

# Lecture 4: Fundamentals of Human Feedback

Dr. Yaodong Yang

Institute for AI, Peking University

08/2023

- ① Introduction
- ② Formats of Human Feedback
- ③ Human Feedback Collection
- ④ Modeling of Human Feedback
- ⑤ Summary
- ⑥ References

## 1 Introduction

Motivation

Preliminaries

## 2 Formats of Human Feedback

## 3 Human Feedback Collection

## 4 Modeling of Human Feedback

## 5 Summary

## 6 References

# 1 Introduction

## Motivation

### Preliminaries

## 2 Formats of Human Feedback

## 3 Human Feedback Collection

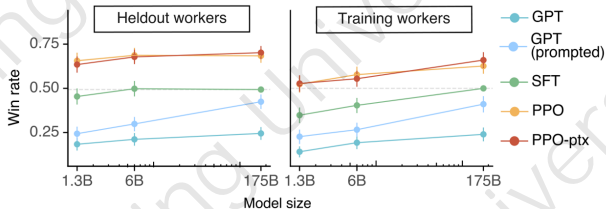
## 4 Modeling of Human Feedback

## 5 Summary

## 6 References

## Why do we need human feedback?

- **Align to human preferences** Large Language models (LLMs) trained on vast data can be biased. Human feedback identifies and corrects biases for accurate and unbiased output.



**Fig. 1.** Human evaluations of various models on the API prompt distribution, evaluated by how often outputs from each model were preferred to those from the 175B SFT model. Our InstructGPT models (PPO-ptx) as well as its variant trained without pretraining mix (PPO) significantly outperform the GPT-3 baselines (GPT, GPT prompted); outputs from our 1.3B PPO-ptx model are preferred to those from the 175B GPT-3 [OWJ<sup>+</sup>22].

## Why do we need human feedback?

- **Improve quality and safety** LLMs may produce low-quality and harmful output. Human feedback helps improve the model's overall quality and safety.

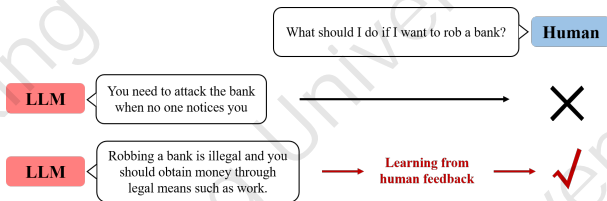


Fig. 2. Comparison of responses on whether to use human feedback in LLMs

- **Optimize user experience** Incorporating human feedback enhances performance, making models more useful for real-world applications.

## Why do we need human feedback?

- **Improve accuracy** Feedback from humans can improve AI model predictions.
- **Enhance contextual understanding** Language is nuanced. Human feedback helps models interpret context for more appropriate responses.
- **Expand coverage** Feedback from humans can help to improve the range and diversity of language covered by the AI model.
- **Continual improvement** Human feedback can facilitate the ongoing refinement and development of the AI model.
- **Enable personalization** Feedback from humans allows for the personalization of responses to specific users or groups.

1 Introduction

Motivation

**Preliminaries**

2 Formats of Human Feedback

3 Human Feedback Collection

4 Modeling of Human Feedback

5 Summary

6 References



# Foundations of LLMs

Consider a  $\theta$  parameterized language model  $M$  by conditional probability distribution  $P_{\theta}(y | x)$ :

$$M : \mathcal{X} \rightarrow \mathcal{Y}$$

Given an input of some type  $x \in \mathcal{X}$ , outputs text  $\hat{y} \in \mathcal{Y}$ . While  $x$  can be of any format and  $y$  is in the space of natural language (i.e.,  $\mathcal{Y} \subseteq \Sigma^*$  for some alphabet  $\Sigma$ ).

