

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

Patrick Fernandes^{1,2,3} Aman Madaan¹ Emmy Liu¹ António Farinhas^{2,3}
 Pedro Henrique Martins⁴ Amanda Bertsch¹ José G. C. de Souza⁴ Shuyan Zhou¹
 Tongshuang Wu¹ Graham Neubig^{1,5} André F. T. Martins^{2,3,4}

¹Carnegie Mellon University ²Instituto Superior Técnico (Lisbon ELLIS Unit)

³Instituto de Telecomunicações ⁴Unbabel ⁵Inspired Cognition

pfernand@cs.cmu.edu

Abstract

Many recent advances in natural language generation have been fueled by training large language models on internet-scale data. However, this paradigm can lead to models that generate toxic, inaccurate, and unhelpful content, and automatic evaluation metrics often fail to identify these behaviors. As models become more capable, *human feedback* is an invaluable signal for evaluating and improving models. This survey aims to provide an overview of the recent research that has leveraged human feedback to improve natural language generation. First, we introduce an encompassing formalization of feedback, and identify and organize existing research into a taxonomy following this formalization. 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 the 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 exploits large language models to make judgments based on a set of principles and minimize the need for human intervention.

feedback (Bai et al., 2022b; Ouyang et al., 2022; OpenAI, 2023a). This feedback serves as a guiding force, steering LLMs 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, when combined with the objective of maximizing the probability of the next token given the previous ones, might result in a *misspecification* of target behavior (Kenton et al., 2021b), and might lead to models that generate toxic, inaccurate, and unhelpful content (Sheng et al., 2019; Bender et al., 2021).

Exacerbating the problem above is the fact that these models are often evaluated using automatic metrics that compare the generated text with some “reference” text using surface-level features (such as word overlap), which often do not correlate with *human-perceived* quality of text (Schluter, 2017; Mathur et al., 2020; Gehrmann et al., 2022a), especially when models are optimized for them (Paulus et al., 2017; Amrhein and Sennrich, 2022). This difficulty in evaluation arises partly because, for many tasks, there is not a single correct answer since the same communicative intent can be conveyed in multiple ways.

Leveraging human assessments to evaluate the quality of texts generated by models is then a popular approach. Crucially, considering human-perceived quality can help close the *gap* between machine and human generated text, and help in addressing the challenges posed by *Goodhart’s law*: “when a measure becomes a target, it ceases to be a good measure” (Goodhart, 1984). This real-

1 Introduction

For generation systems to be widely useful, they must generate text that is not only fluent and high-quality, but also closely aligned with human desires and specifications (Vamplew et al., 2018; Hendrycks et al., 2020; Kenton et al., 2021a; Turner et al., 2022; Ngo, 2022). Achieving such ambitious goals requires modern large language models (LLMs) to evolve beyond traditional training methods. Recent improvements in this space have centered on incorporating human

ization has spurred a growing interest in improving natural language generation systems by leveraging *human feedback* on model-generated outputs, and has led to the emergence of the first widely-used general-purpose language assistants (OpenAI, 2023a). Human feedback not only enhances system performance, but also serves as a mechanism to steer the system in alignment with desired outcomes or goals (Rosenblueth et al., 1943; Wiener, 1948).

Feedback, as a concept, encompasses a wide range of meanings and 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 the notion of *human feedback* and creating a taxonomy of the different types of feedback in the literature, and of how they have been used (§2). We discuss how we can describe feedback by its *format* and its *objective*, in terms of the desired model behavior (§3). We discuss approaches that directly optimize models against human feedback on (their) outputs, for example, using reinforcement learning with human reward functions (§4). We then move to approaches that circumvent the costs of direct feedback optimization by first training *feedback models* to approximate human feedback, and then improving generation using these proxy models (§5). We discuss existing datasets for human-feedback data, how these datasets are typically collected, and the impact that the collection process might have on the behaviour of the models (§6). Finally, we discuss a recent line of work that reduces the need to collect human feedback by leveraging *AI feedback* from large language models (§7).

