Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation

Patrick Fernandes[1,2,3] **Aman Madaan**[1] **Emmy Liu**[1] **António Farinhas**[2,3]
Pedro Henrique Martins[4] **Amanda Bertsch**[1] **José G. C. de Souza**[4] **Shuyan Zhou**[1]
Tongshuang Wu[1] **Graham Neubig**[1,5] **André F. T. Martins**[2,3,4]
[1]Carnegie Mellon University [2]Instituto Superior Técnico (Lisbon ELLIS Unit)
[3]Instituto de Telecomunicações [4]Unbabel [5]Inspired Cognition
pfernand@cs.cmu.edu

Abstract

Natural language generation has witnessed significant advancements due to the training of large language models on vast internet-scale datasets. Despite these advancements, there exists a critical challenge: these models can inadvertently generate content that is toxic, inaccurate, and unhelpful, and existing automatic evaluation metrics often fall short of identifying these shortcomings. As models become more capable, *human feedback* is an invaluable signal for evaluating and improving models. This survey aims to provide an overview of recent research that has leveraged human feedback to improve natural language generation. First, we introduce a taxonomy distilled from existing research to categorize and organize the varied forms of feedback. Next, we discuss how feedback can be described by its format and objective, and cover the two approaches proposed to use feedback (either for training or decoding): directly using feedback or training *feedback models*. We also discuss existing datasets for human-feedback data collection, and concerns surrounding feedback collection. Finally, we provide an overview of the nascent field of *AI feedback*, which uses large language models to make judgments based on a set of principles and minimize the need for human intervention. We also release a website of this survey at feedback-gap-survey.info

1 Introduction

For generation systems to be widely useful, they must generate text that is not only fluent and high-quality, but also well-aligned with human desires and specifications (Vamplew et al., 2018; Hendrycks et al., 2020; Kenton et al., 2021; Turner et al., 2022; Ngo, 2022). Achieving such ambitious goals requires large language models (LLMs) to evolve beyond traditional training methods. Recent improvements in this space have centered on incorporating human feedback (Bai et al., 2022b; Ouyang et al., 2022; OpenAI, 2023a), intended to serve as a guiding force toward the desired outcomes, much like feedback mechanisms in physical machines (Åström and Murray, 2021).

Typically, state-of-the-art language generation systems are obtained by training *probabilistic*, *autoregressive* LLMs on massive amounts of data using *maximum likelihood estimation* (MLE). However, the data used to train these models is generally scraped from the Internet, often containing noise, social biases, and errors (Bolukbasi et al., 2016; Dodge et al., 2021). This combination may result in a *misspecification* of target behavior (Kenton et al., 2021), and may lead to models that generate toxic, inaccurate, and unhelpful content (Sheng et al., 2019; Bender et al., 2021).

The evaluation challenge is compounded as these models are often assessed by automatic metrics that rely on superficial features such as word overlap with reference text. However, these metrics often fail to correlate with *human-perceived* text quality, particularly when models are overly optimized for these metrics (Schluter, 2017; Mathur et al., 2020; Gehrmann et al., 2022; Paulus et al., 2017; Amrhein and Sennrich, 2022)[1]. Considering human-perceived quality can help bridge the *gap* between machine and human generated text and better align the system with desired outcomes (Rosenblueth et al., 1943; Wiener, 1948).

Feedback, as a concept, encompasses a wide range of interpretations (Wiener, 1948); however, some universal characteristics can be identified, such as its format, its intended results, and the ways it is utilized as a part of the model development process. In this survey, we focus on the role of *human feedback* for improving language generation. We start by formalizing *human feedback*,

[1] This is sometimes called *Goodhart's law*: "when a measure becomes a target, it ceases to be a good measure" (Goodhart, 1984)

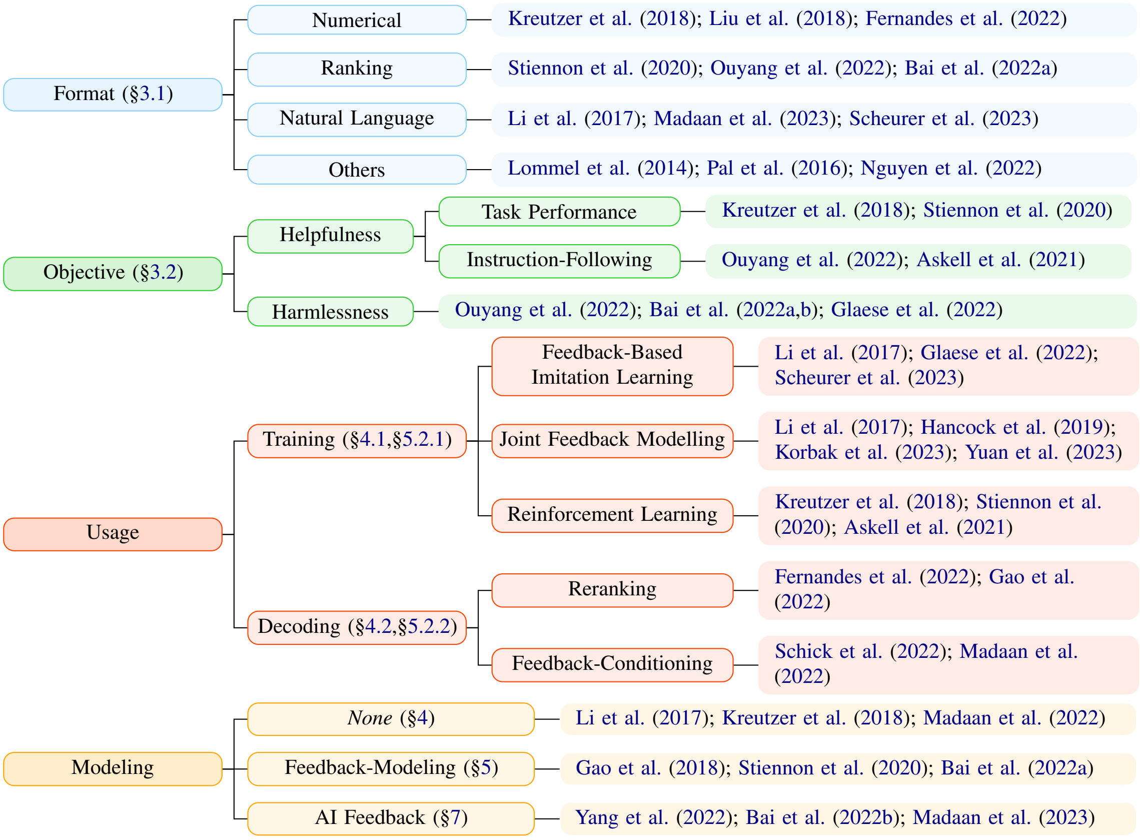
Format (§3.1)
Numerical
Kreutzer et al. (2018); Liu et al. (2018); Fernandes et al. (2022)
Ranking
Stiennon et al. (2020); Ouyang et al. (2022); Bai et al. (2022a)
Natural Language
Li et al. (2017); Madaan et al. (2023); Scheurer et al. (2023)
Others
Lommel et al. (2014); Pal et al. (2016); Nguyen et al. (2022)
Objective (§3.2)
Helpfulness
Task Performance
Kreutzer et al. (2018); Stiennon et al. (2020)
Instruction-Following
Ouyang et al. (2022); Askell et al. (2021)
Harmlessness
Ouyang et al. (2022); Bai et al. (2022a,b); Glaese et al. (2022)
Usage
Training (§4.1,§5.2.1)
Feedback-Based Imitation Learning
Li et al. (2017); Glaese et al. (2022); Scheurer et al. (2023)
Joint Feedback Modelling
Li et al. (2017); Hancock et al. (2019); Korbak et al. (2023); Yuan et al. (2023)
Reinforcement Learning
Kreutzer et al. (2018); Stiennon et al. (2020); Askell et al. (2021)
Decoding (§4.2,§5.2.2)
Reranking
Fernandes et al. (2022); Gao et al. (2022)
Feedback-Conditioning
Schick et al. (2022); Madaan et al. (2022)
Modeling
None (§4)
Li et al. (2017); Kreutzer et al. (2018); Madaan et al. (2022)
Feedback-Modeling (§5)
Gao et al. (2018); Stiennon et al. (2020); Bai et al. (2022a)
AI Feedback (§7)
Yang et al. (2022); Bai et al. (2022b); Madaan et al. (2023)

Figure 1: Taxonomy of methods that leverage human-feedback, with some example representative works.

creating a taxonomy of feedback types and uses (§2). We characterize feedback by its *format* and *objective*, relating to desired model behavior (§3). We explore direct feedback optimization strategies, such as reinforcement learning with human reward functions (§4) and indirect approaches utilizing trained *feedback models* as proxies (§5). We look at human-feedback data datasets and their collection, discussing their influence on models (§6). Lastly, we cover recent work leveraging *AI feedback* from large language models for feedback reduction (§7).