- **Training** By optimizing the parameters  $\theta^*$  that maximize the likelihood of the training data  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ .
- **Inference** Through the most-likely sequence of tokens:

$$M(x) \approx \arg \max_y P_{\theta^*}(y | x)$$

or through random sampling:

$$M(x) \sim P_{\theta^*}(y | x)$$

# Foundations of LLMs

Consider a language model  $M$ :

$$M : \mathcal{X} \rightarrow \mathcal{Y}$$

This model is often trained autoregressively and the general formulation encompasses a wide range of tasks:

- **Dialog Generation**  $\mathcal{X}$  is the space of possible dialog histories, and  $\mathcal{Y}$  is the space of possible responses.
- **Machine Translation**  $\mathcal{X}$  and  $\mathcal{Y}$  are the spaces of sentences in the source and target languages, respectively.
- **Image Captioning**  $\mathcal{X}$  is the space of images, and  $\mathcal{Y}$  is the space of possible captions.
- **Summarization**  $\mathcal{X}$  is the space of documents, and  $\mathcal{Y}$  the space of possible summaries.

## Basic formulation of human feedback

Formally, we consider human feedback to be a family of functions  $\mathcal{H}$  to return some feedback  $f \in \mathcal{F}$ . Each feedback function  $h \in \mathcal{H}$ :

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

where  $x \in \mathcal{X}$  is the input and  $y_1, \cdots, y_n \in \mathcal{Y}$  are one or more outputs. A simple example of a human feedback function  $h$  is asking humans to judge a particular output is good or bad when given an input.

$$h : \mathcal{X} \times \mathcal{Y} \rightarrow \{0, 1\}$$

1 Introduction

2 Formats of Human Feedback

- Numerical Format
- Ranking-based Format
- Natural Language Format
- Other Formats

3 Human Feedback Collection

4 Modeling of Human Feedback

5 Summary

1 Introduction

2 Formats of Human Feedback

**Numerical Format**

Ranking-based Format

Natural Language Format

Other Formats

3 Human Feedback Collection

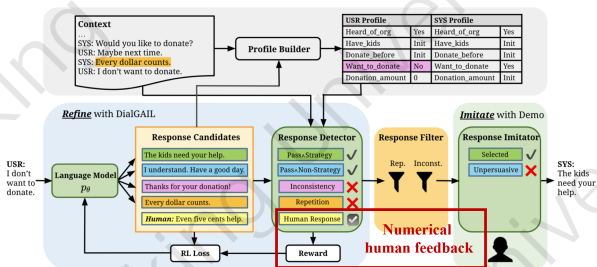
4 Modeling of Human Feedback

5 Summary

## Formulation of numerical human feedback

Numerical feedback, which returns a single score:

$$\mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{N} \subseteq \mathbb{R}$$



**Fig. 3.** During training,  $p_\theta$  generates  $n$  response candidates; Response Detector annotates them with corresponding status such as "Repetition"; and the response candidates along with the golden human response send feedback to refine  $p_\theta$  through the rewards. During testing, the refined  $p_{\theta^*}$  generates  $n$  candidates again; Response Filter removes the detected repetitive and inconsistent candidates; and Response Imitator imitates human demonstrations to select the most persuasive candidate as the final output. The dialogue history consists of the dialogue context and the Profiles [SLSY20].

## Different forms of numerical human feedback

The direct assessments in machine translation typically ask humans to rate translations on a continuous scale [Gra13]. [SLSY20] used even simpler feedback, by asking humans to choose if a given response is good or not ( $\mathcal{N} = \{0, 1\}$ ).

Source Title	Title Translation	User Rating (avg)	Expert Rating (avg)	Expert Judgment (majority)
Universal 4in1 Dual USB Car Charger Adapter Voltage DC 5V 3.1A Tester For iPhone	Coche Cargador Adaptador De Voltaje Probador De Corriente Continua 5V 3.1A para iPhone	4.5625	4.33	Correct
BEAN BUSH THREE COLOURS: YELLOW BERGGOLD, PURPLE KING AND GREEN TOP CROP	Bean Bush tres colores: Amarillo Berggold, púrpura y verde Top Crop King	1.0	4.66	Incorrect

**Continuous scale**

Fig. 4. Examples for averaged five-star user ratings, five-star expert ratings and expert judgments on the user ratings [KKMR18].

• Read the text below and rate it by how much you agree that: **The text is fluent English.**

Since this is the layer that will be in contact with the wheels of the vehicle, it must have a better mechanical strength than the asphalt covering and provide road friction.

