

What makes for a good summary? Roughly speaking, a good summary is a shorter piece of text that has the essence of the original – tries to accomplish the same purpose and conveys the same information as the original post. We would like you to consider these different dimensions of summaries:

### Accuracy

For this axis, answer the question “does the factual information in the summary accurately match the post?” A summary is accurate if it doesn’t say things that aren’t in the article, it doesn’t mix up people, and generally is not misleading.

### Coherence

For this axis, answer the question “how coherent is the summary on its own?” A summary is coherent if, when read by itself, it’s easy to understand and free of English errors. A summary is not coherent if it’s difficult to understand what the summary is trying to say. Generally, it’s more important that the summary is understandable than it being free of grammar errors.

### Coverage

For this axis, answer the question “how well does the summary cover the important information in the post?” A summary has good coverage if it mentions the main information from the post that’s important to understand the situation described in the post. A summary has poor coverage if someone reading only the summary would be missing several important pieces of information about the situation in the post. A summary with good coverage should also match the purpose of the original post (e.g. to ask for advice).

Table 4: Instructions we provided to human labelers on evaluating summarization quality.

## A HUMAN EVALUATION INSTRUCTIONS

For our human evaluations, we provide instructions and metrics definition to the human labelers, asking them to select the preferred output. In order to maintain consistency and build upon prior research (Stiennon et al., 2020; Bai et al., 2022a), we adopt their instructions and definitions of helpfulness, usefulness, and other relevant criteria.

Specifically, the instructions employed in our summarization benchmark are derived from Stiennon et al. (2020). Similarly, for the dialogue task, we derive the instructions based on Bai et al. (2022a). Table 4 provides more detail on the specific instructions given to labelers for comparing summaries, and Table 5 lists our instructions for evaluating dialogues.

## B NATURAL LANGUAGE FEEDBACK

During inference time, we only employ simple positive tokens, while during training, we explored the incorporation of natural language feedback that carries more semantic meaning. This natural feedback is tailored to the specific task and offers increased diversity, as illustrated in Table 6.

## C HYPERPARAMETERS

All models are trained with the Adam (Kingma and Ba, 2014) optimizer, with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.95$ , and an epsilon of  $1.0e-8$ . The batch size for human feedback data is set to 512, while for pretraining data it is set to 2048. The value of  $\lambda$  is 1.5, which determines the relative strength of gradients from the human feedback dataset and the pretraining dataset. The pretraining regularization term is computed using the Pile dataset (Gao et al., 2020). Since we applied random past token masking, dropout is not used in our experiments, as suggested by Liu et al. (2022). When finetuning, we combined three human feedback datasets, and the data was sampled proportionally to their size to ensure balance across the datasets.

What makes for a good dialogue? Roughly speaking, a good dialogue is a conversation that is both helpful and harmless. We would like you to consider these different dimensions of dialogues:

**Helpful**  
 For this axis, assess whether the conversation promotes active listening and open-mindedness. Helpful dialogues typically encourage the exchange of diverse perspectives and ideas. Secondly, look for evidence of empathy and respect in the dialogue; a constructive conversation should foster an atmosphere where all parties feel valued and understood. Thirdly, examine whether the dialogue leads to tangible outcomes or solutions to the issues being discussed. A helpful dialogue often results in progress or resolution, rather than perpetuating conflicts or misunderstandings.

**Harmless**  
 For this axis, response to questions should not contain potentially harmful information such as “information on building a bomb”. Consider the overall tone and language used during the conversation; a helpful dialogue should be free from hostility, personal attacks, or derogatory remarks. A dialogue is harmless if it does not contain any unsafe or potentially harmful information.

Table 5: Instructions we provided to human labelers on evaluating dialogue quality.

Table 6: Examples of Natural language feedback. The task prompts are omitted for simplicity.