2 A Taxonomy for Leveraging (Human) Feedback for Generation

2.1 Background

Consider a model $M : \mathcal{X} \rightarrow \mathcal{Y}$ which, given an input of some type $x \in \mathcal{X}$, outputs text $\hat{y} \in \mathcal{Y}$. Importantly, while x can be of any format, we restrict ourselves to cases where y is in the space of *natural language* (i.e., $\mathcal{Y} \subseteq \Sigma^*$ for some alphabet Σ). This general formulation encompasses a wide range of NLG tasks. For example:

- **Summarization:** \mathcal{X} is the space of docu-

ments, and \mathcal{Y} the space of possible summaries.

- **Machine Translation:** \mathcal{X} and \mathcal{Y} are the spaces of sentences in the source and target languages, respectively.
- **Dialog Generation:** \mathcal{X} is the space of possible dialog histories, and \mathcal{Y} is the space of possible responses.
- **Image Captioning:** \mathcal{X} is the space of images, and \mathcal{Y} is the space of possible 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 an input x is decomposed into the product of the probabilities of each token in the sentence, conditioned on the previous tokens. These models are then trained by finding the parameters θ^* that maximize the likelihood of some training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$. Then, at *inference* time, given an input x , an output \hat{y} is decoded from the learned distribution. This decoding can be done, for example, by approximating the most-likely sequence of tokens ($M(x) \approx \arg \max_y P_{\theta^*}(y|x)$) or by random sampling ($M(x) \sim P_{\theta^*}(y|x)$).

Evaluating the quality of generated text $\hat{y} \in \mathcal{Y}$ can be challenging due to the complexity and subjectivity of natural language. Various automatic metrics have been proposed for various domains/tasks. These metrics traditionally rely on n-gram matching or other simple heuristics that cannot account for complex linguistic phenomena (such as paraphrasing or stylistic variations) and often fail to capture all the nuances of human judgment (Sai et al., 2022; Gehrmann et al., 2022a). For this reason, for many of these tasks, asking for *human feedback* is considered the gold standard for assessing the quality of the generated text, and newer *learned* metrics often aim to approximate the way humans provide feedback (see §5.1).

More formally, we consider **human feedback** to be a family of functions \mathcal{H} such that each *feedback function* $h \in \mathcal{H}$ takes an input¹ $x \in \mathcal{X}$ and

¹ Although feedback can be provided independently of the input (for example for *fluency*), we assume some (potentially empty) input for simplicity of notation.