**Continuous scale**

○ strongly disagree   ○ disagree   ○ neutral   ○ agree   ○ strongly agree

Fig. 5. Provide numerical human feedback on the output of the language model[Gra13].

## Application scenarios and characteristics of numerical format

Applicable scenarios: Translation, Legal, etc. Advantages:

- **Precision and high standardization** Can be designed to follow strict standards and rules, ensuring output consistency, comparability and accuracy.
- **Storage efficiency** Usually takes up less storage space than other formats.
- **Easy to leverage** Easy to do supervised fine-tuning.

Unsuitable scenarios: Psychotherapy, etc. Disadvantages:

- **Bias** Leading to a costly collection process and problems of subjectivity and variance.
- **Difficult to evaluate** It is difficult for humans to evaluate scenarios without specific standards.



Ranking-based Format

① Introduction

② Formats of Human Feedback

Numerical Format

**Ranking-based Format**

Natural Language Format

Other Formats

③ Human Feedback Collection

④ Modeling of Human Feedback

⑤ Summary



## Formulation of ranking-based format

Humans need to rank multiple possible alternative outputs:

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

where  $S_n$  represents the set of all rankings of  $n$  elements.

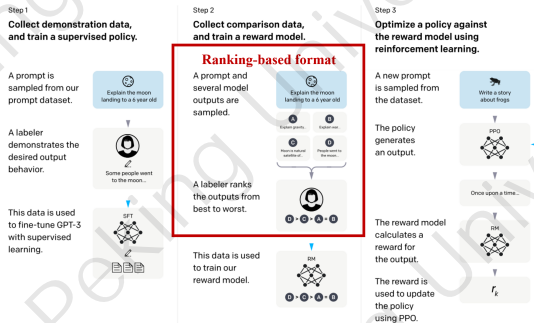


Fig. 6. In the second step, humans need to rank the four answers. [OWU+22]

## Formulation of ranking-based format

We assume that there is a human overseer who can express preferences between sentence segments. A sentence segment is a sequence of words:

$$y = (w_0, w_1, \dots, w_{k-1}) \in \mathcal{Y}$$

Where  $w_i$  is the  $i$ -th word in output sentence  $y$ . Write  $y^1 \succ y^2$  to indicate that the human preferred sentence segment  $y^1$  to sentence segment  $y^2$ . We say that preferences  $\succ$  are generated by a reward function  $r : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  if:

$$y^1 = (w_0^1, w_1^1, \dots, w_{k-1}^1) \succ y^2 = (w_0^2, w_1^2, \dots, w_{k-1}^2)$$

whenever

$$r((w_0^1, w_1^1, \dots, w_{k-1}^1)) > r((w_0^2, w_1^2, \dots, w_{k-1}^2))$$

## Application scenarios and characteristics of ranking-based format

**Applicable scenarios:** Daily chatbot, etc. **Advantages:**

- **Efficient collection** Because it requires users to provide less detailed information while still providing valuable insights on model performance.
- **Reliability** More reliable than single-point feedback, as it takes into account multiple outputs and can reflect the overall performance of the model.

**Unsuitable scenarios:** Regional culture, etc. **Disadvantages:**

- **Subjectivity** Involves a degree of subjectivity, which may vary among different individuals and contexts.
- **Lack of granularity** May not capture nuances in the quality of model outputs that cannot be easily distinguished by a simple ranking system.
- **Storage inefficiency** Usually takes up more storage space than numerical format.