Source	Examples of natural language feedback
Summary	a good summary is: {positive} a bad summary is: {negative}
Summary	a bad summary is: {negative} a good summary is: {positive}
Summary	a good summary is: {positive} a worse summary is: {negative}
Summary	a bad summary is: {negative} a better summary is: {positive}
Shared	a good response is: {positive} a bad response is: {negative}
Shared	a bad response is: {negative} a good response is: {positive}
Shared	a good answer is: {positive} a bad answer is: {negative}
Shared	a bad answer is: {negative} a good answer is: {positive}
Shared	a good answer is: {positive} a worse answer is: {negative}
Shared	a bad answer is: {negative} a better answer is: {positive}
Shared	good: {positive} worse: {negative}
Shared	bad: {negative} better: {positive}
Shared	good: {positive} bad: {negative}
Shared	bad: {positive} good: {negative}
Dialogue	you are a helpful assistant: {positive} you are an unhelpful assistant: {negative}
Dialogue	you are an unhelpful assistant: {positive} you are a helpful assistant: {negative}
Dialogue	you are a respectful and unbiased assistant: {positive} you are a disrespectful and biased assistant: {negative}
Dialogue	you are a disrespectful and biased assistant: {positive} you are a respectful and unbiased assistant: {negative}
Summary	give me a good summary: {positive} give me a worse summary: {negative}
Summary	give me a bad summary: {negative} give me a better summary: {positive}
Summary	let’s generate a good summary: {positive} let’s generate a worse summary: {negative}
Summary	let’s generate a bad summary: {negative} let’s generate a better summary: {positive}
Shared	let’s generate a good answer: {positive} let’s generate a worse answer: {negative}
Shared	let’s generate a bad answer: {negative} let’s generate a better answer: {positive}

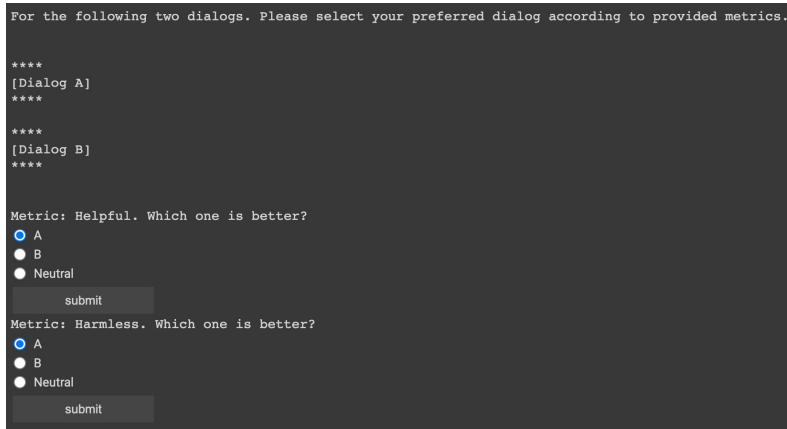


Figure 6: Screenshots of our labeling interface for rating dialog. For each metric, labelers are asked to choose preferred dialog.

