

- adding human responses is not feasible
 - human references can be binding factor for open ended generation.
 like responses in Dolly are very subjective. Dont want to punish
 model generation by evaluating on it.

HREF: HUMAN RESPONSE-GUIDED EVALUATION OF INSTRUCTION FOLLOWING IN LANGUAGE MODELS

Xinxi Lyu¹ Yizhong Wang^{1,2} Hannaneh Hajishirzi^{1,2} Pradeep Dasigi¹

¹Allen Institute for AI ²University of Washington

ABSTRACT

Evaluating the capability of Large Language Models (LLMs) in following instructions has heavily relied on a powerful LLM as the judge, introducing unresolved biases that deviate the judgments from human judges. In this work, we reevaluate various choices for automatic evaluation on a wide range of instruction-following tasks. We experiment with methods that leverage human-written responses and observe that they enhance the reliability of automatic evaluations across a wide range of tasks, resulting in up to a 3.2% improvement in agreement with human judges. We also discovered that human-written responses offer an orthogonal perspective to model-generated responses in following instructions and should be used as an additional context when comparing model responses. Based on these observations, we develop a new evaluation benchmark, Human Response-Guided Evaluation of Instruction Following (HREF), comprising 4,258 samples across 11 task categories with a composite evaluation setup, employing a composite evaluation setup that selects the most reliable method for each category. In addition to providing reliable evaluation, HREF emphasizes individual task performance and is free from contamination. Finally, we study the impact of key design choices in HREF, including the size of the evaluation set, the judge model, the baseline model, and the prompt template. We host a live leaderboard that evaluates LLMs on the private evaluation set of HREF^{1,2}.

1 INTRODUCTION

Automatic evaluations of instruction following abilities in Large Language Models (LLMs) has recently received significant attention (Zheng et al., 2023; Li et al., 2023; Lin et al., 2024; Li et al., 2024; Chiang et al., 2024). To make evaluation efficient and enable rapid iteration over modeling choices during development, prior work has approximated human judgments of the model response quality using by using a powerful language model as a judge (LLM-as-a-Judge). Although model judges have been shown to exhibit biases due to superficial features, such as the length of responses, prior work has indicated that such biases can be addressed (Dubois et al., 2024) to improve the reliability of these judgments. However, the analysis of such biases and the corresponding debiasing techniques developed in prior work are based on a distribution of tasks that is not representative of the full range of applications of instruction-tuned language models.

In this work, we reevaluate various choices for automatic evaluation on a wider range of instruction following tasks (Section 2). We choose a task distribution closely aligned with those typically used to train instruction-tuned models (Ouyang et al., 2022), and measure the agreement between human and model judges by comparing LLM-as-a-Judge and embedding-based similarity approaches. We experiment with using human-written reference responses in the process—by including them as additional context in the LLM-as-a-Judge or measuring embedding similarity between model responses and human responses—and observe that they can and observe that they enhance the reliability of automatic evaluation across many tasks, resulting in up to a 3.2% improvement in agreement with human judges (Section 3.2). Our analysis also provides insights into how human-written responses are helpful. We discovered that human-written responses offer an orthogonal perspective to model-

¹Leaderboard: <https://huggingface.co/spaces/allenai/href>.

²Code: <https://github.com/allenai/href>.