2 A Taxonomy for Leveraging (Human) Feedback for Generation

2.1 Background

Consider a model $M : \mathcal{X} \to \mathcal{Y}$ which, given an input $x \in \mathcal{X}$, outputs *natural language* $\hat{y} \in \mathcal{Y}$. This model encompasses various NLG tasks including **Summarization** ($\mathcal{X}$: documents, $\mathcal{Y}$: summaries), **Machine Translation** ($\mathcal{X}$: source language sentences, $\mathcal{Y}$: target language sentences), **Dialogue Generation** ($\mathcal{X}$: dialogue histories, $\mathcal{Y}$: responses), and **Image Captioning** ($\mathcal{X}$: images, $\mathcal{Y}$: captions).

These models are generally realized as a parameterized, conditional probability distribution $P_\theta(y|x)$, where θ are the model parameters. This distribution is often estimated autoregressively: the probability of a sentence y given input x is decomposed into the product of the probabilities of each token in y, conditioned on the previous tokens. These models are trained by finding $\theta^\star$ that maximizes the likelihood of some training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$. At *inference* time, given input x, an output $\hat{y}$ is decoded from $P_{\theta^\star}$. This decoding can be done, for example, by approximating the most-likely sequence of tokens ($M(x) \approx \arg\max_y P_{\theta^\star}(y|x)$) or by random sampling ($M(x) \sim P_{\theta^\star}(y|x)$).

Evaluating the quality of generated text $\hat{y} \in \mathcal{Y}$ can be challenging due to the complexity and subjectivity of natural language. Although numerous automated metrics have been proposed for different domains, they typically rely on n-gram matching or other rudimentary heuristics. These measures often overlook complex linguistic phenomena, such as paraphrasing or stylistic variations, ultimately failing to align with nuanced human judgments (Sai et al., 2022; Gehrmann et al., 2022). Therefore, for many of these tasks, *human feedback* is considered the gold standard for assessing the quality, and newer *learned* metrics often aim to approximate how humans provide feedback (see §5.1).

Formally, we consider **human feedback** to be a family of functions $\mathcal{H}$ such that $h \in \mathcal{H}$ takes an input[2] $x \in \mathcal{X}$ and one or more outputs $y_1, \cdots, y_n \in \mathcal{Y}$ and returns some *feedback* $f \in \mathcal{F}$:

$$h : \mathcal{X} \times \underbrace{\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n}_{n} \rightarrow \mathcal{F}. \tag{1}$$

A simple example of a feedback function is asking humans to say if, given an input, a particular output is good or bad ($h : \mathcal{X} \times \mathcal{Y} \rightarrow \{0, 1\}$). However, more complex feedback functions, such as rankings or natural language feedback, are also commonly used (see §3.1).

We note that this framing is a *simplification* of the real world: often, different humans might provide different (potentially contradicting) feedback for the same outputs, and a single function may not be able to capture this variability (discussed further in §6). Finally, while our formalization is flexible, it excludes other approaches where models interact with humans to improve learning, such as active learning and other *human-in-the-loop* approaches.

2.2 Taxonomy

Having established a basic mathematical formulation, we now identify four key axes along which we can classify the uses of human feedback: **format**, **objective**, **use** and **modelling**. Figure 1 shows this taxonomy in detail, along with example representative works and how they fit in it. In the next sections we will describe each axis in more detail.

3 Describing Feedback

3.1 Format

An important decision to make when we want to improve language generation systems through human feedback is what *format* to collect this feedback in. This choice has implications on the expressivity of the feedback, the ease of its collection, and how we can use it to improve systems, and the level of *rationality* of said feedback is heavily impacted by this choice (Ghosal et al., 2023). Feedback types are summarized in Table 1 with examples.

Numerical Numerical feedback, which takes an input and output and returns a single score ($\mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{N} \subseteq \mathbb{R}$), is one of the simplest feedback formats to collect and use. Kreutzer et al. (2018) studied using *categorical* feedback, in the form of 5 possible "stars" assigned to a translation, which are averaged to produce a score ($\mathcal{N} = [1, 5]$) to improve the model. Liu et al. (2018) and Shi et al. (2021) used even simpler feedback, by asking humans to choose if a given response is good or not ($\mathcal{N} = \{0, 1\}$). Numerical feedback has also been widely used for evaluation, albeit not with the explicit goal of improving generation. For example, *direct assessments* (Graham et al., 2013) in machine translation ask humans to rate translations on a continuous scale. Some works have attempted to use this data to train feedback models (Sellam et al., 2020; Rei et al., 2020a) and improve generation (Freitag et al., 2022a; Fernandes et al., 2022).

Although easy to leverage, numerical feedback has limitations: reducing feedback to a single score may be a hard and ill-defined task for humans, especially for complex tasks, leading to a costly collection process and problems of *subjectivity* and *variance* (see §6). Furthermore, it may not distinguish well between outputs of similar quality.

Ranking-based An alternative to asking humans to assign a single score to a given input-output pair is asking them to *rank* multiple possible outputs:

$$h : \mathcal{X} \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n \rightarrow S_n$$

where S_n represents the set of all rankings of n elements (optionally allowing ties). This has been used extensively in evaluation (Chaganty et al., 2018). Compared to numerical feedback, this format tends to be easier to collect, and ranking-based feedback is also used to improve model behavior in addition to evaluation (potentially since the former tends to require more feedback data). Ziegler et al. (2019) and Stiennon et al. (2020) asked humans to rank alternative summaries of the system they were trying to improve. Similarly, Ouyang et al. (2022) collected rankings of alternative responses to an *instruction* given to the model. They used these rankings to enhance the model's *instruction-following* capabilities. Subsequent research has also employed ranking-based feedback for the same task (Askell et al., 2021; Bai et al., 2022a,b).

Natural Language Both numerical and ranking-based feedback cannot capture detailed information about problems with the output, which can be crucial for improving generation systems. Natural language feedback typically provides more detailed information, often by suggesting specific shortcomings or revisions for the current output.For example,

[2] Although feedback can be provided independently of the input (for example for *fluency*), we assume some (potentially empty) input for simplicity of notation.

Input	Output(s)	Feedback	Type
A melhor comida do mundo é a portuguesa.	*The worst food in the world are Portuguese.*	0.7	Score
		'worst': `major/accuracy` 'are': `minor/fluency`	MQM
		'worst' → 'best', 'are' → 'is'	Post-Edition
Artificial intelligence has the potential to revolutionize industries (...) but ethical concerns need to be handled.	*AI can change industries.*	`Fluency`: 1 `Relevance`: 0.7	Multi-Aspect
		"Misses the ethical concerns."	Natural Language
Explain the moon landing to a 6 year old	A: *People went to the ...* B: *The moon is a satellite...*	A > B	Ranking

Table 1: Example input and output for three tasks (machine translation, summarization, and instruction following) and possible different (example) feedback that can be given.