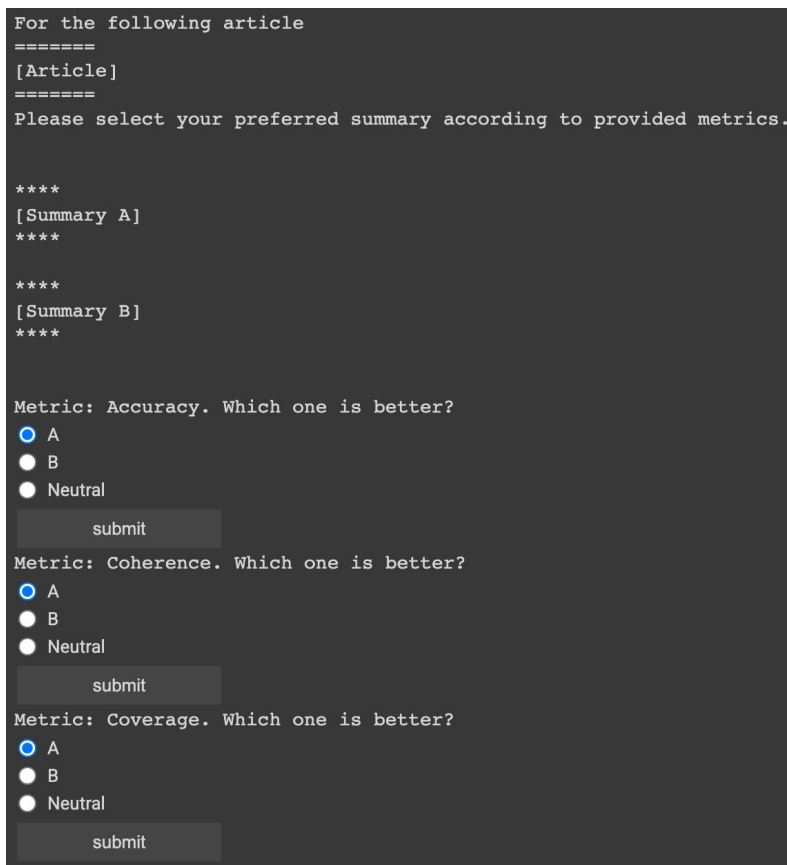


Figure 7: Screenshots of our labeling interface for rating summary. For each metric, labelers are asked to choose preferred summary.

## D HUMAN EVALUATION WEB INTERFACE

In Figure 7 and Figure 6, we show screenshots of our labeling interface, that all of our labelers use to rate data. Labelers can choose the preferred model output or choose tie in cases where two outputs seem to be of similar quality.

## E ADDITIONAL EXPERIMENTAL RESULTS

### E.1 EVALUATION ON CONTROLLABLE GENERATION

The controllable generation results are presented in Figure 8. The models are provided with three instructions to generate summaries of desired quality. The first instruction asks for a standard summary, while the second and third instructions ask for improved summaries conditioned on the previous summary generated by the model. We compare the performance of CoH with that of the RLHF model. The results indicate that while RLHF performs well in modeling human preferences and generates high-scoring summaries by following the first instruction, it fails to follow the second and third instructions, which implies that it cannot comprehend human intentions. On the other hand, the CoH-trained model is capable of understanding the intention of the instructions and generates better summaries in the second and third trials. We note that the controllable generation technique can be further investigated in various evaluation settings to enhance performance.

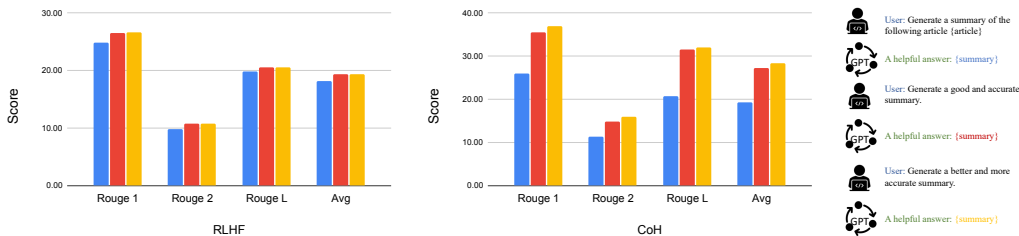


Figure 8: **Controllable generation.** (left): RLHF cannot follow instructions to generate improved summary. (middle): After finetuning on CoH, the model follows instructions to achieve controllable generations. (right): First instruction is standard, while second and third instructions ask for improved summaries.