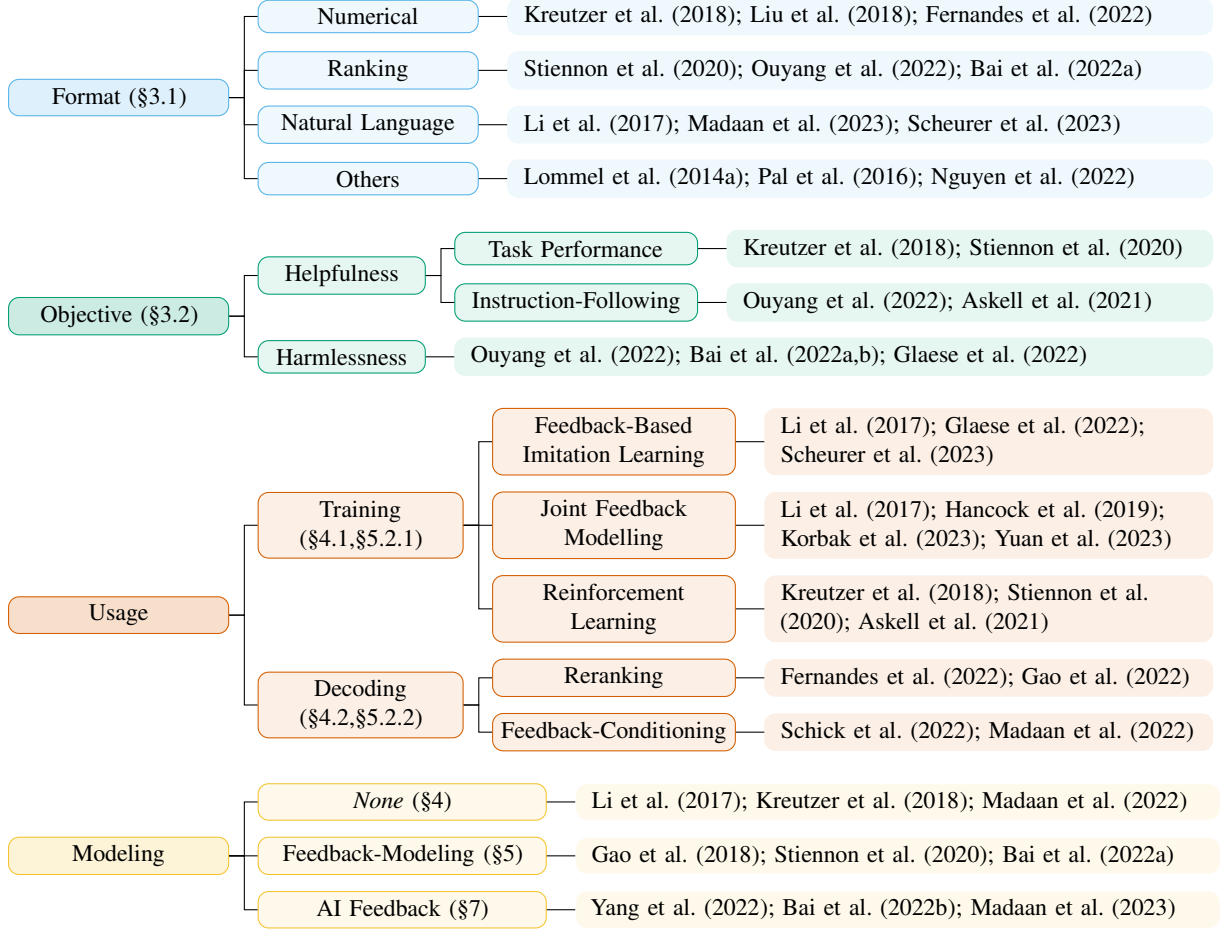


Figure 1: Taxonomy of methods that leverage human-feedback, with some example representative works in the literature that fit in each category.

one or more outputs $y_1, \dots, y_n \in \mathcal{Y}$ and returns some *feedback* $f \in \mathcal{F}$:

$$h : \mathcal{X} \times \underbrace{\mathcal{Y}_1 \times \dots \times \mathcal{Y}_n}_n \rightarrow \mathcal{F}. \quad (1)$$

A simple example of a (human) 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, exist and are commonly used (see §3.1).

We note that this framing is a *simplification* of the real world: often, different humans might provide different (and potentially contradicting) feedback for the same outputs, and a single function might not be able to capture this variability in human opinion (we discuss this 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:

What is the *format* of the feedback? The format of human feedback can vary, including binary judgments, numerical scores, ordinal rankings, or qualitative natural language explanations.

What is its *objective*? Depending on the use case of our model, the feedback can have a variety of purposes, ranging from assessing model performance and accuracy to preventing toxicity and harmful behavior.

When is it *used*? Human feedback can be incorporated into the training stage to optimize the model parameters directly. Alternatively, it can be used at inference time to guide the decoding process.

How is it modeled? While ideally, we would use direct feedback from humans whenever possible, the prohibitive cost of its collection means that it is often useful to instead use *surrogate* models that approximate human preferences.

3 Describing Feedback

3.1 Format

An important decision to make when we want to improve language generation systems through human feedback is in what *format* to collect this feedback in. The choice of format has implications on the expressivity of the feedback, the ease of its collection, and how we can use it to improve systems. In particular, the complexity of the feedback format is an important factor: simpler formats are often easier to collect and use as part of the training/decoding process, but contain less information than more “complex” formats, and might not be able to capture important information for improving the system. The choice of format also has implications in the difficulty for humans to give feedback, its consistency/agreement, and the level of *rationality* of said feedback (Ghosal et al., 2023). Types of feedback 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” that can be assigned to a translation, which are then averaged to produce a score ($\mathcal{N} = [1, 5]$) and used 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 extensively used in the context of evaluation, albeit not with the explicit goal of improving generation. For example, *direct assessments* (Graham et al., 2013) in machine translation (typically) ask humans to rate translations on a continuous scale, and some works have attempted to use this feedback 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 suffers from some limitations: depending on the complexity of the generation task, reducing feed-