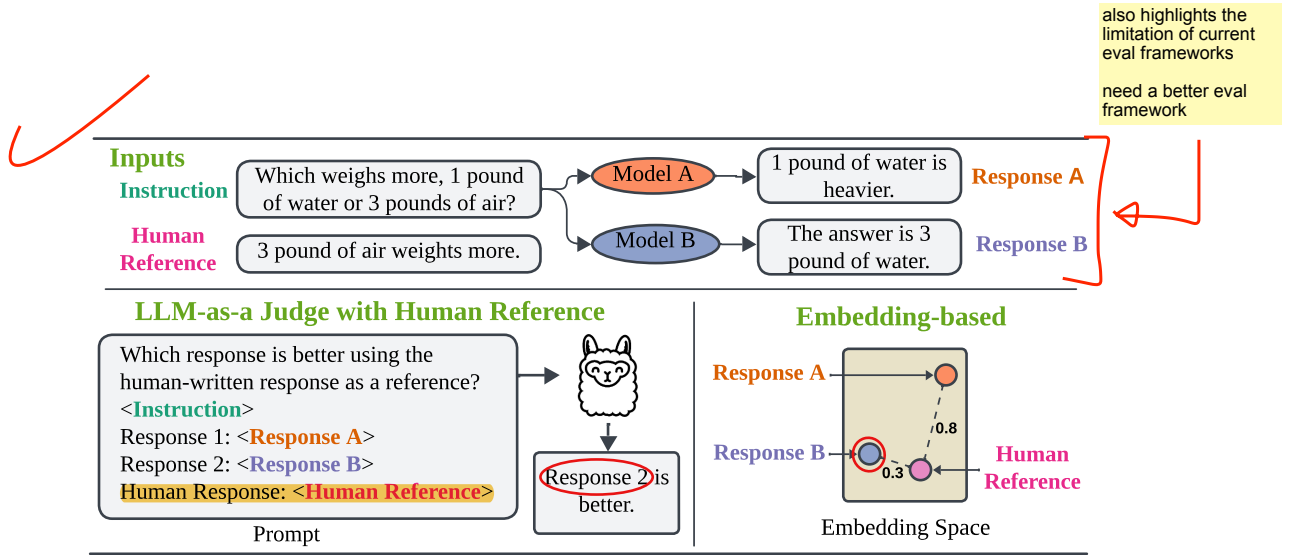


Figure 1: An overview of our composite method leverage the human-written response as reference to judge between two responses given an instruction. The example and the prompt shown in the figure are not exact. See details of these methods in Section 2.2.

generated responses in following instructions and should be used as a complementary reference when comparing model responses.

Based on these observations, we develop a new evaluation benchmark with 4,258 human-written prompts and reference responses spanning 11 task-categories. We use a composite evaluation setup that uses the most reliable evaluation method for each task-category. Given the reliance on human-written responses, we name this benchmark Human Reference-guided Evaluation of instruction Following (**HREF**). Our new benchmark additionally addresses two important limitations in existing instruction-following evaluations: *the ease of contamination* and *a limited focus on individual tasks*.

Ease of contamination. A consequence of the open availability of the existing instruction following evaluation sets is that these datasets can (often inadvertently) end up in the post-training datasets. For instance, Lambert et al. (2024) show that datasets containing real user conversations with language models, like LMSys-Chat 1M contain significant portions of AlpacaEval data in them. Training on such contaminated datasets can lead to inflated model performance on these benchmarks. To deal with this issue, we create separate development and test splits of HREF, and keep the test split private.

Limited focus on individual tasks. Prior instruction-following evaluations either focus on a small set of tasks (Li et al., 2023; Zheng et al., 2023) or use a relatively small sample of real user interactions with language models (Lin et al., 2024) where some tasks are under-represented³. As a result, both these approaches result in evaluation datasets that provide limited actionable insights about the model development process at the individual task level, e.g., which skills to upsample in the training datasets. In contrast, we take a task-centric view of data curation with HREF. We start with a taxonomy of 11 task categories used in Ouyang et al. (2022) and collect more than 100 samples for each task category. We apply a task-specific evaluation method and report the result for each task category separately in order to provide a reliable evaluation and deliver insights about the task direction to improve a LLM on.

We study the impact of our design choices in HREF, including the size of the evaluation set, the choice of the judge model and baseline model, and our prompt template in Section 5. We build a leaderboard that uses the private test split of HREF.

2 EMPIRICAL BASIS FOR THE EVALUATION SETUP

In this section, we describe our experiment settings to explore how human-written response can be utilized to improve the reliability of evaluating the instruction-following capability of LLMs. Specifically, we construct a dataset for evaluating the evaluation methods, collecting human annotations

³WildBench has task categories identified post-hoc, and the smallest category has only 16 instances.