### E.2 ALIGNMENT TAX

We conducted an evaluation on a diverse set of few-shot tasks that are commonly used in previous studies (Brown et al., 2020; Wang and Komatsuzaki, 2021) to assess the effectiveness of aligning models with human preferences. We use Language Model Evaluation Harness<sup>4</sup> for evaluation. The results are reported in Table 7. Interestingly, we found that the average performance of models that were finetuned using SFT decreased after alignment. This decrease could be attributed to the issue known as *alignment tax* in language models (Ouyang et al., 2022), which underscores the importance of human evaluation (Lee et al., 2022). On the other hand, our proposed method, CoH, showed moderate improvements over both the pretrained model and supervised fine-tuned model. This result suggests that CoH is less susceptible to the *alignment tax* issue.

### E.3 COMPARISON AGAINST CHATGPT DISTILLATION

The open-source human preference datasets utilized in this study are curated based on human preferences for model generations. Although these preferences offer valuable learning signals as we have demonstrated in the experiments, the models responsible for these responses are notably less capable than proprietary models like ChatGPT. As a result, the data quality from these open-source datasets falls short when compared to conversations between ChatGPT and users which is shared online on ShareGPT. Given that the ShareGPT data showcases superior quality and greater diversity than the open-source datasets, we are interested in how our approach CoH performs when applied to open-source human preference datasets, in comparison to the SFT approach used on ShareGPT data. To this end, we compared with Koala (Geng et al., 2023) which involves supervised finetuning LLaMA (Touvron et al., 2023) using ShareGPT data. It’s worth noting that we maintained consistency in the model

<sup>4</sup><https://github.com/EleutherAI/lm-evaluation-harness>

Table 7: **Alignment Tax on Few-Shot Benchmark:** The results of our experiments on few-shot NLP benchmarks using the **Language Model Evaluation Harness** are presented in Table 7. We follow the same setup as in previous work (Brown et al., 2020; Wang and Komatsuzaki, 2021), including the splits for each task. The reported numbers for GPT-J are taken from its original paper, while the numbers for other models are reported by us. We average the results over 5 random seeds.

Task	Zero-shot			One-shot			Few-shot		
	GPT-J	SFT	CoH	GPT-J	SFT	CoH	GPT-J	SFT	CoH
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
<b>Average</b>	<b>40.60</b>	<b>40.54</b>	<b>40.95</b>	<b>42.53</b>	<b>42.53</b>	<b>43.14</b>	<b>39.38</b>	<b>39.59</b>	<b>39.98</b>

and training hyperparameters for both SFT and COH when applied to open-source datasets. Additionally, we integrated CoH with Koala by finetuning both the ShareGPT and open-source datasets; here, the open-source datasets provided both positive and negative examples, while ShareGPT contributed solely positive examples. We use the same human evaluation as Koala by hiring third-party human labelers to conduct pairwise comparisons of responses generated by various models. These evaluations were based on questions sourced from a holdout set exclusive to ShareGPT. Results presented in Figure 9 reveal that our approach CoH is on par with Koala in performance. Moreover, the combined approach of CoH +Koala show slightly better performance than Koala based on human ratings. Meanwhile, both C-SFT (conditional SFT) and SFT lag behind Koala considerably. This underscores the efficacy of CoH in leveraging human preferences for learning.

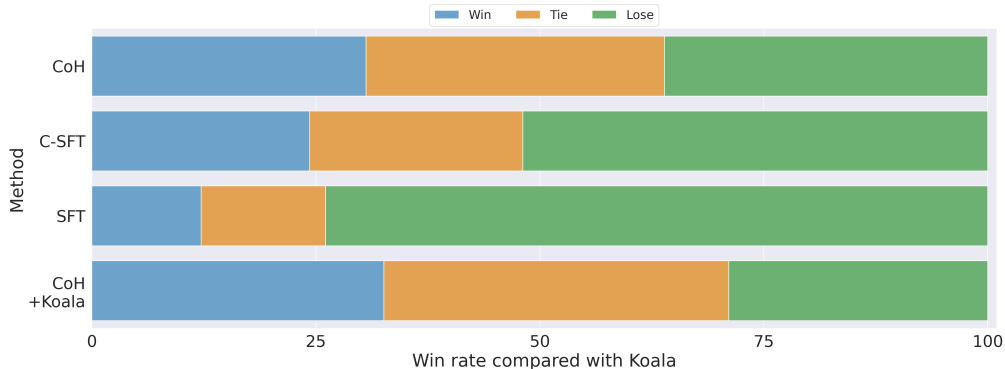


Figure 9: Evaluating various approaches with open source human preference datasets in comparison to ShareGPT’s supervised finetuned Koala.

