerging Minds: The Process of Change in Children’s Thinking*. Oxford University Press, New York, 1996. Introduces the “overlapping waves” theory.
 - [250] Sixth Conference on Machine Translation (WMT21). Shared task: Large-scale multilingual machine translation. <https://www.statmt.org/wmt21/large-scale-multilingual-translation-task.html>, 2021. URL <https://www.statmt.org/wmt21/large-scale-multilingual-translation-task.html>. Accessed 2023-05-29.
 - [251] Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latache, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bower, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Natalia Hübner, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti,

- Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoğlu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting 叶, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. Grambank reveals global patterns in the structural diversity of the world’s languages. *Science Advances*, 9, 2023. doi: 10.1126/sciadv.adg6175.
- [252] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. In *ICLR*, 2025. arXiv preprint arXiv:2408.03314.
- [253] Luca Soldaini and Kyle Lo. pes2o (pretraining efficiently on s2orc) dataset. Technical report, Allen Institute for AI, 2023.
- [254] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxu Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*, 2024. URL <https://arxiv.org/abs/2402.00159>.
- [255] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2023. URL <https://arxiv.org/abs/2206.14486>.
- [256] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey, 2022. URL <https://arxiv.org/abs/2101.10382>.
- [257] Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.
- [258] Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. *arXiv preprint arXiv:2503.19206*, 2025. URL <https://arxiv.org/abs/2503.19206>.
- [259] Gerard J Steen, Alleta G Dorst, Berenike Herrmann, Anna A Kaal, Tina Krenn-

- mayr, and Trynetje Pasma. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins, 2010.
- [260] Chang Su, Shuman Huang, and Yijiang Chen. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311, 2017. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2016.09.030>. URL <https://www.sciencedirect.com/science/article/pii/S0925231216310475>.
 - [261] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.
 - [262] Kaiser Sun and Mark Dredze. Amuro and char: Analyzing the relationship between pre-training and fine-tuning of large language models. *arXiv preprint arXiv:2408.06663*, 2024. doi: 10.48550/arXiv.2408.06663. URL <https://arxiv.org/abs/2408.06663>.
 - [263] Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. Salmon: Self-alignment with instructable reward models. In *The Twelfth International Conference on Learning Representations*, 2024.
 - [264] John Sweller. Discussion of ‘emerging topics in cognitive load research: Using learner and information characteristics in the design of powerful learning environments’. *Applied Cognitive Psychology - APPL COGNITIVE PSYCHOL*, 20:353–357, 04 2006. doi: 10.1002/acp.1251.
 - [265] Zoltán Gendler Szabó. Compositionality. *Stanford Encyclopedia of Philosophy*, 2004. Formal definition of compositionality.
 - [266] Zoltán Gendler Szabó. Compositionality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.
 - [267] Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book, 2024.
 - [268] Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *NAACL 2021*, 2021.
 - [269] Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *arXiv preprint arXiv:2402.10176*, 2024.
 - [270] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*,

- abs/2302.13971, 2023. URL <https://arxiv.org/abs/2302.13971>.
- [271] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. <https://ai.meta.com/llama2/>, 2023. Accessed: 2025-02-21.
- [272] Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. Metaphor detection with cross-lingual model transfer. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1024. URL <https://aclanthology.org/P14-1024/>.
- [273] Unbabel. Why translating idioms is hard. <https://resources.unbabel.com/blog/idioms-translation>, 2019. URL <https://resources.unbabel.com/blog/idioms-translation>. Blog post; accessed 2023-05-26.
- [274] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [275] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [276] Eline Visser. Kalamang dictionary. *Dictionaria*, (13):1–2737, 2020. URL <https://dictionaria.clld.org/contributions/kalamang>.
- [277] Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL <https://aclanthology.org/2020.emnlp-main.14>.
- [278] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- [279] Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. Instance weighting for neural machine translation domain adaptation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1155. URL <https://aclanthology.org/D17-1155/>.
- [280] Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. Hypothesis search: Inductive reasoning with language models, 2023.
- [281] Yichong Wang, Yuxuan Feng, et al. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of ACL*, 2020.
- [282] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training

incentivizes reinforcement learning scaling, 2025.

- [283] Taylor W Webb, Keith J Holyoak, and Hongjing Lu. Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. *PNAS Nexus*, 4(5):pgaf135, 05 2025. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgaf135. URL <https://doi.org/10.1093/pnasnexus/pgaf135>.
- [284] Christine M Weber-Fox and Helen J Neville. Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience*, 8(3):231–256, 1996.
- [285] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- [286] Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. Frequency Effects on Syntactic Rule Learning in Transformers. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. URL <https://aclanthology.org/2021.emnlp-main.72.pdf>.
- [287] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.
- [288] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. arXiv preprint arXiv:2201.11903.
- [289] Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. Inverse scaling can become u-shaped, 2023. URL <https://arxiv.org/abs/2211.02011>.
- [290] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [291] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL <https://aclanthology.org/W17-4413/>.
- [292] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL <https://aclanthology.org/2020.emnlp-demos.6>.