Task	Example
Brainstorming	What should I look for when buying a car?
Open QA	Do tomatoes contain MSG?
Closed QA	What was the first type of anesthesia used in surgery? On October 16, 1846, the first successful public demonstration of the use of ether for surgical anesthesia was performed, making pain-free surgery possible.
Extraction	I want to know the sizes ticks come in in a numbered list. Here is the texting I am talking about: Ticks come in three sizes depending on their life stage. They can be the size of a grain of sand, a poppy seed, or an apple seed.
Generation	Write a poem about tacos.
Rewriting	Rewrite the sentence in active voice. My vegetable garden was eaten by a donkey.
Summarization	Summarize this in one sentence: SOCIETY FOR HUMAN RIGHTS 1924 After being temporarily committed to a mental institution because of his sexual orientation, Henry Gerber, a German immigrant and World War I Army enlistee, establishes the Society for Human Rights, the first American homosexual rights organization.
Classification	If a poem is titled "Hide and Seek" does that sound childish or mature?
Fact Checking	What were the top 5 cited papers in AI from 2022?
Multi-Doc Document Synthesis	According to these reviews from Yelp and Google Maps, determine whether or not Moonshadows is an ideal location for a date night during my upcoming trip to Malibu. Yelp Reviews: \n 1. ... \n 2. ... \n 3. ... \n 4. ...
Reasoning Over Numerical Data	What is the increase in sales from Jan to Feb? \n Month Sales Expenses \n Jan 2,894 2,582 \n Feb 3,820 3,517 \n March 2,009 1,796

Table 1: **Task Examples.** Examples for each the 11 task category.

on the pairwise preference between response pairs (Section 2.1). We introduce three new automatic evaluation methods that leverage human-written responses (Section 2.2), and show the results of the experiment comparing them in Section 3.

2.1 HUMAN AGREEMENT SET CONSTRUCTION

In this subsection, we describe a dataset containing instructions, responses from models and humans, along with human annotated preferences. We refer to this dataset as the **human agreement set** which will be a subset of the final dataset described in Section 4.

2.1.1 INSTRUCTIONS AND RESPONSES COLLECTION

We construct a dataset of instructions, each associated with a human-written response, two candidate model responses, and multiple human judgment annotations indicating which model response is preferred⁴.

Task Selection. Prior benchmarks for evaluating instruction following include sets of instructions that are representative of real user interactions with publicly hosted language models. While evaluating on such datasets can inform how the model would perform in practice, the input distributions tend to be heavily skewed towards a small set of tasks as shown by (Lin et al., 2024; Chiang et al., 2024; Li et al., 2024). Consequently, the decisions regarding the evaluation setup, though based on rigorous human agreement experiments, may be biased towards a small number of tasks. In contrast, we begin with a taxonomy of 11 instruction-following tasks and build a dataset of instructions specifically targeting these tasks. Specifically, we select 8 tasks from the InstructGPT taxonomy (Ouyang et al., 2022)—Brainstorming, Open QA, Closed QA, Extraction, Generation, Rewriting, Summarization, Classification, and 3 additional tasks focused on scientific text understanding—Fact Checking, Multi-Document Synthesis, and Reasoning Over Numerical Data. See Table 1 for the definitions and the examples for each task.

⁴Data Link: https://huggingface.co/datasets/allenai/href_preference.

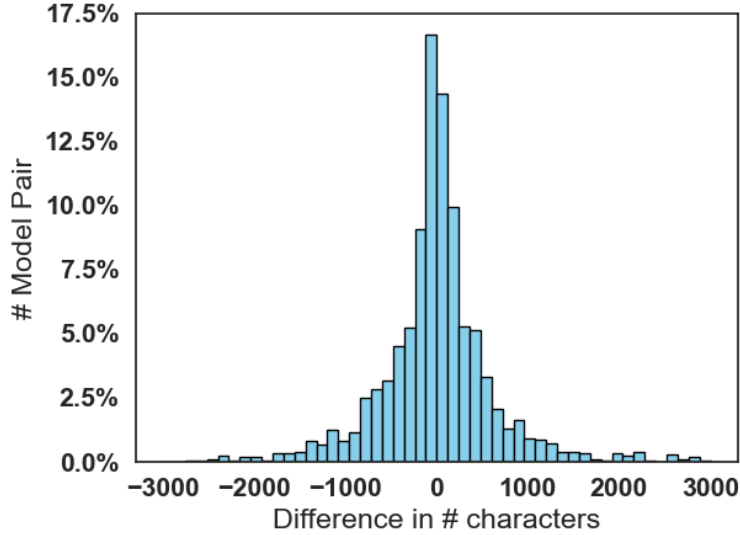


Figure 2: **The distribution of length difference between sampled model responses and the base line model responses.** The distribution is symmetrical.