F QUALITATIVE EXAMPLES

Table 8 and Table 9 show qualitative examples of summaries generated by GPT-J and CoH finetuned GPT-J. The examples are sampled from the validation split of dataset from Stiennon et al. (2020) which is based on TL;DR Reddit dataset (Völske et al., 2017).

Example 1	
Text	
<p>I don't know how to ask this without it sounding really silly, but this is a genuine problem. I created a throwaway because of too much personal detail in my main account. My boyfriend and I are both grad students and have been together for four months now. I absolutely love him. He's sweet, caring, tolerant, liberal, basically every single thing on my 'checklist', if I had one, is there. The problem, though, is he talks VERY slowly. I mean his speed is really slow. If I ask him anything, he takes about 3 seconds before he starts to talk, then when he starts, the words go at a very slow pace, and he even pauses for 3-5 seconds in the middle of his sentence. I get really impatient and sometimes just feel like telling him to hurry up (of course I don't, but I really feel like doing so). I asked him today, very gently, why he spoke so slow, and he said he just gets thoughts very slowly in his brain. He then went on to say he's just generally slow at thinking. I felt really sad when he said this, I don't know why. He's a Math PhD student, and we are at a pretty good school, and his research has been going slow for a while now (but this is totally normal in grad school), so I think that response was kind of a self-deprecating one since he's already low on self-esteem. Anyway, my point is, his speed of talking is just a part of him, it can't be changed. I love him so much otherwise but I just want to know how to deal with this. Conversation is so important in a relationship! How do we manage? He's an introvert and talks very little in general; I suspect this has something to do with his pace of talking?</p>	
GPT-J summary	
<p>A woman is upset because her boyfriend talks too slowly, and she doesn't know what to do. She is bothered by his slow speech, and she thinks it might be due to his low self-esteem and introverted nature.</p>	
CoH-tuned GPT-J	
<p>A woman is in love with her boyfriend who meets all her criteria except he speaks very slowly. He explains that he gets thoughts slowly and is slow at thinking. She wants to know how to manage this in their relationship as conversation is important to her.</p>	
Example 2	
Text	
<p>This is my first post so please be kind :) I know that lots of people often feel confused when they come out of a long-term relationship. They think they have forgotten how to be single, or how to flirt/date. I am one of these people. The problem is, my relationship started when I had just turned 16. I have never been single - as an adult. That might sound silly. But the only time I have ever flirted or dated was as an over-confident, hormone-riddled teenager. Now I have a pretty demanding job, responsibilities blah blah... And I just don't know how to this! I'm no way in a rush to get into a new relationship, but that doesn't mean I want to be completely alone in the mean time. If anyone has experienced anything similar, or just generally has some advice, it would be greatly appreciated!</p>	
GPT-J summary	
<p>Someone is struggling after coming out of a long-term relationship that started when they were 16.</p>	
CoH-tuned GPT-J	
<p>A person is seeking advice after coming out of a long-term relationship that began when they were 16 years old. They feel confused about being single as an adult and are looking for tips on how to flirt and date again, as they don't want to be completely alone during this period.</p>	

Table 8: Qualitative examples of GPT-J and CoH tuned GPT-J on the summarization benchmark. The input texts are sampled from the validation split of the dataset from Stiennon et al. (2020), which is based on the TL;DR Reddit dataset (Völske et al., 2017).

