Languages are Rewards: Hindsight Finetuning using Human Feedback

Hao Liu   Carmelo Sferrazza   Pieter Abbeel
University of California, Berkeley
hao.liu@cs.berkeley.edu

Abstract

Learning from human preferences is important for language models to be helpful and useful for humans, and to align with human and social values. Existing works focus on supervised finetuning of pretrained models, based on curated model generations that are preferred by human labelers. Such works have achieved remarkable successes in understanding and following instructions (*e.g.*, InstructGPT, ChatGPT, etc). However, to date, a key limitation of supervised finetuning is that it cannot learn from negative ratings; models are only trained on positive-rated data, which makes it data inefficient. Because collecting human feedback data is both time consuming and expensive, it is vital for the model to learn from all feedback, akin to the remarkable ability of humans to learn from diverse feedback. In this work, we propose a novel technique called Hindsight Finetuning for making language models learn from diverse human feedback. In fact, our idea is motivated by how humans learn from hindsight experience. We condition the model on a sequence of model generations paired with hindsight feedback, and finetune the model to predict the most preferred output. By doing so, models can learn to identify and correct negative attributes or errors. Applying the method to GPT-J, we observe that it significantly improves results on summarization and dialogue tasks using the same amount of human feedback.

1. Introduction

Large neural network based models have drawn continuously increasing attention in recent years, with applications in everything from natural language understanding (Brown et al., 2020; Devlin et al., 2018) to protein structure prediction (Jumper et al., 2021). However, in order to ensure that these technologies have a positive impact on society, it is of paramount importance for them to be aligned with human values. One of the most critical elements in achieving this is the use of human feedback.

Human feedback allows us to evaluate the performance of such models in a way that is both objective and subjective. It can help to identify issues with accuracy, fairness, and bias, and provide insights into how the model can be improved, in order to ensure that a model's outputs align with societal norms and expectations.

Driven by the importance of incorporating human feedback into language models, researchers have been developing and testing various methods for human-in-the-loop systems. These methods aim to make the process of incorporating human feedback more efficient, resulting in models that are able to achieve improved performance and accuracy, while also providing more fairness and ethical outputs (Hancock et al., 2019; Perez et al., 2019; Yi et al., 2019; Ouyang et al., 2022). For example, InstructGPT (Ouyang et al., 2022) and ChatGPT (OpenAI, 2022).

One key component of these successes is supervised finetuning (SFT) on human annotated data and positive-rated model generation. This method works by pre-training a model on a large dataset and then finetuning it on a smaller dataset with human annotations and positive-rated data. This is effective at improving the model's performance on specific tasks, but it relies on the availability of large amounts of labeled data, which can be costly and time-consuming to acquire.

Despite the successes of supervised finetuning on human feedback, a key limitation is that it cannot use negative-rated model generation. The reason is that directly finetuning model on negative-rated data will make the model less preferred by human, and using minimizing likelihood of negative-rated data conflicts with likelihood maximization. Only utilizing positive-rated data to finetune the model means that the model may not be able to properly identify and correct negative attributes or errors. Additionally, the model may not be able to generalize well to new, unseen data without negative feedback.

To address this issue, we aim to utilize both positive-rated and negative-rated data. The research hypothesis is that by doing so, the model can learn to identify and correct
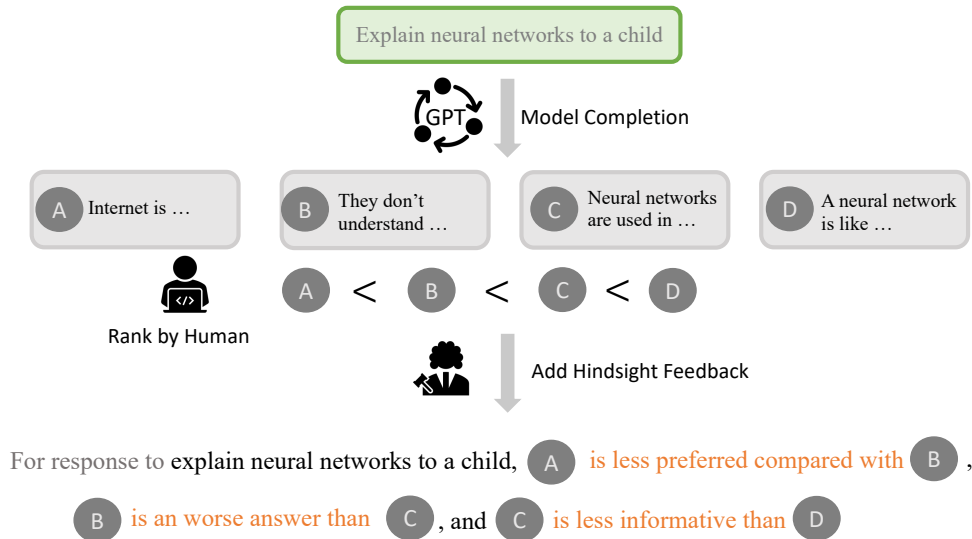
Figure 1. Illustrative example of constructing chain of hindsight sequence from human ranked model generations. The chain of hindsight sequence is then used to finetune model as shown in Figure 2.

negative attributes or errors. Our key idea is motivated by Hindsight Experience Replay (HER) (Andrychowicz et al., 2017). Imagine that you are learning how to figure skating and are trying to pass a test which requires getting at least 90 out of 100 score. You got some elements like lifts and spins correct but did not successful complete all required elements. You have tried multiple times without successes. Conventional supervised finetuning would be that since no attempt leads to a successful figure skating, thus only the ones among them with higher scores or none of them are going to be used, and little to nothing would be learned.