Instruction Set. We sample instructions and human-written responses for 8 of the tasks from the No Robots dataset (Rajani et al., 2023). We sample data primarily from the test set, and for tasks that are not well represented in the test set, we additionally sample from the training set. For the remaining 3 scientific text understanding tasks, we hire human experts to write instructions and associated responses. We ended up with 438 pairs where all 11 categories are reasonably represented (See Figure 5).

Model Pool. In order to ensure the diversity of the responses, we build a model pool with 32 LLMs with sizes from 7B to over 100B and more than 10 different model families. See the full list of models in Appendix D.1.

Response Sampling. For each instruction, we sample responses from 4 distinct models each paired with the large and strong model, Llama-3.1-405B-Instruct-FP8. To avoid introducing implicit bias related to response length (e.g., the longer response is always better), we divide all model responses for each instruction into two groups based on whether they are longer or shorter than the baseline model responses. We then randomly sample 2 response from each of the 2 groups. To ensure high-quality of the response and to avoid repetitions in generation, we use a decoding temperature of 1.0 for all the models Figure 2 shows the resulting distribution of the length difference between sampled model responses and baseline model responses. The symmetrical distribution shows that both the shorter and the longer responses are roughly equally sampled.

2.1.2 HUMAN ANNOTATION COLLECTION.

We collected 4 human preference annotations for each instance in our human agreement set following the procedure described below. Importantly, the annotators are shown only the instructions and the two model responses per each instance, and not the human-written responses.

Annotator Selection. We recruited native English speakers from the U.S., the U.K., and Canada, who have a Bachelor’s degrees or above, and a prior approval rating over 99% from Prolific (First, 2014). We further screened annotators using a qualification test that required them to correctly annotate at least 9 out of 10 instances with easily distinguishable model response pairs. We assign the qualification task to 50 participants, and recruited 16 of them as our final group of annotators and paid them \$16 / hour.

Annotation Guidelines and Interface. We used the annotation guidelines from Li et al. (2023) with the following modifications: We slightly modified checklist of judging aspects, included two example annotations, and importantly allowed the annotators to choose “tie” when both the model

Method	General Category								Science Category			All
	Brn	OQA	CQA	Ext	Gen	Rew	Sum	Cls	FC	MDS	RND	
<i>Simple Baselines</i>												
Random	46.3	49.5	41.1	47.7	54.9	47.3	50.1	50.0	45.3	44.8	50.0	48.2
Shorter	43.8	61.6	65.6	57.6	49.0	45.5	56.6	52.9	57.8	42.8	53.5	53.2
Longer	50.1	33.3	28.3	38.1	49.1	52.6	40.5	43.3	39.0	54.8	45.8	43.1
<i>LLM-as-a-Judge</i>												
GPT4	60.0	49.1	54.8	67.8	66.6	68.6	53.3	52.1	56.3	60.0	69.0	59.4
GPT4-Tb	65.5	57.1	63.6	72.5	63.9	72.5	61.9	62.5	61.5	64.5	69.8	64.8
Llama-7B	59.9	55.5	49.5	64.8	56.1	64.3	57.5	57.8	63.8	56.3	64.5	58.7
Llama-70B	65.3	58.5	64.8	69.7	64.1	73.4	68.0	66.0	69.0	63.5	66.8	66.2
<i>LLM-as-a-Judge with human response</i>												
GPT4	60.1	53.3	63.1	69.3	65.6	70.1	59.0	57.5	59.8	64.8	72.3	62.6
GPT4-Tb	65.5	58.8	67.8	75.6	63.1	72.5	64.4	68.1	64.3	62.5	70.5	66.4
Llama-7B	56.5	53.1	47.2	65.8	53.1	63.0	54.3	55.5	68.5	52.5	63.8	56.8
Llama-70B	62.5	55.3	73.0	75.6	70.3	75.4	66.9	68.1	64.0	66.0	73.8	67.9
<i>Embedding-Based</i>												
RoBERTa-Large	48.4	60.5	68.8	69.9	64.1	64.8	61.0	61.9	71.5	55.5	70.0	62.5
Rouge	52.4	57.3	70.3	65.0	58.0	70.4	55.0	50.0	70.3	61.5	61.0	60.3
<i>Perplexity-based</i>												
Perplexity	47.9	51.5	57.1	51.8	48.8	42.8	52.9	42.1	54.9	52.8	59.7	50.3
<i>Composite</i>												
GPT4	60.1	60.5	70.3	69.9	66.6	70.4	61.0	61.9	71.5	64.8	72.3	65.6
GPT4-Tb	65.5	60.5	70.3	75.6	64.1	72.5	64.4	68.1	71.5	64.5	70.5	67.4
Llama-7B	59.9	60.5	70.3	69.9	64.1	70.4	61.0	61.9	71.5	61.5	70.0	65.0
Llama-70B (Ours)	65.3	60.5	73.0	75.6	70.3	75.4	68.0	68.1	71.5	66.0	73.8	69.4
Human	62.6	59.5	71.9	73.8	66.3	69.9	63.4	69.8	71.0	62.8	74.3	67.0