back to a single score might generally be a hard and ill-defined task for humans, leading to a costly collection process and problems of *subjectivity* and *variance* (see §6.2.1). Furthermore, such feedback might not be suited to distinguish 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 alternative 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 permutations/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, potentially, for this reason, ranking-based feedback tends to be collected to improve model behavior rather than just for evaluation (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 are trying to improve. Similarly, Ouyang et al. (2022) collected rankings of alternative responses to an *instruction* given to the model. They utilized 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 lack the ability to capture detailed information about problems with the output, which can be crucial for improving generation systems. Instead of asking humans to rank or score outputs, we can instead ask for *natural language* feedback. In such cases, the feedback typically provides more detailed information, either highlighting the shortcomings of the current output or suggesting specific actions for improvement. For example, 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 hinting about it. Tandon et al. (2022) and Madaan et al. (2022) gather natural language feedback on errors present in model-generated graphs and the model’s interpretation of a given instruction. Scheurer et al.

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

(2023) improve summarization capabilities of language models by asking humans to provide natural language feedback of summaries of the model.

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 also other feedback types that haven’t been fully leveraged to improve generation: e.g., *Multidimensional Quality Metrics (MQM)* (Lommel et al., 2014b), the standard for evaluating translation quality, asks professional translators to identify errors *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: we might want our summarization model to generate summaries that contain all core information, even if it means they are a bit longer; in *commercial* machine translation, extra

care is given to ensure that models do not mistranslate business-critical information; and in dialogue agents, we might want the model to be able to produce polite and harmless responses. This **alignment objective** has been studied extensively in the *AI safety and alignment* literature (Bostrom, 2014; Amodei et al., 2016; Bommasani et al., 2021). In addition, Kenton et al. (2021b) discuss some behavioral issues in language agents (natural language generation models) arising from a *misspecified* alignment objective (for example, from noisy labels in the training data), and Leike et al. (2018) proposed using feedback models to tackle the difficulty in specifying this objective.

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).²

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 the task well, and so feedback related to **task performance** generally falls under this umbrella. For example, most works in machine translation leverage feedback related to the quality of translation (Kreutzer et al., 2018; Fernandes et al., 2022), which is expected

²We mostly ignore the proposed *honesty* aspect, as none of these works tackle this directly.

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) (in short, the quality of the summary). One particularly well-studied feedback objective is the ability to **follow instructions** (Ouyang et al., 2022): the task of instruction-following can encompass a wide range of other tasks, and using feedback to improve (instruction following) language assistants has been considered a benchmark for the alignment problem (Askell et al., 2021).

Harmlessness Another important alignment objective is *harmlessness*: we want our models not to produce certain types of output or violate certain 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. (2022b) 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: humans would provide the feedback for training or decoding procedures.

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 score ($f \in \mathbb{R}$), we can create the following optimization problem:

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

Where \mathcal{D} is the distribution of possible inputs. Various techniques have been suggested to op-

timize the model parameters, 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:

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

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

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 as 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 preferred dialogues which comply 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. (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 finetune 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 the generations which do not receive positive feedback, which may contain useful information to optimize the model.

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 approach 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 the following loss of the form

$$\mathcal{L}^{(i)}(\theta) = -\log p_{\theta}(y^{(i)}, f^{(i)} | x^{(i)}) \quad (5)$$

Over all examples in \mathcal{D} . These equation can be factorized as $\mathcal{L}^{(i)}(\theta) = -\log p_{\theta}(f^{(i)} | y^{(i)}, x^{(i)}) + \log p_{\theta}(y^{(i)} | x^{(i)})$. 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 of this approach 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 to the model, or giving a hint about the correct answer) to a dialogue question answering model. Then, after having collected the feedback, the model is trained to predict it. Hancock et al. (2019) proposed having an auxiliary model predicting the satisfaction of the human speaking with the model. Then, 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) showed that having summarization models predict the rankings of different summaries helps the model generate better summaries.

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 the positive and negatively-labeled examples to train a classifier head that directs the model away from generating undesirable sequences. Thoppilan et al. (2022a) follow this approach to enforce the model’s quality and safety. First, they collect dialogues between crowd-workers and the proposed language model LaMDA, which are annotated with feedback provided by the crowd-workers. This feedback states

each response’s quality (sensible, specific, and interesting) or safety. Then, LaMDA is fine-tuned to predict the high-quality responses and the rewards given to every response regarding its quality attributes and safety. At inference time, LaMDA is also used to filter out candidate responses for which its safety prediction is below a threshold.