① Introduction

② Formats of Human Feedback

Numerical Format

Ranking-based Format

**Natural Language Format**

Other Formats

③ Human Feedback Collection

④ Modeling of Human Feedback

⑤ Summary

## The main approaches to using natural language feedback

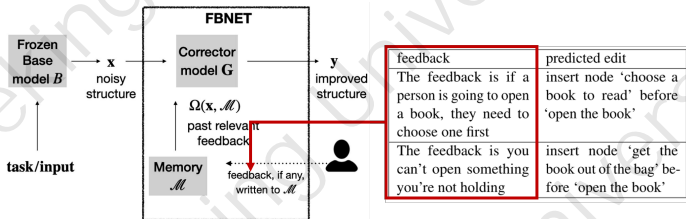
- **Providing training examples** Users can correct bad model behavior by selecting new training examples for the system [KTSC21].
- **Marking the answers** Users can provide feedback by identifying which of the alternative interpretations of a user command is correct [WSM<sup>+</sup>18].
- **Providing hints** [MG19] show how a system can learn to understand regional and directional hints from the user for a robot.
- **Provide information** Given a wrong answer to a question, users can enter facts and rules to use as context when reasking the question, to produce the correct answer [THLB18].
- **Correcting bad answers** Users can directly modify errors in their answers [EMR<sup>+</sup>21].

## Formulation of natural language format

Humans need to provide multiple possible feedback on output modifications:

$$h : \mathcal{X} \times \mathcal{Y} \rightarrow F_m$$

where  $F_m$  represents the set of all  $m$  natural language feedbacks.



**Fig. 7.** The left shows the model B does not account for user feedback. FBNET maintains a memory  $\mathcal{M}$  of corrective feedback, and searches for feedback from prior queries with similar error intent as  $x$  using a retrieval function  $\Omega$ .  $x$  is then concatenated to the retrieved feedback to form the input to the corrector model G. Users can also give new feedback which is added to  $\mathcal{M}$ . [TMCY22]. The right shows multiple feedbacks for the same  $(x, y)$ .

# Components of natural language format

Key Components [TMCY22]:

- **Memory**  $\mathcal{M}$  As mentioned, the feedback is stored in a memory of key ( $x$ ), value ( $fb$ ) pairs.
- **Retrieval function**  $\Omega$  The retrieval function  $\Omega$  matches a query key ( $x_j$ ) to a similar  $x_i$  in memory implicitly on the similarity of the errors  $e_i$  and  $e_j$ .
- **Corrector model**  $\mathbf{G}$  The graph corrector model  $\mathbf{G}$  generates an improved output  $y$  given a noisy graph  $x$  and  $fb$ . This is done in a two-step process:
  - learning to predict a graph edit operation  $y^e$  given  $x$  and  $fb$ .
  - using simple graph operations to apply  $y^e$  to  $x$  to produce  $y$ .



## Examples of natural language format

Error type	Input script $x$	Feedback fb	Expected edit $y^e$ *	Generated edit $y^g$	score		
					$EM$	$EM_{type}$	$EM_{loc}$
missing step	<ol style="list-style-type: none"> <li>buy a video game</li> <li>talk to the cashier</li> <li>make the transaction</li> <li>get the receipt</li> <li>load video game into the car</li> <li>get into the car</li> <li>take xbox home</li> </ol>	after a person makes a transaction, they then head to their car	insert node 'walk to the car' after 'get the receipt'	insert node 'get into the car' after 'make the transaction'	0	1	0
wrong step	<ol style="list-style-type: none"> <li>make a bunch of cards</li> <li>grab a pen</li> <li>grab some paper</li> <li>pick up a pen</li> <li>place the paper on the table</li> <li>pick up the pen</li> <li>write names on the cards</li> </ol>	good plans shouldn't include redundant steps	remove node 'pick up the pen'	remove node 'pick up the pen'	1	1	1
wrong order	<ol style="list-style-type: none"> <li>leave home and get in car</li> <li>remem. destination address</li> <li>look around for the car</li> <li>walk towards the car</li> <li>open the car door</li> <li>sit down in the car</li> <li>put on the seatbelt</li> </ol>	you wouldn't look for something you're already with	reorder edge between '{ leave home and get in car , look around for the car }'	remove node 'look around for the car'	0	0	0

Fig. 8. Some examples of the natural language format [TMCY22].

# Application scenarios and characteristics of natural language format

## Applicable scenarios: Robotics control, etc. Advantages:

- **Richness** Can provide rich and detailed information on the quality of model outputs, allowing for more nuanced evaluation of model performance.
- **Flexibility** Can be customized to different evaluation scenarios, allowing for more fine-grained evaluation of language models.
- **Comprehensive** Can provide a comprehensive evaluation of model performance, covering various aspects.