- [293] P. Wolff and Dedre Gentner. Evidence for role-neutral initial processing of metaphors. *Journal of experimental psychology. Learning, memory, and cognition*, 26 2: 529–41, 2000.
- [294] Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. Predicting performance for natural language processing tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8625–8646, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.764. URL <https://aclanthology.org/2020.acl-main.764/>.
- [295] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining, 2023. URL <https://arxiv.org/abs/2305.10429>.
- [296] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/xiong20b.html>.
- [297] Frank F. Xu, Uri Alon, Graham Neubig, and Vincent J. Hellendoorn. A systematic evaluation of large language models of code, 2022. URL <https://arxiv.org/abs/2202.13169>.
- [298] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yichang Su, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL <https://arxiv.org/abs/2505.09388>. Introducing the Qwen3 model family—dense and MoE variants with “thinking mode,” expansive context, multilingual capabilities, state-of-the-art performance across coding, reasoning, and agentic tasks; Apache 2.0 open-weight release.
- [299] Chunli Yang. Cultural differences on chinese and english idioms of diet and the translation. *English Language Teaching*, 3, 2010.
- [300] John Yang, Kilian Leret, Carlos E. Jimenez, Alexander Wettig, Kabir Khandpur, Yanzhe Zhang, Binyuan Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang. Swe-smith: Scaling data for software engineering agents. arXiv preprint arXiv:2504.21798, 2025. URL <https://arxiv.org/abs/2504.21798>. Open-sourced

pipeline delivering ~50k synthetic SWE agent training instances.

- [301] Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*, 2024.
- [302] Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. Language models as inductive reasoners, 2024.
- [303] Jiasheng Ye, Peiju Liu, Tianxiang Sun, Yunhua Zhou, Jun Zhan, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance, 2024. URL <https://arxiv.org/abs/2403.16952>.
- [304] Zihuiwen Ye, Pengfei Liu, Jinlan Fu, and Graham Neubig. Towards more fine-grained and reliable NLP performance prediction. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3703–3714, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.324. URL <https://aclanthology.org/2021.eacl-main.324/>.
- [305] Zichun Yu, Fei Peng, Jie Lei, Arnold Overwijk, Wen tau Yih, and Chenyan Xiong. Data-efficient pretraining with group-level data influence modeling, 2025. URL <https://arxiv.org/abs/2502.14709>.
- [306] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.
- [307] Andrea Zaninello and Alexandra Birch. Multiword expression aware neural machine translation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.471/>.
- [308] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of ACL*, 2019.
- [309] Kaiwen Zha, Zhihan Gao, Mengxue Shen, Ze-Wei Hong, Duane S Boning, and Dina Katabi. Rl tango: Reinforcing generator and verifier together for language reasoning. *arXiv preprint arXiv:2505.15034*, 2025.
- [310] Biao Zhang and Rico Sennrich. Root mean square layer normalization, 2019. URL <https://arxiv.org/abs/1910.07467>.
- [311] Junlei Zhang, Zichen Ding, Chang Ma, Zijie Chen, Qiushi Sun, Zhenzhong Lan, and Junxian He. Breaking the data barrier – building gui agents through task

- generalization. arXiv preprint arXiv:2504.10127, 2025. URL <https://arxiv.org/abs/2504.10127>. Demonstrates that reasoning-intensive mid-training tasks (even text-only ones) significantly improve GUI agent generalization.
- [312] Junlei Zhang, Zichen Ding, Chang Ma, Zijie Chen, Qiushi Sun, Zhenzhong Lan, and Junxian He. Breaking the data barrier – building gui agents through task generalization, 2025.
- [313] Xuemiao Zhang, Feiyu Duan, Liangyu Xu, Yongwei Zhou, Sirui Wang, Rongxiang Weng, Jingang Wang, and Xunliang Cai. Frame: Boosting llms with a four-quadrant multi-stage pretraining strategy, 2025. URL <https://arxiv.org/abs/2502.05551>.
- [314] Wenting Zhao, Nan Jiang, Celine Lee, Justin T. Chiu, Claire Cardie, Matthias Gall , and Alexander M. Rush. Commit0: Library generation from scratch. In *International Conference on Learning Representations (Poster)*, 2025. URL <https://openreview.net/forum?id=MMwaQEVsAg>.
- [315] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Aritra Ghosh, Liangchen Xiao, Allyson Ettinger, Sheng Chang, Baolin Peng, Yiren Dai, Harsh Jain, Sebastian Cruz, Ananya Gupta, Haokun Zhang, Soundararajan Srinivasan, Taylor Berg-Kirkpatrick, Eduard Hovy, Christopher D. Manning, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023.
- [316] Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. Large language models can learn rules, 2023.
- [317] Nicolas Zucchet, J rg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations. arXiv preprint arXiv:2503.21676, 2025. URL <https://arxiv.org/abs/2503.21676>. Published March 27, 2025.