It is however possible to draw another conclusion, namely that concatenating an attempt with score 30, a feedback (*e.g.*, 'last attempt gets score 30, the following attempt is more preferred by coach and gets score 50'), and an attempt with score 50 together, and treat it as a success example, that would be successful if the model is improving with feedback to be successful. By finetuning the model to predict a higher rated attempt conditioning on one or more lower rated attempts and feedback, this effectively leverages data with all ratings. Therefore, the method is named **Chain of Hindsight Finetuning (CoHF)**.

We evaluate our method on both general NLP benchmarks and human evaluations. *CoHF* performs significantly better than supervised finetuning. To summarize, our contributions are:

- We propose a novel method, Chain of Hindsight Finetuning, for learning from human feedback that effectively uses both positive-rate and negative-rated data.

- We conduct extensive experiments and demonstrate that our method significantly outperforms supervised finetuning in both automatic evaluation and human evaluation.

2. Related Work

Learning from Human Feedback. Prior work have explored using human feedback to improve various NLP tasks, such as summarization (Böhm et al., 2019; Ziegler et al., 2019; Stiennon et al., 2020), dialogue (Yi et al., 2019; Hancock et al., 2019; Bai et al., 2022a;b; Askell et al., 2021), translation (Kreutzer et al., 2018; Bahdanau et al., 2016), semantic parsing (Lawrence & Riezler, 2018), story generation (Zhou & Xu, 2020), review generation (Cho et al., 2018), evidence extraction (Perez et al., 2019), and more recently instruction following (Ouyang et al., 2022). The main techniques behind them can be categorized as supervised finetuning or training (SFT) on filtered human annotations and learning a reward function from human feedback for reinforcement learning (often dubbed as RLHF (Christiano et al., 2017)). Ouyang et al. (2022) demonstrates the effectiveness of SFT and RLHF by first improving models with SFT followed by RLHF. Our work belongs to the category of SFT, and differs from SFT in that our method can learn from negative examples. Because our method is an improved version of SFT, it is complementary to RLHF and both can be directly combined together for further improvement. Using instructions to provide models with human preference and desired behaviors is demonstrated in Bai et al. (2022b), where models are prompted with a set of statements or prin-

ciples and trained with RLHF. In our work, we provide models with feedback on model generations and directly instruct models to generate desired outputs.

Learning from Hindsight. In this paper we explore learning from chain of hindsight with human feedback, an approach that enables a model to learn from errors and revise generations. The key idea of learning from hindsight experience was explored in goal conditioned RL (Kaelbling, 1993; Andrychowicz et al., 2017; Schaul et al., 2015). Andrychowicz et al. (2017) proposes hindsight experience replay (HER) to relabel rewards and transitions retroactively to learn from sparse feedback. In relation to HER (Andrychowicz et al., 2017), our work considers the batch setting rather than online setting. We propose algorithm improvements to construct hindsight experience directly from human rated model generations. HER uses RL to learn from hindsight experience while *CoHF* constructs chain of hindsight experience using human feedback and directly finetunes on it.

Instruction Finetuning. Finetuning on chain of hindsight using human feedback is akin to instruction finetuning. Driven by the impressive in-context learning ability of large language models, finetuning pretrained models on instructions has been shown to improve language models in many benchmarks (see e.g. Wang et al., 2022; Mishra et al., 2021; Ye et al., 2021; Chung et al., 2022; Wei et al., 2021; Sanh et al., 2021, inter alia). Mostly the instructions are reformatted examples from NLP benchmarks (e.g. Wei et al., 2021; Chung et al., 2022; Mishra et al., 2021). These works train models to predict human desired model outputs by following instructions. In relation to them, our work explores chain of hindsight finetuning which consists of both human preferred and non-preferred model completions. Chains of thought prompt (Wei et al., 2022) are widely considered as instructions in prior works (Chung et al., 2022; Wei et al., 2021), specifically in the form of step by step explanations written by humans. In relation to these, our chain of hindsight consists of human written hindsight feedback and ranked model outputs. Our work suggests a promising direction of using hindsight feedback to construct instructions from model outputs, and can be combined with prior instruction finetuning works for further improvement.

3. Chain of Hindsight Finetuning

CoHF uses a standard causal, decoder-only Transformer model architecture (Vaswani et al., 2017), *i.e.*, each timestep can only attend to itself and past timesteps. We illustrate *CoHF* in Figure 2.

Human feedback sequence. Our goal is to train the Transformer on human rated data to learn to achieve higher human preference scores. Human feedback data is in the form of $(x, \{y_i, r_i, z_i\}_{i=1}^n)$ where x is the the prompt (*e.g.*, 'sum-

marizing the following NYT article'), each y_i is a model completion, r_i is human rating on y_i, and z_i is human-provided *hindsight feedback* in natural language, *e.g.*, 'y_i is less preferred compared with y_j' or 'the following output is a more preferred', etc. The form of human feedback varies between datasets, from binary feedback where $n = 2$ to more fine grained scalar feedback where $n > 2$ (see e.g., Ouyang et al., 2022; Stiennon et al., 2020; Bai et al., 2022b). The feedback used in our experiments is provided in Appendix A.

Rather than conventional SFT methods that finetune models on data with high human preference score, we want to leverage both positive-rated data and negative-rated data. Our idea is to feed the model with a sequence of model generations sorted in ascending order according to their human preference scores, and also feed the model with instructions and feedback on which data is more preferred. The model is trained to predict most preferred data conditioning on less preferred data as well as human feedback and explanations. By doing so, we expect that the model can pick up how to generate human preferred data by also considering less preferred data and human feedback.

Given a set of human feedback data,

$$D_h = \{(x, y_i, r_i, z_i)\}_{i=1}^n, \tag{1}$$

without loss of generality, assume y_n is the most human preferred model completion:

$$r_n \geq r_{n-1} \geq \cdots \geq r_0. \tag{2}$$

Our sequence representation is given by:

$$\tau_h = (x, z_i, y_i, z_j, y_j, \cdots, z_n, y_n) \quad 1 \leq i < j < n, \tag{3}$$

Model generations y_i (and y_j, etc) are randomly sampled from D_h. We note that when τ_h only consists of x and y_n (and z_n is empty), that is, $\tau_h = (x, y_n)$, *CoHF* reduces to conventional supervised finetuning. We finetune the model on sequence τ to *only* predict y_n. The rationale is that by only finetuning on more preferred data, model learns to improve upon less preferred behaviors.