Table 2: **Human Agreement Rates of Different Evaluation Methods on 11 Categories.** All numbers are average LOO agreement rates in %. Bold numbers are the highest numbers with Llama-3.1-70B-Instruct for each categories, and we choose their corresponding methods to form the final composite method. When calculating Perplexity, we omit some instances in the human agreement datasets where the perplexity are not available with OpenAI models. Brn → Brainstorming; OQA → Open QA; CQA → Closed QA; Ext → Extraction; Gen → Generation; Rew → Rewriting; Sum → Summarization; Cls → Classification; FC → Fact Checking / Attributed QA; MDS → Multi-Document Synthesis; RND → Reasoning Over Numerical Data.

responses are indistinguishable in quality. See the full guideline in Appendix D.2. We build an website for collecting the annotations (See Appendix 15). To avoid potential bias in order of the responses, we randomly swap the two responses.

Statistics. We collected 4 annotations for each of 1,752 instances . The annotators spend around 180s with a standard deviation of 79s on each annotations on average with a tie rate of 4.9%.

2.2 EVALUATION METHODS

We evaluate a set of *pairwise* evaluation methods (Zheng et al., 2023), i.e., those that select the better response between two candidate model responses, based on their agreement with the human judgments we collected.

LLM-as-a-Judge involves prompting a powerful LLM to judge the better response between a pair of responses from two models. This is the method that prior work have prominently adopted. We experiment with Llama-3.1-7B-Instruct, Llama-3.1-70B-Instruct (Dubey et al., 2024), GPT-4, and GPT-4-Turbo (Achiam et al., 2023) as the judge model in our experiments. See Appendix C for the prompt template we use. Note that we allow the methods to judge ‘tie’ between the two model responses.

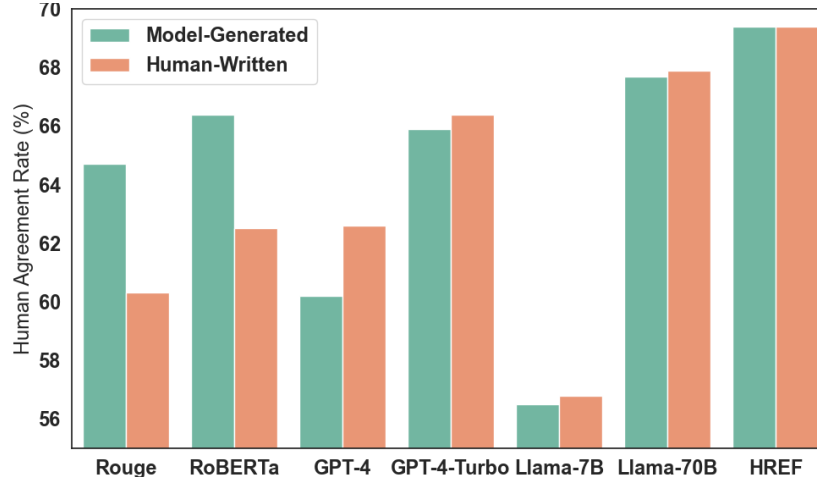


Figure 3: **Human Agreement Rate using model-generated v.s. human-written responses.** Human response outperforms model response for LLM-based evaluation methods but underperforms for embedding-based evaluation methods.

LLM-as-a-Judge with human response is similar to LLM-as-a-Judge except that it embeds human-written response into the prompt and instructs the judge to refer to it. See Appendix C for the prompt template we use. We experiment with Llama-3.1-7B-Instruct, Llama-3.1-70B-Instruct (Dubey et al., 2024), GPT-4, and GPT-4-Turbo (Achiam et al., 2023) as the judge model in our experiments.