Li et al. (2017) asked humans to give natural language feedback to a dialogue question answering model, including positive or negative feedback, but also possibly providing the correct answer to the model or a hint. Tandon et al. (2022) and Madaan et al. (2022) gather natural language feedback on errors in model-generated graphs and the model's interpretation of a given instruction. Scheurer et al. (2022, 2023) improve summarization capabilities of language models by asking humans to provide natural language feedback of the model's summaries. Li et al. (2022) collect natural language feedback (in addition to numerical feedback) for responses from a question answering system.

Others Besides these feedback types, other (potentially domain-specific) types of feedback can be used to improve model behavior. Commonly humans are asked to provide *multi-aspect* feedback ($\mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$ or $\mathcal{F}^d$ more generally), scoring an output or ranking multiple outputs with respect to multiple dimensions (Böhm et al., 2019; Glaese et al., 2022; Madaan et al., 2023; Nguyen et al., 2022). *Post-editions* ask humans to provide corrections to the output in the form of small edits (*e.g.*, *replace X by Y*), and post-edition data has been used to directly improve models (Denkowski et al., 2014) or train *automatic post edition* systems that correct model mistakes (Pal et al., 2016; Mehta and Goldwasser, 2019; Madaan et al., 2021; Talmor et al., 2020; Elgohary et al., 2021). There are other feedback types that haven't been fully leveraged to improve generation: *e.g.*, *Multidimensional Quality Metrics (MQM)* (Lommel et al., 2014), the standard for evaluating translation quality, asks professional translators to identify error *spans* in a translation, alongside severity and type of error.

3.2 Objective

The purpose of collecting feedback is to *align* the model's behavior with some (often ill-defined) *goal behavior*: for example, we might want our summarization model to generate summaries that contain all core information, even if it means they are longer. This **alignment objective** has been studied extensively in the *AI safety and alignment* literature (Bostrom, 2014; Amodei et al., 2016; Bommasani et al., 2021; Kenton et al., 2021), with Leike et al. (2018) proposing the use of feedback models to tackle the difficulty in specifying objectives.

Bai et al. (2022a) explicitly divided the problem of "aligning" a language model into improving its **helpfulness** and increasing its **harmlessness**. Most works implicitly consider either the use of feedback that targets performance factors (such as when targeting overall performance in a task or ability to follow instructions) or harmlessness factors (such as not producing toxic text or providing information that could lead to harm).[3]

Helpfulness Most often, feedback is collected with some *helpfulness* objective in mind: a necessary (but not sufficient) condition for a helpful system is that it performs well, so feedback related to **task performance** generally falls under this umbrella. For example, most works in machine translation leverage feedback related to translation quality (Kreutzer et al., 2018; Fernandes et al., 2022), which is expected to be correlated with its helpfulness in downstream applications. Similarly, in summarization, most works leverage feedback related to aspects such as *relevance*, *consistency* and *accuracy* (Ziegler et al., 2019; Stiennon et al., 2020). One particularly well-studied feedback objective is the ability to **follow instructions** (Ouyang et al., 2022), which encompasses a wide range of tasks.

Harmlessness Another important alignment objective is *harmlessness*: we want our models not to produce certain types of output or violate certain

[3] We mostly ignore the proposed *honesty* aspect, as none of these works tackle this directly.

norms. Feedback collected in Ouyang et al. (2022) considered aspects such as the toxicity of text (besides the overall ability to follow instructions). Bai et al. (2022a) explored the interaction between the helpfulness and harmlessness objectives, showing a trade-off between both. Thoppilan et al. (2022) collected feedback on whether their model violates a set of safety objectives and used it to finetune the model. Glaese et al. (2022) also ask humans to provide feedback on the harmlessness of their system, by defining a set of *rules* and asking humans if the outputs violate these rules. Bai et al. (2022b) showed that feedback produced by LLMs could increase harmlessness without reducing helpfulness.

4 Directly Leveraging Human Feedback

In an ideal scenario, we would directly leverage human feedback to improve generation for both training and decoding.

4.1 Optimizing for Human Feedback

Once human feedback has been collected, one way to use it is by optimizing the model parameters directly. However, this requires the feedback to be "optimizable", *i.e.*, possibly formulated as an optimization problem based on which we can obtain an improved model. For instance, if the feedback is a numerical preference score ($f \in \mathbb{R}$), we can create the following optimization problem:

$$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}}[h(x, M_\theta(x))]. \tag{2}$$

Where $\mathcal{D}$ is the distribution of possible inputs. Various techniques have been suggested to optimize θ using the collected human feedback. These can be divided into three main categories based on the training mechanisms, which we will call **feedback-based imitation learning**, **joint-feedback modeling**, and **reinforcement learning (RL)**.

The **feedback-based imitation learning** approach involves using human feedback to optimize the model by performing supervised learning with a *dataset* composed of positively-labeled generations together with the corresponding inputs, $\mathcal{D}^+$. This can be achieved by minimizing the loss:

$$\theta^\star = \arg\min_\theta \sum_{i=1}^{|\mathcal{D}^+|} \mathcal{L}^{(i)}(\theta) \tag{3}$$

$$\mathcal{L}^{(i)}(\theta) = -\log p_\theta\left(y^{(i)} \mid x^{(i)}\right) \tag{4}$$

An instance of this approach can be found in Li et al. (2017), in which the authors train a dialogue model by maximizing the likelihood of the model's answers labeled correct by humans. Similarly, Kreutzer et al. (2018) trained a machine translation model on a set of positively-labeled translations, and Glaese et al. (2022) performed supervised learning on the dialogues which complied with their pre-defined rules (concerning correctness, harmfulness, and helpfulness), according to humans. A slightly different approach was proposed by Hancock et al. (2019): deploying a chit-chat dialogue model and using the human utterances as targets to fine-tune the model. Scheurer et al. (2022, 2023) leverage the fact that LLMs can follow instructions and start by collecting natural language human feedback about the model generations, which often describes what an improved text would look like. Then, they ask the LM to generate multiple refinements based on the input, previous model generation, and the corresponding feedback. The highest similarity refinements for each generation are then used to fine-tune the LLM. OpenAI's `text-davinci-002` was trained with both human demonstrations and model outputs with the highest possible rating, an approach deemed *FeedME* (OpenAI, 2023b). A downside of these approaches is that they disregard generations which do not receive positive feedback, which may also contain useful information.

On the other hand, **joint-feedback modeling** leverages all the information collected by directly using human feedback to optimize the model. Also, as the feedback is modeled directly by the model, this allows feedback in formats other than numerical or ranking-based (*e.g.*, natural language). Having $\mathcal{D}$ as the *dataset* of inputs x, generations y, and human feedback f collected, this can be achieved by minimizing a loss of the form

$$\mathcal{L}^{(i)}(\theta) = -\log p_\theta\left(y^{(i)}, f^{(i)} \mid x^{(i)}\right) \tag{5}$$

Over all examples in $\mathcal{D}$. This equation can be factorized as $\mathcal{L}^{(i)}(\theta) = -\log p_\theta\left(f^{(i)} \mid y^{(i)}, x^{(i)}\right) + \log p_\theta\left(y^{(i)} \mid x^{(i)}\right)$. Some works simply train the model to predict the feedback given to each generation (Weston, 2016, forward prediction), disregarding the second term of the factorization. One example is the work of Li et al. (2017), in which the authors asked humans to give natural language feedback (*e.g.*, positive/negative feedback, providing the correct answer, or giving a hint about the correct answer) to a dialogue question answering model. Then, the model itself is trained to predict this feedback. Hancock et al. (2019) proposed having an auxiliary model predicting the satisfaction of the human speaking with the model. If the satisfaction score is lower than a pre-defined threshold, the model will ask the human for feedback. The model then leverages the natural language feedback humans give by learning to predict it. Yuan et al. (2023); Rafailov et al. (2023) showed that having summarization models predict rankings of different summaries helps the model generate better summaries, and may even outperform more complicated approaches with feedback models (§5).