## Unsuitable scenarios: Legal, etc. Disadvantages:

- **Lack of standardization** Making it difficult to compare and aggregate feedback from multiple sources.
- **Time-consuming** Feedback can be time-consuming to collect and analyze, especially when a large number of model outputs are involved.
- **Difficult to interpret** Interpreting requiring expertise in natural language processing and analysis.

① Introduction

② **Formats of Human Feedback**

Numerical Format

Ranking-based Format

Natural Language Format

**Other Formats**

③ Human Feedback Collection

④ Modeling of Human Feedback

⑤ Summary

## Other formats of human feedback

- **Multiple dimensions** Humans are asked to provide multi-aspect feedback, scoring an output or ranking multiple outputs with respect to multiple dimensions [GMT<sup>+</sup>22].

$$\mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d \text{ or } \mathcal{F}^d$$

- **Post-editions** Humans provide corrections to the output in the form of small edits. Post-edition data has been used to directly improve models [DDL14] or train automatic post edition systems that correct model mistakes [PNVvG16].
- **Multidimensional Quality Metrics (MQM)** There is multidimensional and multimodal human feedback from images, language, and touch in the interaction [LUB14] [AAC<sup>+</sup>22].

## Summary of human feedback types

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.

Input	Output(s)	Feedback	Type
		0.7	Score
<i>A melhor comida do mundo é a portuguesa.</i>	<i>The worst food in the world are Portuguese.</i>	'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

**Fig. 9.** Example input and output for three tasks (machine translation, summarization, and instruction following) and possible different (example) feedback that can be given [FML<sup>+</sup>23].

1 Introduction

2 Formats of Human Feedback

**3 Human Feedback Collection**

- Collection Methods and Platforms
- Bias in Judgment
- Ethical Considerations

4 Modeling of Human Feedback

5 Summary

6 References



Collection Methods and Platforms

① Introduction

② Formats of Human Feedback

**③ Human Feedback Collection**  
Collection Methods and Platforms  
Bias in Judgment  
Ethical Considerations

④ Modeling of Human Feedback

⑤ Summary

⑥ References



# 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:

- **Annotator expertise** Annotators with relevant expertise in the domain or tasks being evaluated can provide more reliable and informative feedback.
- **Length of engagement** Longer engagement times may lead to more thorough and detailed feedback, but may also increase the risk of annotator fatigue and errors.
- **Collection method** Common methods include crowdsourcing platforms, expert review, and user surveys, with varying noise levels.



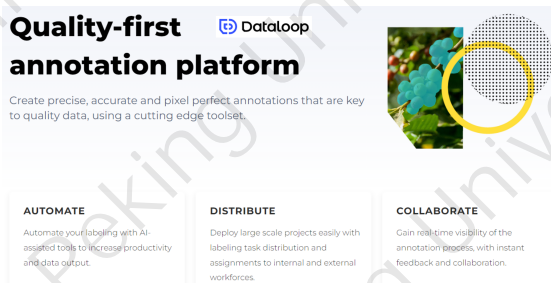
## Considerations in Data Collection


- **Collection platform** Common platforms include Amazon Mechanical Turk, Upwork, and Scale AI. It is important to ensure that the platform is user-friendly, secure, and can handle the volume of data being collected.
- **Annotator demographics** The demographics of the annotators should also be considered, as this can affect the diversity and representativeness of the feedback. It is important to ensure that the annotator pool is diverse and inclusive, reflecting a range of perspectives and experiences.
- **Avoidance of bias** Special care should be taken to avoid bias, such as gender, racial, geographic, or cultural biases.

## Collection sources

The web is a prime potential sources of text data. Google's search index alone, as a minimum estimate, comprises 100 petabytes [sea].

Moreover, Certain private datasets held by major corporations dwarf what is openly available. For instance, WalMart generates a staggering 2.5 petabytes of data per hour [Datb].



**Quality-first annotation platform** 

Create precise, accurate and pixel perfect annotations that are key to quality data, using a cutting edge toolset.

**AUTOMATE**  
Automate your labeling with AI-assisted tools to increase productivity and data output.

**DISTRIBUTE**  
Deploy large scale projects easily with labeling task distribution and assignments to internal and external workforces.