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

$$\mathcal{L}^{(i)}(\theta) = -\log p_{\theta}(y^i | f^i, x^i) \quad (6)$$

Liu et al. (2023) proposed prompt-based fine-tuning, where they create prompts containing previous generations rated by humans, in the order of preference. They also suggest inserting language-based feedback (e.g., “... is a worse answer than ...”) to the prompt, between the generations. Then, the model is fine-tuned to maximize the likelihood of generating the most preferred answer.

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 the REINFORCE algorithm (Williams, 1992), which updates the policy parameters using the following gradient:

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

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 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 conjunction, these approaches are commonly known

as *Reinforcement Learning from Human Feedback (RLHF)*.

4.2 Decoding with Human Feedback

While directly optimizing model parameters provides greater control, modifying them may not always be feasible, particularly in the case of LLMs. Additionally, feedback might be unavailable during model training, limiting the scope for parameter adjustments. In such cases, leveraging human feedback during decoding plays a critical role in enhancing LLMs’s performance. This type of feedback, derived from interactions between LLMs and users in practical scenarios, enables models to learn from their errors and offers opportunities for ongoing refinement without altering model parameters. In addition, the feedback functions as a guiding mechanism, allowing the model to generate more desirable outputs by leveraging its existing capabilities.

There are two broad categories in which human feedback is used in this setup: 1. *Feedback Memory*: Feedback Memory Utilization involves maintaining a repository of feedback from prior sessions. Then, when processing new inputs, the system uses relevant feedback from similar inputs in its memory to guide the model toward generating more desirable outputs based on past experiences and user preferences. While a classical concept (Riesbeck, 1981; Schank, 1983), recent work has shown the promise of such a memory-augmented approach in both finetuning (Weston et al., 2014; Wu et al., 2018; Tandon et al., 2022) and few-shot setups (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 providing feedback on the outputs.

These two techniques are not mutually exclusive. They can be combined to achieve even better performance, creating a more adaptive and respon-

sive system that caters to user expectations and requirements.

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 alternative approach to obtaining human feedback is to develop models that can predict or approximate it. Although these models may not be perfect, they offer the advantage of providing feedback at a low cost after training, thereby enabling the scaling of feedback-dependent techniques.

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

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

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

For example, if the feedback function we are trying to model is also numerical ($h : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$), 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* models³ $\hat{h}_\phi(x, y_n)$ 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 subsequently been used in other works (Ouyang et al., 2022; Askell et al.,

³We 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.

2021; Liu et al., 2022; Qin et al., 2022; Yuan et al., 2023).

The problem of feedback modeling has been studied extensively in the context of *metric learning* for NLP. Zhang et al. (2019) and Zhou et al. (2023) utilized pre-trained masked LMs to compute similarity scores between the generated text or code snippets and their references. In MT, Sellam et al. (2020) and Rei et al. (2020a) trained BLEURT and COMET, respectively, to regress on human quality assessments of translation quality. 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 included in 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). It is notable that these reward models were not trained with the intent of improving generation directly, although some of them were used for that purpose later, as discussed in §5.2.

Recently, there has been a growing interest in developing feedback models directly with the aim of using them to improve generation (Böhm et al., 2019; Ziegler et al., 2019). As a first step, these models are typically initialized with weights from either the target LM that requires improvement or from a model of the same family (e.g., of a smaller size) (Askell et al., 2021; Bai et al., 2022a; Ouyang et al., 2022). One key consideration in the initialization is the size of the pretrained model: while scaling up may improve overall performance (Askell et al., 2021; Bai et al., 2022a), Ouyang et al. (2022) find that larger models may be less stable for future finetuning.

Next, the feedback model is finetuned on a dataset of human feedback. This dataset is typically collected by asking annotators to provide feedback on outputs from an earlier version of the model being improved. However, it is also possible to first finetune the feedback model on naturally occurring implicit feedback, such as from user interactions on websites (e.g., Reddit, Stack-Overflow). Though less accurate than explicitly-collected feedback, it allows feedback models to be trained on much more data. Askell et al. (2021) found that naturally occurring feedback data benefits models larger than 1B parameters, but of-

ten has diminishing returns when the number of explicit-collected feedback increases.