Other works train the model to predict the generations and the corresponding human feedback. Xu et al. (2022) proposed using the DIRECTOR model introduced by Arora et al. (2022) to leverage human feedback. As this model has a unified decoder-classifier architecture, Xu et al. (2022) proposed using positively-labeled examples to train its language modeling head (similarly to feedback-based imitation learning) and using both positive and negatively-labeled examples to train a classifier head that directs the model away from generating undesirable sequences. Thoppilan et al. (2022) follow this approach to enforce the model's quality and safety: using collected dialogues between crowd-workers and the model LaMDA, annotated with the crowd-workers' feedback on each response's quality, LaMDA is fine-tuned to predict the high-quality responses alongside each response's quality attributes and safety.

Finally, this can also be achieved by training the model to predict generation and conditioning on the feedback. This corresponds to minimizing:

$$\mathcal{L}^{(i)}(\theta) = -\log p_\theta \left(y^i \mid f^i, x^i\right) \tag{6}$$

Liu et al. (2023) proposed prompt-based fine-tuning, where they create prompts containing previous generations rated by humans, in the order of preference and insert language-based feedback (*e.g.*, "... is a worse than ...") to the prompt, between the generations. Then, the model is fine-tuned on the preferred answers. In Section 5.2.1, we discuss scnearios where the feedback f is sourced from a feedback model instead of humans.

Finally, **reinforcement learning (RL)** offers a more versatile approach, allowing for direct optimization of a model's parameters based on human feedback, regardless of the feedback's differentiability. A common RL algorithm used in this context is REINFORCE (Williams, 1992), which updates the policy parameters using the following gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_{x\sim\mathcal{D},y\sim p_\theta}[h(x,y)\nabla_\theta \log p_\theta(y \mid x)] \tag{7}$$

Here, $\mathcal{D}$ represents the set of inputs x, and p_θ is the policy. This flexibility enables RL to handle various types of feedback h and better align the generated output with human preferences. For instance, Kreutzer et al. (2018) proposed using task-based implicit feedback from user queries as a reward signal to train a machine translation model using a word-level variant of minimum risk training (Shen et al., 2016), while Jaques et al. (2019) used implicit human reactions in chat to improve open-domain dialog systems through off-policy Q-learning (Watkins and Dayan, 1992). Given that collecting human feedback can be expensive and time-consuming, learning is done offline from logged data, which is typically more favorable than on-policy settings that need feedback on the fly. Later in §5.2.1, we discuss several works that attempt to optimize feedback models using RL instead of directly optimizing human feedback. In conjuction, these aproaches are commonly known as *Reinforcement Learning from Human Feedback* (**RLHF**).

4.2 Decoding with Human Feedback

Directly adjusting model parameters might not always be feasible, especially for LLMs. Moreover, during the training phase, consistent and meaningful feedback may not always be readily available. In such settings, leveraging human feedback during decoding becomes crucial. There are two primary approaches in this realm: 1. *Feedback Memory:* This involves maintaining past feedback and incorporating relevant aspects when processing new inputs, guiding the model toward preferential outputs.

To illustrate, imagine a scenario where the model produces an output that is either biased or factually incorrect. Upon receiving feedback highlighting this flaw, a model without feedback memory capabilities would still be prone to making the same error on similar inputs. In contrast, a model equipped with a robust feedback memory mechanism can actively reference this feedback. When faced with a comparable input or context in the future, it can thus reduce the likelihood of reproducing the same error. This feedback memory can be conceptualized as a repository or "bank" where past feedback is stored. Depending on the implementation, this could be in the form of plain text entries (Madaan et al., 2022) or dense vector representations (Tandon et al., 2022). When processing new inputs, the model first probes this memory bank to identify if a similar input or context exists. If a match or a close approximation is found, the model retrieves the corresponding feedback. This feedback can then be factored (e.g., by concatenating the feedback to the prompt) in to produce a refined output.

While the notion of learning from past experiences or feedback traces its roots to early cognitive theories and computational models (Riesbeck, 1981; Schank, 1983), its effectiveness in finetuning language models and few-shot learning settings has been shown in recent work (Weston et al., 2014; Wu et al., 2018; Tandon et al., 2022; Madaan et al., 2022).

2. *Iterative Output Refinement:* This method employs human feedback to refine the model's output iteratively. Users can provide feedback on intermediate responses, enabling the model to adjust its output until it meets the user's satisfaction. This process allows the model to better understand user preferences and produce more suitable outcomes (Reid and Neubig, 2022; Saunders et al., 2022; Schick et al., 2022; Nijkamp et al., 2022). Feedback can also be provided on model attributes such as the decoding strategy (Passali et al., 2021), rather than directly on its outputs.

5 Improving Generation using Human Feedback Models

Directly using human feedback to improve model behavior is not feasible in the general case: asking humans to provide feedback for *every* model output is both expensive and time-consuming.

5.1 Learning Models of Human Feedback

An alternate approach to obtaining human feedback is to develop models that can predict or approximate it. Although they may not be perfect, they can provide feedback at a low cost after training, enabling feedback-dependent techniques at scale.

More formally, given a feedback function $h : \mathcal{X} \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n \to \mathcal{F}$, we want to learn a *parametric* (numerical) feedback model $\hat{h}_\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ (with parameters ϕ) that "agrees" with human feedback. This agreement is expressed through a loss which the model is trained to minimize.

$$\phi_\star = \arg\min_\phi \mathbb{E}_{x,y_1,\cdots,y_n \sim \mathcal{D}_f} \left[\mathcal{L}(\phi)\right] \quad (8)$$

$$\mathcal{L}(\phi) = \text{loss}\left(\hat{h}_\phi(x, y_1), \cdots, h(x, y_{1:n})\right) \quad (9)$$

For example, if the feedback function we are trying to model is also numerical ($h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$), then this loss can just be any standard regression loss, such as the squared difference between the human feedback and model feedback $\mathcal{L}(\phi) = \left(\hat{h}_\phi(x, y) - h(x, y)\right)^2$. Importantly, while the feedback model is (generally) numerical, the human feedback can be in any other format, as long as a suitable loss function can be specified. Stiennon et al. (2020) train *preference* models [4] $\hat{h}_\phi(x, y_n)$ on ranking-based feedback, using a loss of the form

$$\mathcal{L}(\phi) = \log\left(\sigma\left(\hat{h}_\phi(x, y_{+1}) - \hat{h}_\phi(x, y_{-1})\right)\right) \quad (10)$$

such that sample y_{+1} was preferred to y_{-1} for the same input x: $h(x, y_{-1}, y_{+1}) = (y_{-1} < y_{+1})$. Variants of this loss have been used in subsequent works (Ouyang et al., 2022; Askell et al., 2021; Liu et al., 2022; Qin et al., 2022; Yuan et al., 2023).

Feedback modeling has been studied extensively in the context of *metric learning* for NLP. In MT, Sellam et al. (2020) and Rei et al. (2020a) trained BLEURT and COMET, respectively, to regress on human translation quality assessments. For summarization, Zopf (2018) leveraged annotated pairwise preferences to train a preference model and Peyrard et al. (2017) learned a summary-level metric from a set of human judgements from older summarization datasets (*e.g.,* TAC-2008). These metrics have been shown to correlate much better with human judgments than widely used lexical-metrics such as BLEU and ROUGE (Freitag et al., 2022b). Notably, these reward models were not trained with the intent of improving generation directly, though somewere used for that purpose later (§5.2).

Recently, there has been interest in developing feedback models directly for improving generation (Böhm et al., 2019; Ziegler et al., 2019). Initialized from either the target LM to improve or a smaller one from the same family, the feedback model finetuned on (collected) human feedback. This data is typically collected by asking annotators to provide

[4] We specify the feedback model with respect to the human feedback format, *i.e.*, *reward* and *preference* model for numerical and ranking-based human feedback, respectively.

feedback on outputs from an earlier version of the model being improved. It is also possible to first finetune the feedback model on naturally occurring implicit feedback, such as user interactions on sites (e.g., Reddit, StackOverflow), which greatly increases data size at the cost of noisier data.