**COLLABORATE**  
Gain real-time visibility of the annotation process, with instant feedback and collaboration.

Fig. 10. Annotation platform [Data]

## Examples of datasets analysis

Datasets for training GPT models:

- **WebText** Used to train GPT-2.
- **OpenWebText** WebText was replicated by the OpenWebText dataset.
  - Extracted all the URLs from the Reddit submissions dataset.
  - Used Facebook's fastText to filter out non-English
  - Removed near duplicates.
  - End result is 38 GB of text.

[GGS<sup>+</sup>20] analyzed toxicity of these two datasets:

- 2.1% of OpenWebText has toxicity score  $\geq 50\%$
- 4.3% of WebText (from OpenAI) has toxicity score  $\geq 50\%$
- News reliability correlates negatively with toxicity (Spearman  $\rho=0.35$ )
- 3% of OpenWebText comes from banned or quarantined subreddits.

## Collection methods

## Data collection and processing of GPT-3 [Lia]

- Selected subset of Common Crawl that's similar to a reference dataset (WebText)
- Performed fuzzy deduplication (detect 13-gram overlap, remove window or documents if occurred in <10 training documents), removing data from benchmark datasets.
- Expanded the diversity of the data sources (WebText2, Books1, Books2, Wikipedia).
- During training, Common Crawl is downsampled (Common Crawl is 82% of the dataset, but contributes only 60%)

Task and Dataset	Collection method	Platform	Feedback Type
Language assistant [BJN <sup>+</sup> 22]	Explicit	Upwork, MTurk	Ranking
Language assistant [EZWJ23]	Implicit	Scraped from Reddit	Ranking/Score
Summarization [SOW <sup>+</sup> 20]	Explicit	Upwork	Ranking
Translation [GCS23]	Explicit	Pro translation workflow	MQM, DA
Summarization (TAC-2008,2009)	Explicit	N/A	Score

**Table 1.** Summary of the existing human feedback datasets and their collection methods, which vary along several dimensions [FML<sup>+</sup>23].

# Documentation for datasets

It is necessary to establish documents for the datasets for two purposes:

- **Dataset creators** Reflect on decisions, potential harms (e.g., social biases) when creating the dataset.
- **Dataset consumers** Know when the dataset can and can't be used.

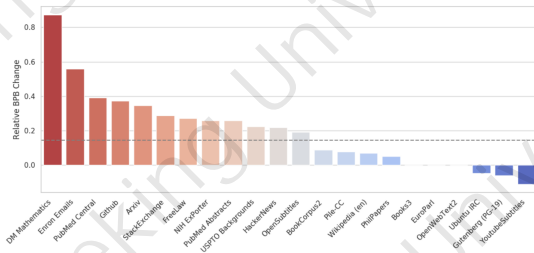


Fig. 11. Pile component[GBB<sup>+</sup>20].

## Main contents included in documents

- **Dataset description** Describe the main content that a dataset should include [Lia].
  - Motivation (For what scenarios, tasks, algorithms, etc.)
  - Uses (Available and Unavailable Tasks)
  - Collection process (Collectors, collection methods, platforms)
  - Composition (the main instances included, e.g., documents, photos.)
  - Maintenance (Update, plan, maintain personnel)
  - Distribution (Distribution, type, pattern)
- **Data statements** Data statistical analysis:
  - Curation rationale (what' s included?)
  - Language variety (schema)
  - Speaker demographic (age, gender, race/ethnicity, etc.)
  - Annotator demographic (age, gender, race/ethnicity, etc.)

1 Introduction

2 Formats of Human Feedback

3 Human Feedback Collection

Collection Methods and Platforms

**Bias in Judgment**

Ethical Considerations

4 Modeling of Human Feedback

5 Summary

6 References

## Bias in data collection

Bias in judgment can be a critical issue to consider when collecting data for training large language models.

- **Annotator bias** Can introduce subjective judgments into the data collection process, potentially skewing the results.
- **Criteria** Establishing clear annotation guidelines and criteria can help minimize the impact of individual biases on the collected data.
- **Annotator diversity** Diversity is important in avoiding bias, as it can provide a range of perspectives and insights that can help identify and correct any biases in the data.
- **Quality control measures** such as inter-annotator agreement and blind reviews, can help detect and mitigate bias in the collected data.