Embedding-based methods computes the similarity between the text embeddings of a model response and a human-written response, using the resulting score to select the response with the higher similarity. To obtain the embedding, we experiment with the text embedding space output from RoBERTa-Large (Liu et al., 2019) and Rouge (Lin, 2004).

Perplexity-based methods calculates the perplexities of the human-written answer conditioned on the instruction for both the models, and selects the model with lower perplexity.

Random uniformly select one of the two responses.

Shorter / Longer naively prefers the shorter / longer response.

Composite select the best method from LLM-as-a-Judge, LLM-as-a-Judge with human response, and embedding-based methods for each category.

2.3 COMPUTING HUMAN AGREEMENT

Following Li et al. (2023), we use the **Leave-One-Out (LOO) agreement rate** to evaluate the agreement between a method’s output and the 4 annotations for each sample. Concretely, we compute the frequency with which the evaluation method’s output matches the mode of each combination of 3 out of 4 human annotations, then average the results across all 4 possible combinations. We report the *human agreement rate* as the average of LOO agreement rate over the all response pairs. To calculate the agreement rate within the human annotator themselves, we treat the remaining annotation as the “model” prediction for each combination of 3 annotations and perform the same calculation. See Appendix D.4 for more details.

3 RESULTS

In this section, we present the results from the experiment described in Section 2, and we provide additional insights into why human-written responses are helpful in improving the evaluation methods.

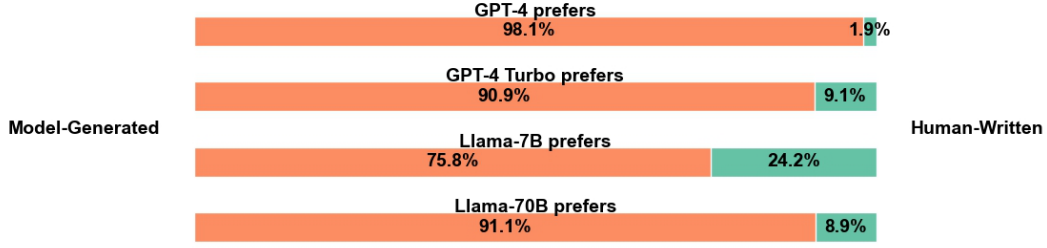


Figure 4: **Judges Preference Between Model-Generated Responses v.s. Human-Written Responses.** Model-generated responses are in great favor of all the judges.

3.1 MAIN RESULTS

Table 2 presents the results of the human agreement analysis.

Human agreement rates varies across task categories. Tasks such as Brainstorming, Open QA, Summarization, and Multi-Document Synthesis, tend to have responses that vary in multiple dimensions, including general content, level of details, tone, etc. We observe that both the inner-agreement rate among human annotators and the agreement rates across all evaluation methods are lower within these task categories, indicating that humans apply divergent standards for judging LLM responses and weights various dimension of such open-ended responses differently. Conversely, categories that tends to have easily verifiable answers, including Close QA, Extraction, Classification, and Reasoning Over Numerical Data, appears to have higher agreements. Note that although Rewrite contains many open-ended instructions, a large portion of the instructions are verifiable as they ask for specific tone or format of the response. These findings highlight the importance to evaluate LLMs on specific task categories.

Llama-3.1-70B-Instruct is the best judge. Llama-3.1-70B-Instruct outperforms GPT-4 by 6% and GPT-4-Turbo by 1.5% without human responses, achieving the closest agreement rate compared to the human. It also outperforms GPT-4 by 4.2%, GPT-4-Turbo by 1.3%, and even humans by 0.9% using human responses on average.

Human-written responses improve agreement with human judgments. Across all models except Llama-3.1-7B-Instruct, embedding human-written responses into the prompts and using them as additional context frequently improves agreement with human judgments. The performance drop with Llama-3.1-7B-Instruct is likely because LLMs have to reach a certain capability threshold so that they understand how to properly utilize the human-written responses. In task categories Close QA, Extraction, Generation, Rewriting, Classification, Multi- Document Synthesis, and Reasoning Over Numerical Data, using human-written responses brings an increment of 4.8% on average in agreement with human for using Llama-3.1-70B-Instruct as the judge. For OpenQA, Summarization, and Fact Checking, we observe that human-written response improves agreement with human judgement for GPT-4 and GPT-4-Turbo but not for Llama models. This suggests that the capability of properly leverage human-written responses as additional context is inconsistent across different models for these task categories. We also see that RoBERTa-Large is able to deliver the highest agreement rate with human on Open QA and Fact Checking. These results show that, despite that the annotators who write the human response and the ones who annotate the preference are two different groups, a human-written response can help improve the judgment by serving as an additional context or a comparable reference. We will talk about the insights around the usefulness of human-written responses in the following Section 3.2.