Nguyen et al. (2022) train a preference model based on rankings on three human-designed objectives: whether the summary has an appropriate topic, length, and quality, combining these three into a single objective using a distance-based ranking loss. Interestingly, automatic post-editing (APE) systems in MT (*e.g.*, Simard et al. (2007); Correia and Martins (2019)) can also be seen as feedback models (albeit non-numerical). Their aim is to automatically correct the output of an MT system and they are trained on human post-editions.

5.2 Leveraging Feedback Models to Improve Generation

After training a feedback model, we can use it almost exactly as we would use human feedback: either by leveraging this feedback model during training of the generation model, or by incorporating the feedback model during decoding.

5.2.1 Optimizing for Feedback Models

Similar to optimizing for human feedback, one way to use the feedback model is to optimize the model with respect to the feedback it gives. If the feedback model outputs numerical feedback ($\hat{h}_\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$) we can define an optimization problem similar to Equation 2. However, due feedback models being imperfect proxies, typically a *regularization* term R is introduced to avoid *"overfitting"* to the feedback model (Ziegler et al., 2019):

$$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}} \left[\hat{h}_\phi(x, M_\theta(x)) - \beta R(\theta)\right] \tag{11}$$

Due to the similarities between both optimization problems, approaches to tackle Equation 11 can be divided into two of the three categories in §4.2: **joint-feedback modeling** and **reinforcement learning**. Recall that while in §4.2 we discuss approaches for directly optimizing for human feedback (h), while this section is focused on cases where a model of human feedback ($\hat{h}$) is used instead.

Unlike when using human feedback directly, most works attempt to optimize feedback models using **reinforcement learning**. Gao et al. (2018); Böhm et al. (2019) use the feedback collected in other works to train reward and preference models, and use reinforcement learning to optimize against these models. They show that humans preferred their summarization model to other supervised and RL-trained baselines. Ziegler et al. (2019) proposed a similar approach, but trained preference models with feedback collected on the model being improved, and introduced a *KL* regularization term

$$R(\theta) = \log\left[P_\theta(y|x)/P_{\theta_{\text{SL}}}(y|x)\right] \tag{12}$$

to avoid the optimized model deviating too much from the original (supervised) model with parameters θ_{SL}[5]. Stiennon et al. (2020) extended this work, by *scaling* both the summarization and preference models, showing that their model was highly preferred by humans, and generalized better than supervised baselines. Ouyang et al. (2022) also used reinforcement learning with preference models to LLMs' ability to follow instructions, but combined the RL objective with the original pretraining objective on public NLP benchmarks. Other works have also used reinforcement learning with preference models similarly (Askell et al., 2021; Bai et al., 2022a; Wu et al., 2021; Nguyen et al., 2022). In these methods, generally the model is first trained with imitation-learning on human demonstrations, improving performance compared to using RL directly on the pretrained policy. Glaese et al. (2022) compared doing feedback-based imitation learning with human feedback (§4.1) with doing reinforcement learning with a feedback model, finding that the latter led to a better preference rate and lower rule violation rate.

In terms of the impact of the underlying RL algorithm, PPO (Schulman et al., 2017) is by far the most used algorithm, and the one where tricks for its success are more widely known (see Zheng et al. (2023)). However, it is unclear how important a role it plays in the success of RLHF, and some works have proposed alternative RL algorithms that claim better performance (Donato et al., 2022).

Casper et al. (2023) identify several intrinsic limitations of RLHF, including human evaluation inconsistencies, the potential for feedback manipulation, trade-offs between feedback depth and efficiency, difficulties in capturing diverse human values in reward functions, and risks of reward hacking and policy deployment shortcomings.

[5] Note that this KL term is different from other algorithm-specific regularization terms, such as the KL terms in PPO (Schulman et al., 2017).

The **joint-feedback modeling** with feedback *models* was explored by Korbak et al. (2023), who study pre-training an LLMs with a loss similar to Equation 6, based on feedback from a preference model trained on ranking-based feedback for toxicity. They showed that this leads to models producing less toxic generations, when compared to pretraining a model with vanilla MLE. Note that this is different from techniques discussed in 5.1, as the focus there was to train models with real human feedback, not their model.

5.2.2 Decoding with Feedback Models

As mentioned, feedback models can be queried cheaply for feedback once trained. Perhaps for this reason, most approaches use feedback models by sampling a large number of candidate generations, and rerank them with the feedback model:

$$\mathcal{C} = \{\bar{y}_1, \cdots, \bar{y}_S\} \text{ where } \bar{y}_i \sim P_\theta(y|x)$$

$$\hat{y} = \arg\max_{\bar{y} \in \mathcal{C}} \hat{h}_\phi(x, \bar{y})$$

where $\hat{h}_\phi$ is a trained (numerical) feedback model and $\mathcal{C}$ is a set of candidate generations given by the model (for example, by sampling multiple times).

In machine translation, Fernandes et al. (2022) and Freitag et al. (2022a) use feedback model training, involving a two-stage process of candidate generation and scoring with quality metrics learned from human judgments (Rei et al., 2020a,b). Top-rated candidates are selected using reranking or MBR decoding (Kumar and Byrne, 2002). Similarly, Li et al. (2022) improves a QA system by gathering numerical and natural language feedback, then refining a pretrained model on this feedback to rerank predictions. Works like Bhattacharyya et al. (2022) also demonstrate efficiency in enhancing machine translation outputs via automatic post-editing systems.

Feedback Model Overoptimization One problem that arises when optimizing a system with a feedback model is that the model is an imperfect proxy for human feedback. Therefore, systems may overoptimize for good model scores, but not humans. This is known as the *overoptimization* problem, and is the main reason for the regularization term in Equation 11. Gao et al. (2022) studies the overoptimization problem in preference models, by both optimizing against it with RL (training) and reranking outputs with it (decoding), finding that both lead to similar levels of overoptimization.

5.3 Comparing the Effectiveness of Approaches

While numerous approaches have been proposed for incorporating human feedback, it is difficult to directly compare their relative effectiveness, as dealing with feedback introduces many additional experimental variables (the quality of the original generation model/policy, the feedback collection process, etc). Nevertheless, a few studies compare a subset of these approaches within a consistent experimental setup, allowing for some high-level conclusions to be drawn:

- Comparing approaches that leverage feedback to optimize the model, while RLHF seems to be the predominant technique for current SotA LLMs (Glaese et al., 2022; OpenAI, 2023a; Touvron et al., 2023), some works claim that simpler approaches (such as joint-feedback modelling *directly with human preferences*) can lead to better or comparable performance (Yuan et al., 2023; Rafailov et al., 2023).
- It is also not clear if optimizing the model's with the feedback (model) is necessary to obtain the best performance, and instead using the feedback model to rerank the outputs of the model leads to comparable results (Gao et al., 2022).

6 Collecting and Using Human Feedback

Collecting human feedback can be expensive and may present issues for the inexperienced. We highlight existing datasets, their collection methods, and considerations for those creating their own preference datasets. Annotator variability remains largely unexplored (Plank, 2022; Gehrmann et al., 2023), though evidence suggests well-constructed annotation guidelines are necessary (Ziegler et al., 2019; Parmar et al., 2023) to avoid systemic bias away from the intended task.