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)), trained on human post-edits with the intent of automatically correcting the output of an MT system, can also be seen as feedback models (albeit non-numerical).

5.2 Leveraging Feedback Models to Improve Generation

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

5.2.1 Optimizing for Feedback Models

Similarly to optimizing for human feedback, one possible way to use the feedback model is to optimize model parameters 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 to the limitations of feedback models as imperfect proxies, typically a *regularization* term R is introduced to avoid “overfitting” to the feedback model (Ziegler et al., 2019) (more on this at the end of this section):

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

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, while this section is focused on cases where a model of human feedback is used instead.

Unlike when using human feedback directly, most works attempt to optimize for feedback models using **reinforcement learning**. Gao et al. (2018); Böhm et al. (2019) use the (numerical) feedback collected in other works to train reward and preference models, and use reinforcement

learning to optimize against these models, showing 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 using feedback collected on the model being improved, and introduced a KL regularization term

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

to avoid the optimized model deviating too much from the original (supervised) model with parameters θ_{SL} ⁴. 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 improve the ability of LLMs to follow instructions, but combined the RL objective with the original pretraining objective to avoid performance regressions in public NLP benchmarks. Other works have also used reinforcement learning with preference models in a similar manner (Askell et al., 2021; Bai et al., 2022a; Wu et al., 2021; Nguyen et al., 2022). Underlying all these methods is that generally the model is first trained with imitation-learning on human demonstrations, which improves performance compared to using reinforcement learning 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.

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.

In an approach outside these main categories, Peyrard and Gurevych (2018) use a scoring function learned from human judgments as a fitness function for a genetic algorithm to generate summaries of input texts.

⁴Note that this KL term is different from other algorithm-specific regularization terms, such as the KL terms in PPO (Schulman et al., 2017).

5.2.2 Decoding with Feedback Models

As mentioned, feedback models have the advantage that they can be queried cheaply for feedback once trained. Perhaps for this reason, most approaches that leverage feedback models by sampling a large number of candidate generations, and reranking them according to the feedback model:

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

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

where \hat{h}_ϕ is a trained (numerical) feedback model and \mathcal{C} is a set of S candidate generations given by the model (for example, by sampling from its distribution multiple times).

In machine translation, Fernandes et al. (2022) and Freitag et al. (2022a) build upon recent advances in automatic quality estimation and evaluation via feedback model training to improve generation. Their framework comprises a candidate generation stage followed by a ranking stage, in which the candidates are scored using quality metrics trained to regress on human assessments (reward models) (Rei et al., 2020a,b) via N -best list reranking or minimum Bayes risk (MBR) decoding (Kumar and Byrne, 2002). The highest-scoring candidate is then chosen as the final translation.

Gao et al. (2022) also used this approach to study the scaling properties of feedback models and the problem of "overoptimization" (see below).

Additionally, there are several works combining MT and APE systems at decoding time, in which the output of an MT system is further improved by an APE system (Bhattacharyya et al., 2022).

Feedback Model Overoptimization One problem that arises when optimizing a system with a feedback model is that this model is only an imperfect proxy for the ground truth human feedback, therefore, "overoptimizing" for them can lead to systems that receive good feedback from the model, but not humans. This problem 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 reinforcement learning (training) and reranking outputs with it (decoding). They found that both using preference models during

training or decoding led to similar levels of overoptimization, and that the scale of the generation model helps little with this problem.

6 Collecting and Using Human Feedback

Collecting human feedback can be rather expensive and may present issues for the inexperienced, making it important to leverage existing resources and consider additional data collection carefully. We present an introduction to existing datasets and their collection methods, along with considerations for experimenters creating preference datasets for their own use cases. Additionally, we discuss ethical considerations in the use and collection of human feedback.

In future, richer types of feedback may be collected and we may find ways to make use of this signal. For instance, most existing datasets consist of ranking or numerical scores, but humans prefer to provide richer feedback than labelling (Stumpf et al., 2007; Amershi et al., 2014a; Ghai et al., 2021). Furthermore, variability between human annotators has also not been fully explored (Plank, 2022; Gehrmann et al., 2022b).

6.1 Considerations in Data Collection

There are multiple facets to consider when collecting human feedback data for a generation task; a non-exhaustive list of axes along which data collection can vary is presented below.