Therefore, the objective is given as follows:

$$\max_{\phi} \sum_{\tau_h \sim D_h} \log p_{\phi}(y_n | x, z_i, y_i, z_j, y_j, \cdots, z_n, y_n). \tag{4}$$

At test time, following existing methods, we prompt the model with dataset questions, and evaluate model generations.

In the following, we remark two techniques that help finetuning using human feedback.
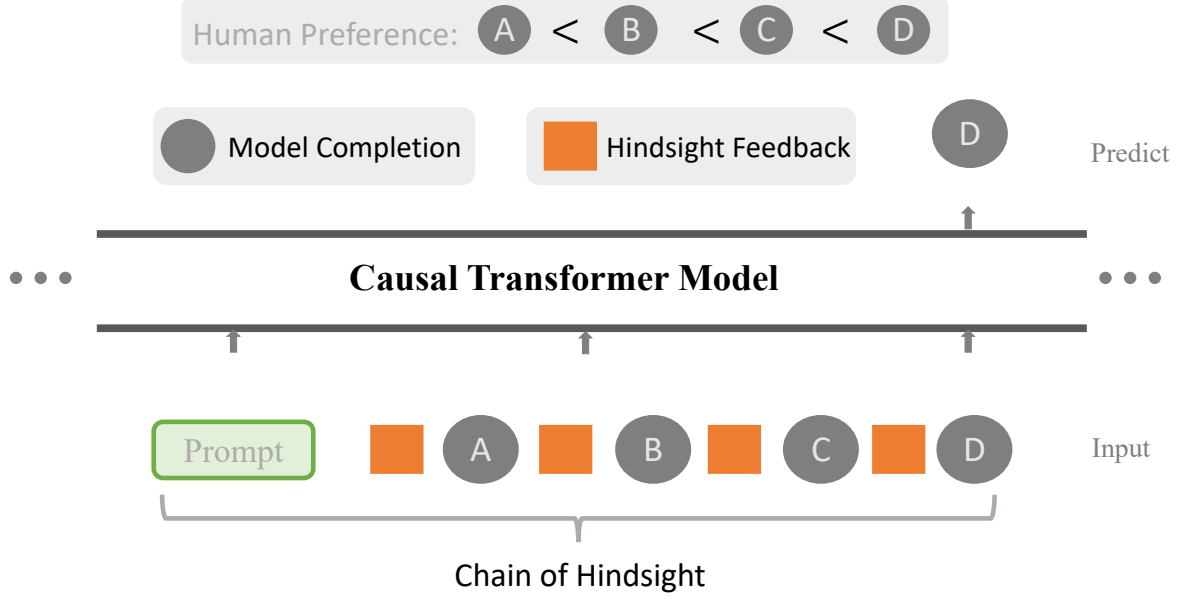
Figure 2. Model takes task question prompt and chain of hindsight as input and predicts model output that is most preferred by human. In this diagram, model outputs A, B, C, and D are ranked by human labeler and then a chain of hindsight is constructed by incorporating hindsight feedback. Hindsight feedback comes with various forms (*e.g.*, A is less preferred compared with B) to inform model relative rankings between model outputs. An illustrative example is shown in Figure 1.

Prevents shortcut. Human preferred data often have small differences with non-preferred data, *e.g.*, a negative-rated dialogue may be fairly accurate and coherent in most aspects but missed one or two key words. Because of this, 'hindsight' sequences share many common words, and directly finetuning the model on them makes it easy for this to just shortcut copy tokens. To mitigate this issue, we randomly mask 15% of past tokens similar to Liu et al. (2022). In our experiments, we find that this regularization improves results.

Specifically, denoting the chain of hindsight sequences as $\mathbf{s} = [s_1, \cdots, s_n]$, the standard causal language modeling objective is defined to maximize the log likelihood of s autoregressively:

$$\log p(\mathbf{s}) = \log \prod_{i=1}^{n} p(s_i|s_1, s_2, \ldots, s_{i-1})$$
$$= \log \prod_{i=1}^{n} p(s_i|\mathbf{s}_{<i}) := \log \prod_{i=1}^{n} p(s_i|[s_j]_{j=0}^{i-1}). \quad (5)$$

Following (Liu et al., 2022), we use $\eta = 0.15$ throughout the experiments unless otherwise mentioned. The model is asked to predict each token $s_i \in \mathbf{s}$, and can only attend to tokens in $\mathbf{s}_{<i}$ that are not sampled. Concretely, the training objective is given by:

$$\log p(\mathbf{s}) = \log \prod_{i=1}^{n} p(s_i|[I[m_j > \eta] \cdot s_j]_{j=0}^{i-1}), \quad (6)$$

where $m_j \sim \mathcal{U}(0, 1)$, and I is the indicator function.

Prevent overfitting. The diversity of human annotations and model-generated data is limited and this can cause overfitting. To mitigate this issue, similar to Ouyang et al. (2022), we also minimize the negative log likelihood of the pretraining dataset. In our experiments, we find that this regularization enables generating more natural sentences. The objective is therefore a combination of the finetuning objective, defined through Equation 4 and 6, and the λ weighted pretraining objective.

$$\max_{\phi} \sum_{\tau_h} \log p_\phi(y_n|x, z_i, y_i, z_j, y_j, \cdots, z_n, y_n)$$
$$+ \lambda \sum_{\tau_p} \log p_\phi(\tau_p) \quad (7)$$
$$\text{where} \quad \tau_p \in D_p, \tau_h = (x, z_i, y_i, z_j, y_j, \cdots, z_n, y_n) \sim D_h.$$

Once again, D_h is human rated model generation dataset and D_p is pretraining dataset. λ is a hyperparameter to balance between pretraining objective and finetuning objective.