In general, there is not much discussion in the literature as to how choices made in the feedback collection process impact the final generalization ability of the model, as this has not been studied in a controlled setting. However, there is an appreciation for the importance of data quality in feedback collection, as researchers make efforts to filter out annotators based on their agreement with gold labels, as well as based on inter-annotator agreement (Stiennon et al., 2020; Bai et al., 2022a). However, Bai et al. (2022a) note that judging data quality is difficult for more open-ended forms of feedback such as dialogues. Despite this, they were able to

Task	Dataset & their descriptions	Collection method	Platform	Feedback Type
Language assistant	HH-RLHF (Bai et al., 2022a; Perez et al., 2022a)	Explicit	Upwork, MTurk	Ranking
Language assistant	SHP (Ethayarajh et al., 2023)	Implicit	Scraped from Reddit	Ranking/Score
Summarization	summarize-from-feedback (Stiennon et al., 2020)	Explicit	Upwork	Ranking
Question Answering	FeedbackQA (Li et al., 2022)	Explicit	MTurk	Score, NL
Question Answering	StackExchange (Lambert et al., 2023)	Implicit	StackOverflow	Ranking
Translation	WMT Metrics Shared Task (Freitag et al., 2022b)	Explicit	Pro translation workflow	MQM, DA

Table 2: Summary of existing human feedback datasets and their collection methods, which vary along several dimensions. Refer to Table 1 for definitions of feedback types. A separation is drawn between datasets explicly designed to capture human preferences for model improvement, and datasets designed for evaluation, such as MQM/DA datasets in MT. **N/A** means we could not find information.

achieve good results without detailed data quality controls. The impact of collection methods on final results may be a direction for future research, but for now, we present considerations for data collection along different axes below that experimenters should keep in mind, based on previous studies:

1. **Annotator expertise**: Depending on task and training (Snow et al., 2008; Sheng et al., 2008; Clark et al., 2021; Gillick and Liu, 2010; Freitag et al., 2021), annotators can be domain experts to crowdworkers or even models. Expert feedback is usually more reliable but considerably more expensive (due to recruitment difficulty) (Kulkarni et al., 2014). In many tasks, like translation or summarization, using crowdworkers and models can be sufficient and, if given the correct instruction, they can even help mimic expert opinions (Moore et al., 2020). [6]
2. **Length of engagement**: Involves one-time or long-term collaborations with annotators, with preference datasets often involving extended partnerships (Stiennon et al., 2020; Bai et al., 2022a; Freitag et al., 2021).
3. **Type of feedback:** Existing datasets generally use rankings or scores, but future work may investigate the value of more detailed feedback, which humans usually prefer to provide (Stumpf et al., 2007; Amershi et al., 2014; Ghai et al., 2021).
4. **Collection method**: Data can be gathered explicitly through experiments or implicitly from online sources/user interactions, with varying noise (Kreutzer et al., 2018; Freitag et al., 2021). When explicitly annotating, the choice of feedback type generally influences the type of collection: surveys generally are oriented toward more numerical or ranking-based feedback, while user studies or interviews are used to collect more detailed, open-ended feedback.
5. **Collection platform**: Platforms include Amazon Mechanical Turk, Upwork, Scale AI, and — when the collection is implicit — some discussion platforms where human interactions and preferences emerge organically. Alternately, researchers may collect their own feedback through online forms or interfaces, and recruit participants from their own institution.
6. **Annotator demographics**: Annotator identities can influence labeling; in some cases, demographics are collected during data collection. (Sap et al., 2022; Ding et al., 2022).

Note that some of these dimensions are shared more broadly across various tasks that involve humans in the loop, including human evaluation (Gehrmann et al., 2023; Liao and Varshney, 2021), interactive model debugging (Lertvittayakumjorn and Toni, 2021), data collection (Suhr et al., 2021), etc. For example the evaluation on text generation can sometimes be viewed similar to preference collection: hosted on crowdsourcing platforms, acquired from non-experts, collected in the form of ranking feedback (e.g., reads better better than the text from a baseline generator). In our enumeration above, we mostly focused on how these dimensions are implemented specifically in the context of feedback collection, and leave cross-comparison with other human-in-the-loop approaches to the reader (Wang et al., 2021). Table 2 shows some existing human feedback datasets.

Variance in judgement Considering K annotators with feedback functions $h_{i}{}_{i=1}^{K}$, judgments are given on data $\mathcal{D} = d_1, ..., d_N$. Inter-rater reliability metrics, such as Cohen's Kappa, Fleiss' Kappa, or Krippendorff's alpha, can assess annotator agreement (Hayes and Krippendorff, 2007; Fleiss, 1971;

[6] In some cases, when data is collected from user interaction or mined from existing data sources, it may not be possible to control for expertise of annotators.

Cohen, 1960). Low reliability may result from unclear tasks or evaluation criteria (Gehrmann et al., 2023; Thomson and Reiter, 2021), underqualified annotators, inherent subjectivity, or multiple plausible interpretations (Plank, 2022; Nie et al., 2020; Gordon et al., 2022). Mitigation strategies include viewing humans as making noisily-rational choices (Ghosal et al., 2023), learning the reliability level of feedback from multiple humans (Yamagata et al., 2021), augmenting evaluation metrics like COMET with confidence intervals (Glushkova et al., 2021; Zerva et al., 2022), and requiring annotators to justify their rankings (Ziegler et al., 2019).

(In)experienced annotators There is generally a trade-off between the effort needed to create the datasets and the reliability of judgments collected. While some have claimed that a small number of crowdworkers can replace a domain expert in certain tasks such as affect recognition, recognizing textual entailment, or word-sense disambiguation (Snow et al., 2008; Sheng et al., 2008), this heavily depends on the task. Untrained crowdworkers may rely on more surface heuristics to evaluate text (Clark et al., 2021; Gillick and Liu, 2010), and one comparison of MT model evaluations performed by expert translators and crowdworkers found low agreement between the groups led to a completely different ranking of the models, with crowdworker evaluation being less reliable than automatic embedding-based metrics (Freitag et al., 2021). Generally, high-stakes applications or applications dependent on specific linguistic or specialized domain knowledge may need to rely on feedback from human experts, and extended partnerships with annotators can provide consistency of annotations. Crowdworkers or AI feedback may be acceptable substitutes in other situations; for general alignment with human preferences, it may instead be prudent to recruit a large and diverse group of annotators to avoid overfitting to the preferences of specific annotators or demographics. As the difficulty of tasks increases, it may become more difficult for non-experts to provide feedback, and evaluation of difficult tasks such as writing secure code may require designing feedback methods that incorporate human-AI teams, or rigorous criteria for evaluating feedback (Saunders et al., 2022; Perry et al., 2022; Bowman et al., 2022).

Subjectivity in judgment Some subjectivity in annotator judgment can arise from differences in cultural or social groups (Santurkar et al., 2023). Several works observe that tuning with human feedback increases the model's alignment with US liberal views on controversial topics (Perez et al. (2022b), Hartmann et al. (2023)). Annotators with different backgrounds may disagree on what qualifies as toxic content (Sap et al. (2022), Ding et al. (2022)). This is pronounced when annotators are asked to make ethical judgments (Jiang et al. (2022), Talat et al. (2022)). Some work has critiqued the idea of a unified human preference (Prabhakaran et al., 2021; Casper et al., 2023), suggesting that some variance in judgment is both expected and potentially useful signal.

Biases in judgement Presenting annotators with isolated text can lead to oversight of superior alternatives, mistakenly marking the text as high-quality (Bansal et al., 2021). When generating text, *anchoring bias* can influence writing style (Jakesch et al., 2023; Lehmann et al., 2022) and the given suggestions or corrections. Fundamentally, there may be a difference between what is correct and what humans want to hear, which may lead models to imitate persuasive behaviour, which may influence humans to rate an output more highly if it "feels familiar" (Hasher et al., 1977; Griffin et al., 2023). Mitigation strategies entail ranking diverse outputs and defining explicit evaluation criteria.

When giving feedback in traditional RL environments, users tend to give much more positive feedback than negative (a *positivity bias*), which may lead the agent to avoid the goal they are actually trying to reach (Amershi et al., 2014; Knox and Stone, 2013; Thomaz and Breazeal, 2008).