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.
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. **Collection method:** Data can be gathered explicitly through experiments or implicitly from online sources or user interactions, with varying noise levels (Kreutzer et al., 2018; Freitag et al., 2021).
4. **Collection platform:** Common platforms include Amazon Mechanical Turk, Upwork, and Scale AI.
5. **Annotator demographics:** Different groups may have varying opinions on quality generations; demographics may be collected during

data collection.

There is generally a trade-off between the effort needed to create the datasets and the reliability of judgments collected. For higher-stakes applications in specific domains, it may be worth the effort to consult expert annotators in an extended partnership. For general alignment with human preferences, it may instead be prudent to recruit a diverse group of annotators to avoid overfitting to the preferences of specific demographics that may be more accessible in recruitment.

6.2 Pitfalls and Ethical Considerations of Human Feedback

Although we have focused on the idealized form of human feedback in §2.1, actual feedback may be low-quality, contradictory, or adversarial.⁵ As discussed in §3.2, we must carefully specify annotation guidelines so that feedback is aligned towards the actual goals for the model (Ziegler et al., 2019). Even in the case where human experts are available, different groups of experts may not agree (Kahneman et al., 2021). In this section, we enumerate possible issues with human feedback, most of which are shared with other annotation tasks. We also touch on possible mitigation strategies.

6.2.1 Subjectivity and variance in judgment

Considering K annotators with feedback functions $h_{i=1}^K$, judgments are given on data $\mathcal{D} = d_1, \dots, d_N$. 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; Cohen, 1960). Low reliability may result from unclear tasks or evaluation criteria (Gehrmann et al., 2022b; Thomson and Reiter, 2021), 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), and augmenting evaluation metrics like COMET with confidence intervals (Glushkova et al., 2021; Zerva et al., 2022). Clear annotation guidelines and including rationales with rankings can reduce biases and improve clarity (Ziegler et al., 2019).

⁵By adversarial feedback, we mean feedback that intentionally inverts a user’s preferences, or is designed to mislead a model in some systematic way, rather than just noisy data.

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
Translation	WMT Metrics Shared Task (Freitag et al., 2022b)	Explicit	Pro translation workflow	MQM, DA
Summarization	TAC Shared Tasks (TAC-2008, TAC-2009)	Explicit	N/A	Score

Table 2: Summary of existing human feedback datasets and their collection methods, which vary along several dimensions. Refer to Table 1 for definitions related to feedback types. A separation is drawn between datasets that were explicitly designed to capture human preferences in a general sense, and datasets designed for more specific use cases, such as MQM/DA datasets in MT. N/A means we could not find information.

6.2.2 Bias in judgment

Even if all K annotators agree on a particular judgment for a certain data point, they may all be mistaken. There are well-known biases in human reasoning which may cause all annotators or a large percentage of annotators to be mistaken, or not take evidence into account. Furthermore, even if annotators are technically unbiased in terms of the task they were instructed to evaluate, instructions can be underspecified or lead the annotators to evaluate a slightly different task, leading to the appearance of systematic bias away from the originally intended task (Parmar et al., 2023).

Anchoring/Confirmation bias: When annotators are presented with a text in isolation, they may fail to consider better alternatives and erroneously label the text as high-quality (Bansal et al., 2021). When asked to generate text, anchoring bias can cause people to write in a different manner than usual (Jakesch et al., 2023; Lehmann et al., 2022), which may influence what types of suggestions or corrections they give. Mitigation strategies include asking people to rank several diverse outputs and being explicit about the dimensions people are asked to evaluate.

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

6.2.3 Ethical considerations

Some subjectivity in annotator judgment can arise from differences across cultural or social groups. Santurkar et al. (2023) measure opinions in language model generations, demonstrating varying

degrees of representation of demographic groups. Several works observe that tuning with human feedback increases the alignment of generated outputs with US liberal views on controversial topics (Perez et al. (2022b), Hartmann et al. (2023)). Annotators with different demographic or political backgrounds may disagree on what qualifies as toxic content (Sap et al. (2022), Ding et al. (2022)). This is particularly pronounced when annotators are asked to make ethical judgments, which may vary with cultural context and personal sensibilities (Jiang et al. (2022), Talat et al. (2022)).