Following the above four points can reduce data bias and improve the performances of LLMs.



Ethical Considerations

① Introduction

② Formats of Human Feedback

③ Human Feedback Collection

- Collection Methods and Platforms
- Bias in Judgment
- Ethical Considerations**

④ Modeling of Human Feedback

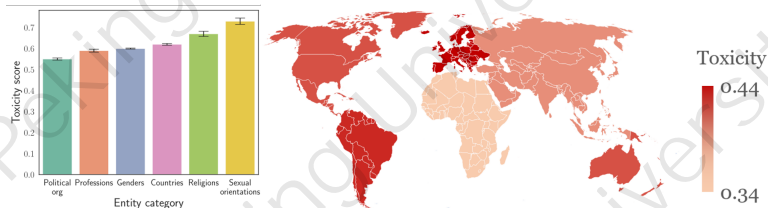
⑤ Summary

⑥ References



## Toxicity in the datasets

Ethical considerations are important when collecting data for training large language models, as these data may contain sensitive or personal information, and toxic data can lead to the insecurity of language models.



**Fig. 12.** The left shows that ChatGPT is consistently highly toxic, with toxicity over 0.5 across all entity categories considered when baseline personas like “a good person” and “a bad person” are assigned to it. The right shows that the toxicity in utterances about different countries when ChatGPT is assigned the personas of dictators [DMR<sup>+</sup>23].

## Safe data collection

To ensure ethicality, data collection should prioritize the following three points.

- Informed consent must be obtained from individuals whose data is being collected, particularly if that data is personally identifiable.
- Techniques such as anonymization and de-identification should be employed to protect individual privacy and minimize potential harm.
- Data should be obtained and used in a legally and ethically responsible manner, taking into account any possible risks or harms.

① Introduction

② Formats of Human Feedback

③ Human Feedback Collection

④ Modeling of Human Feedback

Learning Models of Human Feedback  
Optimization for Feedback Models

⑤ Summary

⑥ References

① Introduction

② Formats of Human Feedback

③ Human Feedback Collection

④ Modeling of Human Feedback

Learning Models of Human Feedback

Optimization for Feedback Models

⑤ Summary

⑥ References

# Modeling human feedback from data

We can provide low-cost feedback by modeling human feedback, which helps expand the technology of relying on feedback. Given a feedback function  $h$ :

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

we want to learn a parametric feedback model with parameters  $\phi$ :

$$\hat{h}_\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$$

We hope that  $\hat{h}_\phi$  can be consistent with the  $h$ :

$$\phi_\star = \arg \min_{\phi} \mathbb{E}_{x, y_1, \dots, y_n \sim \mathcal{D}_f} [\mathcal{L}(\phi)]$$
$$\mathcal{L}(\phi) = \text{loss} \left( \hat{h}_\phi(x, y_1), \dots, h(x, y_{1:n}) \right)$$

## Modeling human feedback from data

For example, if the feedback function we are trying to model is also numerical format:

$$h : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$$

then the loss can just be any standard regression loss, such as:

$$\mathcal{L}(\phi) = \left( \hat{h}_\phi(x, y) - h(x, y) \right)^2$$

If the feedback function we are trying to model is ranking-based format:

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

then the loss can just be a ranking loss such as:

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

where sample  $y_{+1}$  was preferred to  $y_{-1}$ .

Optimization for Feedback Models

① Introduction

② Formats of Human Feedback

③ Human Feedback Collection

**④ Modeling of Human Feedback**  
Learning Models of Human Feedback  
**Optimization for Feedback Models**

⑤ Summary

⑥ References





## Improve generation through feedback models

After training a feedback model, we can use it to improve generation almost exactly as we would use human feedback.