Training. We are given a dataset of model outputs and their human preference feedback. We sample minibatches

Algorithm 1 Hindsight Finetuning from Human Feedback

   Required: Human Feedback Dataset, GPT Model
   Required: Max Iterations m, Max Number of Chain of Hindsight n
   Initialize
   for $i = 1$ **to** $m - 1$ **do**
      Randomly sample an example from dataset
      Randomly sample j from 1 to n
      Randomly sample j model outputs from the example
      Sort j outputs ascending according to their human ratings
      Generate hindsight feedback for the j outputs
      Concatenate example prompt, model outputs and hindsight feedback as a sequence as shown in Figure 2
      Finetune model on hindsight experience to predict the best model output.
   end for

of model outputs from the dataset. Hindsight feedback in natural language is generated by sampling the feedback format and accounting for the human ratings. Hindsight feedback and model outputs are packed together as chain of hindsight. The prediction is autoregressively predicting the most preferred model output sequence, and the cross-entropy loss is averaged for each timestep in the last model output sequence. We did not find predicting hindsight feedback tokens or other model output sequences to improve performance, although it is easily permissible within our framework and would be an interesting study for future work. The algorithm is shown in Algorithm 1.

4. Evaluation Setup

Evaluation Tasks and Metrics. For automatic evaluation, we follow prior works Brown et al. (2020); Wang & Komatsuzaki (2021) and consider a diverse suite of standard NLP tasks, including SuperGLUE (Sarlin et al., 2020), ANLI (Nie et al., 2019), LAMBADA (Paperno et al., 2016), StoryCloze (Mostafazadeh et al., 2016), PIQA (Bisk et al., 2019), and more. The full task list is shown in Table 1. We use Language Model Evaluation Harness[1] for evaluation.

Following prior works on learning from human feedback (Stiennon et al., 2020; Nakano et al., 2021; Bai et al., 2022a), we also consider two tasks that are best evaluated with human preference. The first one is summarization on TL;DRs dataset (Völske et al., 2017). The original TL;DR dataset contains about 3 million posts from reddit.com across a variety of topics (subreddits), as well summaries of the posts written by the original poster (TL;DRs). We use the filtered version provided by Stiennon et al. (2020), which

contains 123,169 posts[2]. We evaluate the performance on the validation set. For evaluation metrics, labelers rated summaries for coverage (how much important information from the original post is covered), accuracy (to what degree the statements in the summary are part of the post), coherence (how easy the summary is to read on its own), and overall quality. More details about evaluation dimensions and instructions for human labelers are available in Appendix B.

The second task is dialogue. Our evaluation dataset is based on the dataset from Bai et al. (2022a)[3]. Each example in the dataset consists of a pair of conversations between large language model and human, where one of them is preferred by the human and the other is not. In our evaluation, since deploying our finetuned model to chat with humans for data collection is extremely expensive and time consuming, we use positive examples to construct 'pseudo' dialogues between language model and human. Concretely, we replace each model response from a previous dialogue with our model's output. The output is generated by conditioning the model on human response and past model outputs. As in prior work, the metrics we consider for evaluating dialogue are helpfulness and harmlessness. To be helpful, the model should follow instructions, but also infer intention from a few-shot prompt or another interpretable pattern. Since a given prompt's intention can be unclear or ambiguous, we rely on judgment from our labelers, and our main metric is labelers preference ratings. However, since the labelers are not the users who generated the prompts, there could be a divergence between what a user actually intended and what the labeler thought was intended from only reading the prompt.

Training Datasets. We use a combination of three datasets for finetuning. The three datasets are:

WebGPT comparisons. This dataset is from Nakano et al. (2021).[4] There are 19,578 comparisons in total. Each example in the dataset contains a pair of model answers for a question, and the associated metadata. Each answer has a preference score from humans that can be used to determine which of the two answers is better.

Human Preference. This dataset is from Ganguli et al. (2022); Bai et al. (2022a) and contains human rated dialogues.[3] An example consists of a pair of conversations between human and chatbots. One of the two conversations is more preferred by human.

Summarize from feedback. This dataset is provided by Stiennon et al. (2020) and contains human feedback on model generated summarizations[2]. There are two parts of this dataset: comparisons and axis. In the comparisons part, human annotators were asked to choose the best out of two summaries. In the axis part, human annotators gave scores on a Likert scale for the quality of a summary. The comparisons part only has a train and validation split, and the axis part only has a test and validation split. The summaries used for training the reward model in the paper come from the TL;DR dataset. Additional validation and test data come from the TL;DR dataset, CNN articles, and Daily Mail articles.

Model Architectures. We use the same model and architecture as GPT-J (Wang & Komatsuzaki, 2021), including the modified activation (Shazeer, 2020), multi-query attention (Shazeer, 2019), parallel layers (Wang & Komatsuzaki, 2021) and RoPE embeddings (Su et al., 2021) described therein.

Baselines. Our baselines are pretrained model and SFT. In our experiments, the pretrained model is GPT-J 6B (Wang & Komatsuzaki, 2021), which is also the base model of SFT and *CoHF*. The SFT method finetunes the model on positive-rated data, *e.g.*, that is, only on human preferred summarization or dialogue. Prior works have shown its effectiveness in learning from human feedback (see e.g., Ouyang et al., 2022; Stiennon et al., 2020; Bai et al., 2022a). Concretely, the objective of supervised finetuning is given by the following equation:

$$\max_{\phi} \sum_{\tau_h} \log p_\phi(y_n|x, y_n) + \lambda \sum_{\tau_p} \log p_\phi(\tau_p), \quad (8)$$

where $\tau_p \in \mathrm{D}_p, \tau_h = (x, \cdots, y_n) \sim \mathrm{D}_h$. To ensure fair comparison, we use the same training techniques and hyperparameters for both SFT and *CoHF*.