Steiger et al. (2021) survey moderators of toxic content, identifying harms ranging from slight discomfort to lasting psychological harm from the prolonged performance of content moderation tasks; however, the severity and frequency of toxic content examined in content moderation likely exceeds that in other types of human feedback annotation. Shmueli et al. (2021) identify toxicity classification and generation from open-ended inputs as two NLP annotation tasks that may trigger harmful responses in annotators. They further argue that this moves beyond the “minimal risk” requirement for Institutional Review Board exemption in the United States and encourage academic researchers using crowdworker annotation to file for this ethical review of their work.

Media attention has also focused on fair pay for annotators, with one *TIME* article⁶ describing annotators paid \$2 USD or less per hour to review toxic content and provide harmfulness annotations for model training. 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 inad-

⁶<https://time.com/6247678/openai-chatgpt-kenya-work>

equate pay, especially for workers in lower-resourced regions, can be a form of worker exploitation.

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 minor improvements, with outcomes approximating chance. Moreover, employing static feedback can create consistency and accuracy challenges, as the integration of feedback leads to changes in the model’s output distribution.

AI-generated feedback, an emerging research area, focuses on harnessing the large language model’s own abilities to evaluate and improve its output, enhancing the model without constant 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. Intriguingly, the capacity to offer valuable AI feedback may depend on the model being trained with human feedback.

External AI Feedback: The second approach employs a separate model to provide feedback on the model’s outputs which is being improved. In this setting, the task model is often paired with a separately trained feedback model (Yasunaga and Liang, 2020; Madaan et al., 2021; Welleck et al., 2022; Bai et al., 2022b). An advantage of this approach is that the feedback model does not need to be a large, general-purpose model like GPT-4. Thus, training smaller feedback models becomes an attractive alternative when a large amount of feedback is available.

8 Conclusion

In this paper, we provided an overview of recent research that has leveraged human feedback to improve natural language generation, highlighting different ways it can be defined, collected and leveraged, and respective advantages and disadvantages. 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 hope this survey can help researchers understand the current state of the art, and identify new and existing sources of feedback and ways of leveraging it.

Acknowledgments

This work was supported by EU’s Horizon Europe Research and Innovation Actions (UTTER, contract 101070631) and by the projects MAIA and NextGenAI (LISBOA-01-0247-FEDER-045909 and 2022-C05i0102-02).

References

- Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014a. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120.
- Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014b. 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, Avaniika 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.
- Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. 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 Groeneweld, 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.
- 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, Chiklu 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, Abhinava 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. 2022a. 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. 2022b. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. arXiv preprint arXiv:2202.06935.
- 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.
- 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.
- 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.

- Daniel Kahneman, Sibony Olivier, and Cass R. Sunstein. 2021. *Little, Brown Spark*, New York.
- Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021a. Alignment of language agents. arXiv preprint arXiv:2103.14659.
- Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021b. 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.
- 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.
- 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.
- 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.
- 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 Fine-tuning 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, Aljoscha Burchardt, and Hans Uszkoreit. 2014a. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. Tradumàtica: tecnologies de la traducció, 0:455–463.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014b. 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.
- 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.
- 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, Kamilè Lukošiušė, 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.
- 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.
- Maxime Peyrard and Iryna Gurevych. 2018. Objective function learning to match human judgments for optimization-based summarization. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 654–660, New Orleans, Louisiana. Association for Computational Linguistics.
- Barbara Plank. 2022. The ‘problem’ of human label variation: On ground truth in data, modeling and evaluation.
- Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2022. T5score: Discriminative fine-tuning of generative evaluation metrics.
- 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 DiS-alvo, Ben Rydal Shapiro, and Betsy DiS-alvo. 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, 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.
- Weiyang 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 transla-

- tion 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 crowdsourcing 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 Venkataragi, 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.
- Zeera 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. 2022a. Lambda: Language models for dialog applications.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna,

- Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022b. Lamda: Language models for dialog applications. CoRR, abs/2201.08239.
- 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.
- 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.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.
- Shuyan Zhou, Uri Alon, Sumit Agarwal, and Graham Neubig. 2023. Codebertscore: Evaluating code generation with pretrained models of code. arXiv preprint arXiv:2302.05527.
- 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.