Ethical considerations Prolonged content moderation tasks can be harmful (Steiger et al., 2021). Tasks involving toxicity classification and generation from open-ended inputs may particularly affect annotators (Shmueli et al., 2021). Media attention has focused on fair pay for annotators, with one *TIME* article[7] describing annotators paid $2 USD/hr or less to provide annotations of toxic content for RLHF datasets. Research on crowdsourcing (Shmueli et al. (2021); Rothschild et al. (2022); Soratana et al. (2022); Toxtli et al. (2021); Hornuf and Vrankar (2022)) cautions that inadequate pay, especially in lower-resourced regions, is a form of worker exploitation.

[7] https://time.com/6247678/openai-chatgpt-kenya-workers/

7 AI Feedback

Feedback models have been crucial in advancing generation techniques by effectively leveraging feedback. However, they are heavily reliant on human input: for example, Gao et al. (2022) found that across various preference model sizes, utilizing fewer than 1,000 comparisons resulted in only chance improvements. Moreover, employing static feedback can make consistency challenging, causing changes in the model's output distribution. AI-generated feedback, an emerging research area, focuses on harnessing the LLM's own abilities to enhance the model without human intervention. Two primary approaches have emerged in this domain:

Self AI Feedback The first approach involves using the same model to provide feedback and improve its output. In this scenario, the model engages in a continuous self-improvement process, learning from its evaluations and refining its capabilities accordingly. Examples of this approach include prompting models to generate harmful responses and revising them for harmlessness (Bai et al., 2022b), or employing rule-based reward models for RLHF fine-tuning (OpenAI, 2023a). Techniques such as iterative output revision through few-shot prompting (Peng et al., 2023; Shinn et al., 2023; Chen et al., 2023; Paul et al., 2023; Madaan et al., 2023; Yang et al., 2022) have been explored using LLMs like GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023a). Notably, these techniques demonstrate potential when applied to LLMs trained to adhere to human instructions and align outputs with human preferences. This suggests that incorporating human feedback during training equips AI models to comprehend task requirements better, align outputs with directives, and function as dependable feedback mechanisms, thereby minimizing human intervention. The capacity to offer valuable AI feedback may depend on the model being trained with human feedback.

External AI Feedback: This approach utilizes a separate feedback model to critique the task model's outputs (Yasunaga and Liang, 2020; Madaan et al., 2021; Welleck et al., 2022; Bai et al., 2022b; Akyürek et al., 2023). A key advantage is that the feedback model need not be large or general-purpose, making smaller feedback models an appealing option when abundant feedback is available.

8 Conclusion

Recent developments in large language models have emphasised the need for human feedback to ensure models have desirable behaviour and generate helpful and harmless text. We provide an overview of a recent line of research on leveraging (human) feedback to improve natural language generation. Despite the relative infancy of this field, several important observations emerge when considering existing works:

1. Current models often underutilize more expressive forms of feedback like natural language, favouring ranking-based or numerical feedback.
2. A trade-off exists between effort spent creating datasets and the reliability of judgments. Enlisting expert and diverse annotators can be beneficial for high-stakes applications.
3. The value of leveraging feedback lies primarily in the feedback itself rather than the specific method. While Reinforcement Learning from Human Feedback (RLHF) has been popular, other methods report notable improvements and might be simpler to apply (Gao et al., 2022; Rafailov et al., 2023; Zhou et al., 2023; Zhao et al., 2023). However, large-scale comparative analysis remains necessary.

Acknowledgments

This work was supported by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the projects DECOLLAGE (ERC-2022-CoG 101088763); Project Center for Responsible AI, grant 62, under PRR funded by NextGenerationEU (2022-C05i0102-02); and by a grant from the Singapore Defense Science and Technology Agency.

References

Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. 2023. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. Ai Magazine, 35(4):105–120.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. CoRR, abs/1606.06565.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1125–1141, Online only. Association for Computational Linguistics.

Kushal Arora, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. 2022. Director: Generator-Classifiers For Supervised Language Modeling. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861.

Karl Johan Åström and Richard M Murray. 2021. Feedback systems: an introduction for scientists and engineers. Princeton university press.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–16.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2022. Findings of the WMT 2022 shared task on automatic post-editing. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 109–117, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In 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 3110–3120, Hong

Kong, China. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. ArXiv, abs/2108.07258.

Nick Bostrom. 2014. Superintelligence: Paths, Dangers, Strategies, 1st edition. Oxford University Press, Inc., USA.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. 2022. Measuring progress on scalable oversight for large language models.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. arXiv preprint arXiv:2304.05128.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A.

Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7282–7296, Online. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20:37–46.

Gonçalo M. Correia and André F. T. Martins. 2019. A simple and effective approach to automatic post-editing with transfer learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3050–3056, Florence, Italy. Association for Computational Linguistics.

Michael Denkowski, Chris Dyer, and Alon Lavie. 2014. Learning from post-editing: Online model adaptation for statistical machine translation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 395–404, Gothenburg, Sweden. Association for Computational Linguistics.

Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of annotator demographics on sentiment dataset labeling. Proc. ACM Hum.-Comput. Interact., 6(CSCW2).

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Domenic Donato, Lei Yu, Wang Ling, and Chris Dyer. 2022. Mad for robust reinforcement learning in machine translation.

Ahmed Elgohary, Christopher Meek, Matthew Richardson, Adam Fourney, Gonzalo Ramos, and Ahmed Hassan Awadallah. 2021. Nl-edit: Correcting semantic parse errors through natural language interaction. arXiv preprint arXiv:2103.14540.

Kawin Ethayarajh, Heidi Zhang, Yizhong Wang, and Dan Jurafsky. 2023. Stanford human preferences dataset.

Patrick Fernandes, António Farinhas, Ricardo Rei, José De Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

J.L Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76:378–382.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics, 9:1460–1474.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. Transactions of the Association for Computational Linguistics, 10:811–825.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization.

Yang Gao, Christian M. Meyer, and Iryna Gurevych. 2018. APRIL: Interactively learning to summarise by combining active preference learning and reinforcement learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4120–4130, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina McMillan-Major, Anna Shvets, Ashish Upadhyay, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir R. Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh D. Dhole, Khyathi Raghavi Chandu, Leonardo F. R. Ribeiro, Lewis Tunstall, Li Zhang, Mahima Pushkarna, Mathias Creutz, Michael White, Mihir Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qinqin Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja vStajner, Sébastien Montella, Shailza, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin P. Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Yi Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. Gemv2: Multilingual nlg benchmarking in a single line of code. In Conference on Empirical Methods in Natural Language Processing.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. Journal of Artificial Intelligence Research, 77:103–166.

Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW3):1–28.

Gaurav R. Ghosal, Matthew Zurek, Daniel S. Brown, and Anca D. Dragan. 2023. The effect of modeling human rationality level on learning rewards from multiple feedback types.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 148–151, Los Angeles. Association for Computational Linguistics.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. arXiv preprint arXiv:2209.14375.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics.

C. A. E. Goodhart. 1984. Problems of Monetary Management: The UK Experience. Macmillan Education UK, London.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In CHI Conference on Human Factors in Computing Systems. ACM.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Lewis D Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly T Mai, Maria Vau, Matthew Caldwell, and Augustine Marvor-Parker. 2023. Susceptibility to influence of large language models.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019.

Learning from Dialogue after Deployment: Feed Yourself, Chatbot! In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3667–3684.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation.

Lynn Hasher, David Goldstein, and Thomas Toppino. 1977. Frequency and the conference of referential validity. Journal of Verbal Learning and Verbal Behavior, 16(1):107–112.

A.F Hayes and K Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1:77–89.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. arXiv preprint arXiv:2008.02275.

Lars Hornuf and Daniel Vrankar. 2022. Hourly wages in crowdworking: A meta-analysis. Business & Information Systems Engineering, 64(5):553–573.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. ArXiv, abs/2302.00560.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Àgata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind W. Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. CoRR, abs/1907.00456.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2022. Can machines learn morality? the delphi experiment.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. CoRR, abs/2103.14659.

W. Bradley Knox and Peter Stone. 2013. Learning non-myopically from human-generated reward. In Proceedings of the 2013 International Conference on Intelligent User Interfaces, IUI '13, page 191–202, New York, NY, USA. Association for Computing Machinery.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback? In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), pages 92–105, New Orleans - Louisiana. Association for Computational Linguistics.