5. Main Results

5.1. Automatic Evaluation

First automatic evaluation is on diverse suite of NLP tasks used in prior work (Brown et al., 2020; Wang & Komatsuzaki, 2021). We notice that the average performance of supervised finetuning is decreased after finetuning (see Table 1), which would be related to the *alignment tax* issue in language models (Ouyang et al., 2022), suggesting the necessity of human evaluation (Lee et al., 2022). We observe *CoHF* improves moderately over both pretrained model and supervised finetuned model.

In Figure 4, we show the ROUGE scores of our models on the TL;DR dataset. *CoHF* significantly outperforms both pretrained model and supervised finetuning. This suggests
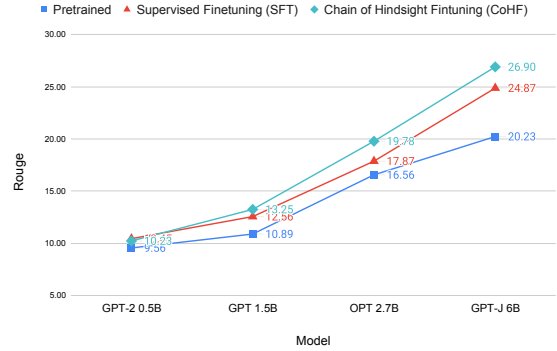

Figure 3. **Scaling trend**. Comparison of supervised finetuning and chain of hindsight finetuning on summarization task with different model sizes. *CoHF* scales better than SFT.
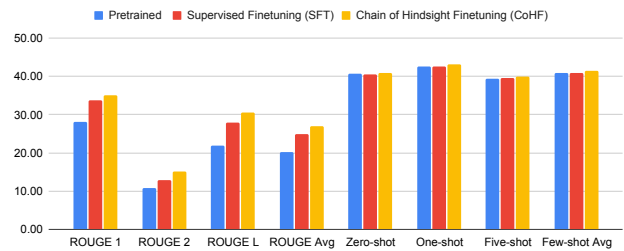

Figure 4. **Automatic evaluation**. Chain of hindsight finetuning outperforms supervised finetuning on automatic evaluation. The metrics are ROUGE score on TL;DR summary task and average (0, 1, 5)-shot performance on 22 few-shot learning tasks.

CoHF is more effective at learning from human feedback.

Figure 3 shows the results under different model sizes, at smaller model sizes, *CoHF* performs slightly worse than SFT, but as model size increases, *CoHF* consistently outperform SFT and shows a promising scaling trend.

5.2. Human Evaluation

Following prior work Stiennon et al. (2020); Bai et al. (2022a), we proceed with human evaluations on summarization and dialogue tasks. 15 human labelers who are proficient in English are hired from a third-party platform to provide ratings.

For the summarization task, human labelers are presented with two summaries, one is generated by SFT (or pretrained GPT-J) and one is generated by *CoHF*. Labelers are instructed to select the best (or neutral) among the two according to the three metrics mentioned above.

The results on the summarization task are presented in Table 3 (Top) and Table 2 (Top). We observe that *CoHF* is significantly more preferred by human labelers than pretrained model and supervised finetuned model.

Table 1. Results across few-shot NLP benchmarks. We use the same setup as in Brown et al. (2020); Wang & Komatsuzaki (2021), including the splits for each task. GPT-J numbers are from its original paper. Other models' numbers are from us. Results are averaged over 5 random seeds. SFT: Supervised Finetuning. *CoHF*: Chain of Hindsight Finetuning.

	Zero-shot			One-shot			Few-shot		
Task	GPT-J	SFT	*CoHF*	GPT-J	SFT	*CoHF*	GPT-J	SFT	*CoHF*
ANLI R1	34.00	33.50	33.80	33.50	33.50	33.60	32.70	32.60	32.70
ANLI R2	32.00	32.00	32.10	34.40	34.10	34.20	33.90	34.20	34.10
ANLI R3	34.00	34.30	36.80	34.80	34.60	36.90	35.40	35.60	36.80
ARC-C	27.00	26.80	27.60	32.20	32.50	33.80	33.10	33.50	34.20
ARC-E	54.30	54.20	54.40	62.80	62.50	62.50	66.50	66.50	66.50
BoolQ	58.50	61.50	61.30	57.20	57.10	58.10	42.50	42.30	42.90
CB	41.10	41.00	40.50	41.10	41.10	40.50	42.90	42.10	42.00
COPA	71.00	70.50	69.90	80.00	80.10	80.50	82.00	82.20	81.50
HeadQA	23.50	23.00	23.80	24.00	23.80	24.30	23.90	22.50	22.80
HellaSwag	42.60	42.30	42.00	46.20	46.10	46.10	46.10	46.00	46.70
MultiRC	3.00	3.10	4.10	6.50	6.70	7.40	6.60	6.90	7.50
ReCORD	85.80	85.60	85.60	86.20	86.00	86.40	58.60	58.80	58.60
RTE	51.20	50.50	50.00	55.60	55.50	55.90	52.00	52.00	52.00
WiC	45.00	45.00	45.00	44.50	44.20	44.10	50.00	50.50	50.00
WSC	36.50	36.90	42.80	37.50	38.10	43.70	35.80	37.60	41.30
LAMBADA (openai)	5.50	5.70	5.70	5.30	5.40	5.40	2.50	2.70	3.60
LAMBADA (standard)	2.10	0.90	0.90	3.00	2.20	1.90	3.20	3.30	3.30
LogiQA	21.50	20.00	20.00	20.70	20.90	20.90	19.00	20.60	20.10
WinoGrande	49.70	50.40	51.20	50.70	51.80	53.50	50.70	51.10	52.80
SciQ	86.40	86.00	86.00	89.10	89.10	89.10	54.00	55.00	55.00
OpenBookQA	16.00	16.20	15.40	16.80	16.70	16.70	20.80	20.90	21.10
PIQA	72.40	72.40	72.00	73.60	73.70	73.50	74.20	74.00	74.00
Average	40.60	40.54	**40.95**	42.53	42.53	**43.14**	39.38	39.59	**39.98**

For the dialogue task, we replace the model response parts from each dialogue in the data with new generations from our models. For instance, if the original dialogue is two-turns dialogue as [human-1][chabot-1][human-2][chatbot-2], the new dialogue would be [human-1][newbot-1][human-2][newbot-2], where [newbot-1] is generated by feeding the model with [human-1] and [newbot-2] is generated by feeding the model with [human-1][newbot-1][human-2].