## Example 3

Text

Just wanted to say sorry again for wasting your time helping me out! I know I'm probably blind or something. > . < I read the FAQ, and I was able to do some of the things to get started but many of the things I am still confused on. I was able to calculate my BMR (By the way is there one calculator that is better than the others? I got a few different numbers.) However, when I went to the link on the FAQ to calculate my BF, the page was gone. Is this necessary? However, I know even without my BF that I would be classified as "skinny fat" at 5'9 and 156 pounds with a decent layer of fat. The next part I am confused on is diet. Most of the provided resources provide diets to try don't seem like they would lend themselves well to everyday/lifetime eating. I don't know how to eat healthy. The last time I tried to truly get in good shape I did it all wrong. I'm talking eating only 800-900 calories a day, counting obsessively, cutting out all my favorite foods, and interspersed with bouts of binging on aforementioned foods. Today, counting calories freaks me out because I fear that I will go back to that. I am trying to get into real fitness, like going to a gym or bodyweight fitness, but I cannot afford a gym membership and I can't seem to find something sturdy enough to use as a pullup bar for bodyweight. What are my other options because I do want to build muscle. My final question is that I have no idea what I am truly aiming for. I want to lose my layer of fat and build muscle, but I am not sure if there's a goal weight involved or what sort of time table I am looking at or how many calories I would need to eat. This brings us full circle as I restate that I have no idea what I am doing.

GPT-J summary

The individual is expressing regret for taking up the reader's time in attempting to assist them. They are unable to financially support a gym membership and lack a durable apparatus for executing bodyweight exercises, such as pull-ups.

CoH-tuned GPT-J

The person is apologetic for taking up the reader's time and expresses confusion about some of the steps involved in getting started with a fitness journey. They were able to calculate their BMR but couldn't find a working link to calculate their BF%. They feel they are "skinny fat" and are unsure about how to eat healthily without resorting to obsessive calorie-counting. They can't afford a gym membership or a sturdy pull-up bar for bodyweight exercises. They are also uncertain about their goals for losing fat and building muscle, including their target weight, time frame, and caloric intake.

## Example 4

Text

This happened a couple of hours ago, and i'm still a bit nervous about all this. I'm what most would consider a hardcore gamer, and as most people into this, they usually cherish and collect their favorite games and such. Today I skipped work because i woke up feeling like something had banged on my left side, hurting all over (ear infection, it's a reoccurring thing, I need to get this checked more often than not.) and decided to stay home and rest a bit more. I stayed home all day and at about 6 my dad rolls around to pick me up to go to the ER to see what's going on with my ear. During the following hour and a half I spent on the ER to get this checked and pick up the prescription the doctor gave me, I get home and find some clothing on my porch, lying on the path to my entrance. My dog was feasting on some plastic bag and at first I thought it was just some kids who'd have thrown a bag of old clothing onto my yard. As soon as I get up to my room though, things took a dire turn for me. I froze, looking baffled at what was right in front of me. My monitor was all crooked and the cables were off, and I look to the left and see this gaping hole where all my consoles were. Then it finally sunk into me: Someone robbed my house. [Some words have been omitted because they may appear inappropriate.]

GPT-J summary

A gamer skipped work due to an ear infection and got robbed. The shock and disbelief of the situation sink in, leaving the person feeling violated and anxious about the loss of their cherished belongings. The incident serves as a reminder of the vulnerability and unpredictability of life, leaving them with a sense of unease and distrust towards others.

CoH-tuned GPT-J

A gamer stayed home from work due to an ear infection and later went to the ER. Upon returning home, they discovered their house had been robbed and their game consoles were missing.

Table 9: (Cont. from previous page) Qualitative examples of GPT-J and CoH tuned GPT-J on the summarization benchmark. The input texts are sampled from the validation split of the dataset from Stiennon et al. (2020), which is based on the TL;DR Reddit dataset (Völske et al., 2017).