Anand Kulkarni, Prayag Narula, David Rolnitzky, and Nathan Kontny. 2014. Wish: Amplifying creative ability with expert crowds. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 2, pages 112–120.

Shankar Kumar and William Byrne. 2002. Minimum bayes-risk word alignments of bilingual texts. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, page 140–147, USA. Association for Computational Linguistics.

Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. 2023. Huggingface h4 stack exchange preference dataset.

Florian Lehmann, Niklas Markert, Hai Dang, and Daniel Buschek. 2022. Suggestion lists vs. continuous generation: Interaction design for writing with generative models on mobile devices affect text length, wording and perceived authorship. Proceedings of Mensch und Computer 2022.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. ArXiv, abs/1811.07871.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of nlp models: A survey. Transactions of the Association for Computational Linguistics, 9:1508–1528.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017. Dialogue Learning With Human-in-the-Loop. In International Conference on Learning Representations.

Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie Cheung, and Siva Reddy. 2022. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. In Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics.

Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. arXiv preprint arXiv:2110.10790.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2060–2069, New Orleans, Louisiana. Association for Computational Linguistics.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023. Languages are Rewards: Hindsight Finetuning using Human Feedback. arXiv preprint arXiv:2302.02676.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. Revista Tradumàtica: tecnologies de la traducció.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve GPT-3 after deployment. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2833–2861, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Aman Madaan, Niket Tandon, Dheeraj Rajagopal, Peter Clark, Yiming Yang, and Eduard Hovy. 2021. Think about it! improving defeasible reasoning by first modeling the question scenario. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6291–6310, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4984–4997, Online. Association for Computational Linguistics.

Nikhil Mehta and Dan Goldwasser. 2019. Improving natural language interaction with robots using advice. arXiv preprint arXiv:1905.04655.

Steven Moore, Huy A Nguyen, and John Stamper. 2020. Towards crowdsourcing the identification of knowledge components. In Proceedings of the Seventh ACM Conference on Learning@ Scale, pages 245–248.

Richard Ngo. 2022. The alignment problem from a deep learning perspective. arXiv preprint arXiv:2209.00626.

Duy-Hung Nguyen, Nguyen Viet Dung Nghiem, Bao-Sinh Nguyen, Dung Tien Tien Le, Shahab Sabahi, Minh-Tien Nguyen, and Hung Le. 2022. Make the most of prior data: A solution for interactive text summarization with preference feedback. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1919–1930, Seattle, United States. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? arXiv preprint arXiv:2010.03532.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. A conversational paradigm for program synthesis. arXiv e-prints, pages arXiv–2203.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Model index for researchers. Accessed: 2023-05-01.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 281–286, Berlin, Germany. Association for Computational Linguistics.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics.

Tatiana Passali, Alexios Gidiotis, Efstathios Chatzikyriakidis, and Grigorios Tsoumakas. 2021. Towards human-centered summarization: A case study on financial news. In Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 21–27, Online. Association for Computational Linguistics.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. arXiv preprint arXiv:2304.01904.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. CoRR, abs/1705.04304.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Lidén, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. ArXiv, abs/2302.12813.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022a. Red teaming language models with language models.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022b. Discovering language model behaviors with model-written evaluations.

Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. 2022. Do users write more insecure code with ai assistants?

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In Proceedings of the Workshop on New Frontiers in Summarization, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Barbara Plank. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2022. T5score: Discriminative fine-tuning of generative evaluation metrics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In Proceedings of the Fifth Conference on Machine Translation, pages 911–920, Online. Association for Computational Linguistics.

Machel Reid and Graham Neubig. 2022. Learning to model editing processes. arXiv preprint arXiv:2205.12374.

Christopher Riesbeck. 1981. Failure-driven reminding for incremental learning. In IJCAI, pages 115–120. Citeseer.

Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow. 1943. Behavior, purpose and teleology. Philosophy of science, 10(1):18–24.

Annabel Rothschild, Justin Booker, Christa Davoll, Jessica Hill, Venise Ivey, Carl DiSalvo, Ben Rydal Shapiro, and Betsy DiSalvo. 2022. Towards fair and pro-social employment of digital pieceworkers for sourcing machine learning training data. In CHI Conference on Human Factors in Computing Systems Extended Abstracts, pages 1–9.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. ACM Comput. Surv., 55(2).

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators.

Roger C Schank. 1983. Dynamic memory: A theory of reminding and learning in computers and people. cambridge university press.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with language feedback.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale.

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. Peer: A collaborative language model. ArXiv, abs/2208.11663.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7881–7892, Online. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, page 614–622, New York, NY, USA. Association for Computing Machinery.

Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. 2021. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3478–3492, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3758–3769, Online. Association for Computational Linguistics.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 203–206.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Teerachart Soratana, Yili Liu, and X Jessie Yang. 2022. Effects of payment rate and country's income level on attitude toward acrowdsourcing task. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 66, pages 2220–2224. SAGE Publications Sage CA: Los Angeles, CA.

Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, New York, NY, USA. Association for Computing Machinery.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback.

Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In Proceedings of the 12th international conference on Intelligent user interfaces, pages 82–91.

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar, Samuel Bowman, and Yoav Artzi. 2021. Crowdsourcing beyond annotation: Case studies in benchmark data collection. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 1–6.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 769–779, Seattle, United States. Association for Computational Linguistics.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Teaching pre-trained models to systematically reason over implicit knowledge. arXiv preprint arXiv:2006.06609, 4(6).

Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. 2022. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 339–352, Seattle, United States. Association for Computational Linguistics.

Andrea L. Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. Artificial Intelligence, 172(6):716–737.

Craig Thomson and Ehud Reiter. 2021. Generation challenges: Results of the accuracy evaluation shared task. In Proceedings of the 14th International Conference on Natural Language Generation, pages 240–248, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the invisible labor in crowd work. Proceedings of the ACM on human-computer interaction, 5(CSCW2):1–26.

Alexander Matt Turner, Aseem Saxena, and Prasad Tadepalli. 2022. Formalizing the problem of side effect regularization. In NeurIPS ML Safety Workshop.

Peter Vamplew, Richard Dazeley, Cameron Foale, Sally Firmin, and Jane Mummery. 2018. Human-aligned artificial intelligence is a multiobjective problem. Ethics and Information Technology, 20:27–40.

Zijie J Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. arXiv preprint arXiv:2103.04044.

Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. Machine learning, 8:279–292.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. arXiv preprint arXiv:2211.00053.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. arXiv preprint arXiv:1410.3916.

Jason E Weston. 2016. Dialog-based language learning. Advances in Neural Information Processing Systems.

Norbert Wiener. 1948. Cybernetics; or control and communication in the animal and the machine.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn., 8(3–4):229–256.

Bin Wu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2018. Query suggestion with feedback memory network. In Proceedings of the 2018 World Wide Web Conference, pages 1563–1571.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2022. Learning New Skills after Deployment: Improving open-domain internet-driven dialogue with human feedback.

Taku Yamagata, Ryan McConville, and Raul Santos-Rodriguez. 2021. Reinforcement learning with feedback from multiple humans with diverse skills.

Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In Conference on Empirical Methods in Natural Language Processing.

Michihiro Yasunaga and Percy Liang. 2020. Graph-based, self-supervised program repair from diagnostic feedback. 37th Int. Conf. Mach. Learn. ICML 2020, PartF168147-14:10730–10739.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. arXiv preprint arXiv:2304.05302.

Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. 2022. Disentangling uncertainty in machine translation evaluation.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. arXiv preprint arXiv:2305.10425.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. 2023. Secrets of rlhf in large language models part i: Ppo.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. CoRR, abs/1909.08593.

Markus Zopf. 2018. Estimating summary quality with pairwise preferences. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1687–1696, New Orleans, Louisiana. Association for Computational Linguistics.