The purpose of doing so instead of having humans directly chat with the finetuned model is to reuse human generated data, since collecting interactive data is very costly and is prone to low data quality issues.

The results are presented in Table 3 (bottom) and Table 2 (bottom). While more than 50% of the labelers is neutral between SFT and *CoHF*, *CoHF* is still more favorable to human labelers compared to SFT.

5.3. Model Variations

To evaluate the importance of the different components of *CoHF*, we varied our default configuration in different ways, measuring the change in performance on multiple metrics. We present these results in Table 4, where NLP Avg denotes the average score on the same suite of tasks as in Table 1. Rouge Avg denotes ROUGE scores on the

Table 2. **Human evaluation between pretrained and *CoHF*.** The comparisons between pretrained model and chain of hindsight finetuning (*CoHF*). **Top**: TL;DR summarization task based on Stiennon et al. (2020). **Bottom**: dialogue task based on Bai et al. (2022a). The metrics used in human evaluation follow definitions from prior works. Δ denotes the releative improvement of *CoHF* over SFT.

Summary (%)	Pretrained	Neutral	*CoHF*	Δ
Accuracy	24.5	23.6	51.9	27.4
Coherence	18.9	18.5	62.6	43.7
Coverage	31.8	20.5	47.7	15.9
Average	25.1	20.9	54.1	29.0
Dialogue (%)				
Helpful	15.8	34.8	49.4	33.6
Harmless	14.5	35.9	49.6	35.1
Average	15.2	35.3	49.5	34.4

Table 3. **Human evaluation between SFT and *CoHF*.** The comparisons between supervised finetuning (SFT) and chain of hindsight finetuning (*CoHF*). Evaluation setting is the same as Table 2.

Summary (%)	SFT	Neutral	*CoHF*	Δ
Accuracy	29.5	32.6	37.9	8.4
Coherence	21.7	25.6	52.7	31.0
Coverage	30.5	25.4	44.1	13.6
Average	27.2	27.9	44.9	17.7
Dialogue (%)				
Helpful	19.6	55.3	25.1	5.5
Harmless	12.5	57.4	30.1	17.6
Average	16.1	56.3	27.6	11.6

filtered TL;DR dataset from Stiennon et al. (2020). HH Avg denotes model performance of model prompted to classify human preference on the *Human Preference* dataset (Bai et al., 2022a) validation split.

In Table 4 rows (A), we vary the mask ratio. Performance decreases when using a large mask ratio, and decreases significantly when mask ratio equals 0, this suggests using causal masking is crucial for preventing model from simply copying similar tokens.

In Table 4 rows (B), we vary the number of hindsight feedback used in constructing chain of hindsight. We observe that using a more diverse set of hindsight feedback is helpful for better results.

In Table 4 rows (C), instead of ascending sorting, we finetune on descending sorted sequences (with reversed hindsight feedback too). We observe that the scores of Rouge and HH are about the same as default setting, suggesting that the model can learn from reversed chain of hindsight. We leave exploring learning from both reversed and non-

Table 4. **Variations on *CoHF***. Unlisted values are identical to those of the default model. NLP Avg: average performance across the same set of diverse tasks from Table 1. Rouge Avg: average rouge scores on summarization dataset. HH Avg: average classification accuracy on human feedback dataset.

Variants	HF diversity	FCM	Reverse HF	Variable HF	Mix Pretrain	Adv HF	NLP Avg	Rouge Avg	HH Avg(%)
Default	20	0.15	false	true	true	false	41.36	26.90	78.8
(A)		0					40.25	25.35	75.2
		0.3					41.05	25.40	70.5
(B)	1						41.16	24.78	60.6
	5						41.19	25.79	64.5
	15						41.36	26.90	73.8
(C)			true				40.88	26.51	77.6
(D)				false			41.88	23.35	74.7
(E)					false		38.75	20.89	60.85
(F)						true	40.87	11.27	18.87
(H)	SFT with unlikelihood						38.35	19.97	41.5
	SFT on all data						40.44	20.35	49.8
	Supervised Finetuning (SFT)						40.89	24.87	62.3
(I)	Pretrained Model						40.84	20.23	43.8

[handwritten: "Mask Ratio" pointing to row (A)]

[handwritten: "How many in hindsight" pointing to row (B)]

reversed chain of hindsight as future work.

In Table 4 rows (D), instead of using variable length chain of hindsight, we use maximum length. We observe a decrease in Rouge and HH scores. The reason is probably that variable length chain of hindsight reduces the gap between training/finetuning and inference, since currently at inference time the model only does one-turn generation.

[handwritten: "Clarifies how this done"]

In Table 4 rows (E), we set $\lambda = 0$ which disables pretraining dataset regularization, we observe strong overfitting to finetuning dataset with NLP Avg score decreasing significantly. We further observe HH score decreases a lot, suggesting that the generalization is worse without pretraining datset regularization.

In Table 4 rows (F), we experiment with adversarial hindsight feedback, where a more preferred model generation is described as less preferred in chain of hindsight. We observe that both Rouge and HH scores show significant decrease. This suggests that the model can follow adversarial instructions encoded in chain of hindsight.

In Table 4 rows (H), we experiment with two variants of SFT.