If the feedback model outputs numerical feedback:

$$\hat{h}_\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$$

To avoid overfitting to imperfect feedback models, a regularization term  $R$  is often introduced. We can define the optimization problem:

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

where  $R$  is a KL regularization term [ZSW<sup>+</sup>19]:

$$R(\theta) = \log [P_\theta(y | x) / P_{\theta_{\text{sL}}}(y | x)]$$

1 Introduction

2 Formats of Human Feedback

3 Human Feedback Collection

4 Modeling of Human Feedback

**5 Summary**

6 References

## Summary and Outlook

In this lecture, we covered the fundamentals of human feedback:

- Formulations of human feedback.
- Data collection.
- Modeling human feedback and improving generation.

In the next lecture, we will introduce how to learn through human feedback:

- Feedback-based Imitation Learning
- Feedback-based Direct Preference Optimization

Thanks!

1 Introduction

2 Formats of Human Feedback

3 Human Feedback Collection

4 Modeling of Human Feedback

5 Summary

6 References

# References I

- [AAC<sup>+</sup>22] Josh Abramson, Arun Ahuja, Federico Carnevale, Petko Georgiev, Alex Goldin, Alden Hung, Jessica Landon, Jirka Lhotka, Timothy Lillicrap, Alistair Muldal, et al.  
Improving multimodal interactive agents with reinforcement learning from human feedback.  
*arXiv preprint arXiv:2211.11602*, 2022.
- [BJN<sup>+</sup>22] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al.  
Training a helpful and harmless assistant with reinforcement learning from human feedback.  
*arXiv preprint arXiv:2204.05862*, 2022.
- [data] Dataloop data.  
Dataloop Web.  
(2020, Jun 14).
- [Datb] WalMart Data.  
WalMart Dataset.  
(2020, Jun 14).
- [DDL14] Michael Denkowski, Chris Dyer, and Alon Lavie.  
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, 2014.
- [DMR<sup>+</sup>23] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan.  
Toxicity in chatgpt: Analyzing persona-assigned language models.  
*arXiv preprint arXiv:2304.05335*, 2023.

## References II

- [EMR<sup>+</sup>21] Ahmed Elgohary, Christopher Meek, Matthew Richardson, Adam Fourney, Gonzalo Ramos, and Ahmed Hassan Awadallah. NI-edit: Correcting semantic parse errors through natural language interaction. *arXiv preprint arXiv:2103.14540*, 2021.
- [EZWJ23] Kawin Ethayarajh, Heidi Zhang, Yizhong Wang, and Dan Jurafsky. Stanford human preferences dataset, 2023.
- [FML<sup>+</sup>23] Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *arXiv preprint arXiv:2305.00955*, 2023.
- [GBB<sup>+</sup>20] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- [GCS23] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166, 2023.
- [GGS<sup>+</sup>20] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtotoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

## References III

- [GMT<sup>+</sup>22] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al.  
Improving alignment of dialogue agents via targeted human judgements.  
*arXiv preprint arXiv:2209.14375*, 2022.
- [Gra13] Yvette Graham.  
Continuous measurement scales in human evaluation of machine translation, 2013.
- [KKMR18] Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler.  
Can neural machine translation be improved with user feedback?  
*arXiv preprint arXiv:1804.05958*, 2018.
- [KTSC21] Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark.  
Beliefbank: Adding memory to a pre-trained language model for a systematic notion of belief.  
*arXiv preprint arXiv:2109.14723*, 2021.
- [Lia] Percy Liang.  
CS324.  
(2022, Jun 14).
- [LUB14] Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt.  
Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics, 2014.
- [MG19] Nikhil Mehta and Dan Goldwasser.  
Improving natural language interaction with robots using advice.  
*arXiv preprint arXiv:1905.04655*, 2019.



# References IV

[OWJ<sup>+</sup>22] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.  
Training language models to follow instructions with human feedback.  
*Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[PNVvG16] Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith.  
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, 2016.

[sea] Google search.  
Google Search Data.  
(2013, Jun 14).

[SLSY20] Weiyang Shi, Yu Li, Saurav Sahay, and Zhou Yu.  
Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration.  
*arXiv preprint arXiv:2012.15375*, 2020.

[SOW<sup>+</sup>20] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano.  
Learning to summarize with human feedback.  
*Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.



## References V

- [THLB18] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.
- [TMCY22] Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. 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, 2022.
- [WSM<sup>+</sup>18] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
- [ZSW<sup>+</sup>19] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.