SFT on all data denotes applying SFT on not just human preferred examples but also human rejected examples.

SFT with unlikelihood denotes adding an unlikelihood of human rejected examples to standard SFT.

$$\max_{\phi} \sum_{\tau_h} \log p_{\phi}(y_n|x, y_n) + \lambda \sum_{\tau_p} \log p_{\phi}(\tau_p)$$
$$- \gamma \sum_{\tau_h, i \neq n} \log p_{\phi}(y_i|x, y_i),$$

[handwritten: "→ Interesting comparison"]

where $\tau_p \in D_p, \tau_h = (x, \cdots, y_n) \sim D_h$. Unlikelihood has been shown to be helpful at steering models away of unwanted behaviors in controllable generation (Welleck et al., 2019; Li et al., 2019).

We observe that *SFT on all data* has minimal negative impact on NLP avg score but *SFT with unlikelihood* decreases performance. This suggests that unlikelihood training on human feedback data may hurt the pretrained model. On the other two scores, both variants are outperformed by standard SFT quite a lot, indicating that training on positive data seems more effective than combining negative data for SFT.

6. Conclusion

We propose Chain of Hindsight Finetuning (*CoHF*), a simple technique for finetuning language models with human feedback. *CoHF* can effectively leverage both negative-rated and positive-rated examples.

For summarization and dialogue tasks, *CoHF* significantly outperforms supervised finetuning. For automatic evaluation of a diverse suite of tasks, *CoHF* achieves better results than supervised finetuning.

We are excited about the future of *CoHF* and plan to apply it to other forms of human feedback and automatic feedback.

References

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., and Zaremba, W. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Aroca-Ouellette, S., Paik, C., Roncone, A., and Kann,

K. PROST: Physical reasoning about objects through space and time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4597–4608, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.404. URL https://aclanthology.org/2021.findings-acl.404.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., and Bengio, Y. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*, 2016.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., and Choi, Y. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv: Arxiv-1911.11641*, 2019.

Bisk, Y., Zellers, R., Le bras, R., Gao, J., and Choi, Y. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 04 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Böhm, F., Gao, Y., Meyer, C. M., Shapira, O., Dagan, I., and Gurevych, I. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214*, 2019.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cho, W. S., Zhang, P., Zhang, Y., Li, X., Galley, M., Brockett, C., Wang, M., and Gao, J. Towards coherent and cohesive long-form text generation. *arXiv preprint arXiv:1811.00511*, 2018.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pp. 4299–4307, 2017.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try ARC, the AI2 Reasoning Challenge. *Computing Research Repository*, arXiv:1803.05457, 2018. version 1.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The Pile: An 800GB dataset of diverse text for language modeling. *Computing Research Repository*, arXiv:2101.00027, 2020. version 1.

Hancock, B., Bordes, A., Mazare, P.-E., and Weston, J. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415*, 2019.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kaelbling, L. P. Learning to achieve goals. In *IJCAI*, volume 2, pp. 1094–8. Citeseer, 1993.

Kreutzer, J., Khadivi, S., Matusov, E., and Riezler, S. Can neural machine translation be improved with user feedback? *arXiv preprint arXiv:1804.05958*, 2018.

Lawrence, C. and Riezler, S. Improving a neural semantic parser by counterfactual learning from human bandit feedback. *arXiv preprint arXiv:1805.01252*, 2018.

Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., Gerard-Ursin, I., Li, X. L., Ladhak, F., Rong, F., et al. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*, 2022.

Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y.-L., Cho, K., and Weston, J. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. *arXiv preprint arXiv:1911.03860*, 2019.

Liu, H., Geng, X., Lee, L., Mordatch, I., Levine, S., Narang, S., and Abbeel, P. Fcm: Forgetful causal masking makes causal language models better zero-shot learners. *arXiv preprint arXiv:2210.13432*, 2022.

Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., and Zhang, Y. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In Bessiere, C. (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/501. URL https://www.ijcai.org/proceedings/2020/501.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260.

Mishra, S., Khashabi, D., Baral, C., and Hajishirzi, H. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv: Arxiv-1604.01696*, 2016.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv: Arxiv-1910.14599*, 2019.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

OpenAI. ChatGPT, OpenAI. https://openai.com/blog/chatgpt/, 2022. [Online; accessed 2-Feb-2023].

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv: Arxiv-1606.06031*, 2016.

Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., and Morante, R. QA4MRE 2011-2013: Overview of question answering for machine reading evaluation. In Forner, P., Müller, H., Paredes, R., Rosso, P., and Stein, B. (eds.), *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pp. 303–320, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40802-1. doi: 10.1007/978-3-642-40802-1_29.

Perez, E., Karamcheti, S., Fergus, R., Weston, J., Kiela, D., and Cho, K. Finding generalizable evidence by learning to convince q&a models. *arXiv preprint arXiv:1909.05863*, 2019.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 8732–8740. AAAI Press, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

Sarlin, P., DeTone, D., Malisiewicz, T., and Rabinovich, A. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 4937–4946. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00499. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Sarlin_SuperGlue_Learning_Feature_Matching_With_Graph_Neural_Networks_CVPR_2020_paper.html.

Schaul, T., Horgan, D., Gregor, K., and Silver, D. Universal value function approximators. In *International conference on machine learning*, pp. 1312–1320. PMLR, 2015.

Shazeer, N. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv: Arxiv-1911.02150*, 2019.

Shazeer, N. Glu variants improve transformer. *arXiv preprint arXiv: Arxiv-2002.05202*, 2020.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv: Arxiv-2104.09864*, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vilares, D. and Gómez-Rodríguez, C. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 960–966, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1092. URL https://aclanthology.org/P19-1092.

Völske, M., Potthast, M., Syed, S., and Stein, B. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, 2017.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 3266–3280. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Wang, B. and Komatsuzaki, A. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *URL https://arxiv. org/abs/2204.07705*, 2022.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Welbl, J., Liu, N. F., and Gardner, M. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413.

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., and Weston, J. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

Ye, Q., Lin, B. Y., and Ren, X. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. *arXiv preprint arXiv:2104.08835*, 2021.

Yi, S., Goel, R., Khatri, C., Cervone, A., Chung, T., Hedayatnia, B., Venkatesh, A., Gabriel, R., and Hakkani-Tur, D. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. *arXiv preprint arXiv:1904.13015*, 2019.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of*

the Association for Computational Linguistics, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Zhou, W. and Xu, K. Learning to compare for better training and evaluation of open domain natural language generation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9717–9724, 2020.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Hindsight Feedback

Table 5. Hindsight feedback used in the formatting human feedback data. Note that here we omit task prompt and other context information such as cited sources in WebGPT comparison data, because they are dataset specific and are available from dataset, one can just prefix them to input sequence.

Dataset	Hindsight feedback
Shared	compared to {neg} the following is preferred {pos}
Dialogue	the following conversation {neg} is worse than the following conversation {pos}
Shared	{neg} is worse than {pos}
Dialogue	{neg} is a less preferred conversation with human than {pos}
Dialogue	{neg} is a worse conversation with human than {pos}
Summary	{neg} is a worse summary compared with {pos}
Summary	{neg} can be improved compared with {pos}
Summary	The following is an example summary {neg}, it can be improved to a better summary as {pos}
Summary	The following is a summary {neg}, a more accurate and concise summary is {pos}
Dialogue	The following is an example dialogue around certain topic {neg}, it can be improved to a better dialogue as {pos}
Dialogue	{neg} is a dialogue, a more helpful and harmless dialogue is {pos}
Comparison	The following example is rejected by human {neg}, a better example is given by {pos}
Comparison	The following answer can be improved {neg}, for instance, a better answer is given by {pos}
Comparison	{neg} is an answer and a more accurate answer is {pos}
Shared	{neg} is less preferred compare with {pos}
Shared	{neg} is not as good as {pos}
Shared	a better alternative of {neg} is {pos}
Shared	Compared with {neg} a more accurate and readable choice is {pos}
Comparison	{pos(tie)} is an equally good answer as {neg(tie)}
Comparison	{neg(tie)} is an equally good answer as {pos(tie)}

B. Human Evaluation Instructions

For human evaluations, we instruct human labelers to select preferred output. We follow prior work Stiennon et al. (2020); Bai et al. (2022a) and resue their instructions and definitions of helpful, useful, etc. The instructions we use in summarization task are from Stiennon et al. (2020) which is publicly available at `https://docs.google.com/document/d/1MJCqDNjzD04UbcnVZ-LmeXJ04-TKEICDAepXyMCBUb8/edit#`. The instructions we use for dialogue task is from Bai et al. (2022a), we refer the readers to original paper for details.

C. Task List and Prompt Format

For an automatic evaluation of model's ability on diverse NLP tasks, we evaluate our model on a diverse collection of standard language model evaluation datasets: ANLI (Nie et al., 2020), ARC (Clark et al., 2018), HeadQA (English) (Vilares & Gómez-Rodríguez, 2019), HellaSwag (Zellers et al., 2019), LAMBDADA (Paperno et al., 2016), LogiQA (Liu et al., 2020), OpenBookQA (Mihaylov et al., 2018), PiQA (Bisk et al., 2020), PROST (Aroca-Ouellette et al., 2021), QA4MRE (Peñas et al., 2013) (2013), SciQ (Welbl et al., 2017), TriviaQA (Joshi et al., 2017), Winogrande (Sakaguchi et al., 2020), and the SuperGlue version of the Winograd Schemas Challenge (WSC) (Wang et al., 2019).

Two other tasks are summarization (Stiennon et al., 2020) and dialogue (Bai et al., 2022a). In our ablation study, we consider prompting model to evaluate model on whether it knows an example dialogue is preferred or not preferred by human using the dialogue dataset (Bai et al., 2022a).

The majority of prompt formats follow GPT-3 (Brown et al., 2020) which are made available by `https://github.com/EleutherAI/lm-evaluation-harness`. We follow the prompt formats used in Bai et al. (2022a) and Stiennon et al. (2020) for dialogue and summarization tasks.

```
For the following two dialogs. Please select your preferred dialog according to provided metrics.


****
[Dialog A]
****

****
[Dialog B]
****


Metric: Helpful. Which one is better?
  ○ A
  ● B
  ● Neutral
        submit
Metric: Harmless. Which one is better?
  ○ A
  ● B
  ● Neutral
        submit
```

Figure 5. Screenshots of our labeling interface for rating dialog. For each metric, labelers are asked to choose preferred dialog.

D. Hyperparameters

All models are trained with the Adam optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and an epsilon of $1.0e{-}8$. We finetune model for 10 epochs with residual dropout of 0.1. We use batch size 512 for human feedback data and batch size 512 for pretraining data. λ equals 1.5 which control the relative strength of gradients from SFT or *CoHF* on human feedback dataset and pretraining dataset. The Pile (Gao et al., 2020) is used as pretraining dataset for the pretraining regularization term.

E. Web UI

In Figure 6 and Figure 5, we show screenshots of our labeling interface, that all of our labelers use to rate data. Labelers can choose preferred model output or choose neutral in cases where two outputs seem to be of similar quality.

```
For the following article
=======
[Article]
=======
Please select your preferred summary according to provided metrics.


****
[Summary A]
****

****
[Summary B]
****


Metric: Accuracy. Which one is better?
  ● A
  ● B
  ● Neutral
          submit
Metric: Coherence. Which one is better?
  ● A
  ● B
  ● Neutral
          submit
Metric: Coverage. Which one is better?
  ● A
  ● B
  ● Neutral
          submit
```

Figure 6. Screenshots of our labeling interface for rating summary. For each metric, labelers are asked to choose preferred summary.