Choosing the best method for each category. With the new set of evaluation methods that leverage human-written responses, we are provided with the option to select the best evaluation methods for each task categories and compose the final composite methods. Overall, the resulting composite method with Llama-3.1-70B-Instruct achieves 1.5% higher in human agreement rate than only using

Benchmark	Size	Eval	BM	Judge	TaskCent	Private	HumResp
MT-Bench	80	Score	—	gpt4	✓	✗	✗
AlpacaEval 2.0	805	PWC	gpt4-turbo	gpt4-turbo	✗	✗	✗
Chatbot Arena	—	PWC	—	Human	✗	✓	✗
Arena-Hard	500	PWC	gpt4-0314	gpt4-turbo	✗	✗	✗
WildBench	1,024	Score/PWC	gpt4-turbo	three models	✗	✗	✗
HREF	4,258	PWC	Llama-405B	Llama-70B	✓	✓	✓

Table 3: **Benchmark Comparision.** A comparison between the existing instructional LLM evaluation benchmarks and HREF. **TaskOrit** refers to whether the instructions are task-oriented. **PWC** refers to the paired comparison. HREF has the largest evaluation set, is the only benchmark that uses open-weight models (Llama-3.1 Instruct) as both the baseline model (BM) and the judge, is built with task-centric instruction, is completely private, and uses human-written responses (HumResp) to facilitate preference judgment.

Llama-3.1-70B-Instruct as a judge with human reference, outperforming human annotators’ inner agreement rate by 2.4%.

3.2 ANALYSIS: LEVERAGING HUMAN REFERENCES

In order to understand the unique value of human-written responses, we compare them directly against *model-generated response* proposed in Zheng et al. (2023).

Human-written responses are more useful than model-generated responses with LLM-as-a-Judge. We use generate responses from GPT-4-Turbo for the instructions in the human agreement set and repeat the experiments in Section 2 with model-generated responses. Figure 3 demonstrates a comparison between using human-written responses and model-generated responses. We observe that with LLM-as-a-judge methods, human-written responses display higher agreement rates than model-generated responses across all judge models. This demonstrates that **references written by humans are consistently more useful than those generated by even the strongest LLMs**. With embedding-based evaluation methods (RoBERTa and Rouge), using model-generated responses display higher agreements than human-written responses. This is due to the fact that model-generated responses are syntactically and stylistically more similar to each other than to human-written ones, likely biasing these simpler evaluation methods.

Why not directly compare against human responses? We experimented with a setup where we prompt each LLM judge in Section 2 to directly compare model responses with human responses. Figure 4 shows that, surprisingly, all the judge models strongly prefer model responses over human responses despite their judgments being more aligned with those of human annotators when using human responses as additional context. This is likely because that the judge models strongly prefer the stylistic characteristics of model-generated responses. However, humans may prefer the style of human-written responses and other impactful dimensions, such as correctness, which are overlooked by the judge models. This demonstrates that **human-written responses are much more effective as additional context or additional reference for comparing model responses, rather than serving as the sole reference for direct comparison in evaluating response quality**.

4 NEW BENCHMARK: HREF

Based on the insights that human-written responses significantly improves the evaluation of LLMs’ instruction-following capability, we construct a new evaluation benchmark, **Human Response-guided Evaluation of instruction Following (HREF)**. See Table 3 for an overview of the comparison between HREF and similar existing benchmarks. We release two evaluation sets in addition to the human agreement set we used for experiments described in Section 2: a private evaluation set and a public development set.

Public Development Set We adopt a subset of the No Robots (Rajani et al., 2023) test split as the development set, which contains 430 human-written instruction and response pairs covering 8 out of the same 11 task categories as the evaluation set (See Figure 5). The remaining three scientific text

Dataset	Size	# Category	Annotation	Release
Evaluation	4,258	11	✗	✗
Development	430	8	✗	✓
Agreement	438	11	✓	✓

Table 4: **HREF Subsets Comparison.** An comparison of important aspects among the three subsets.

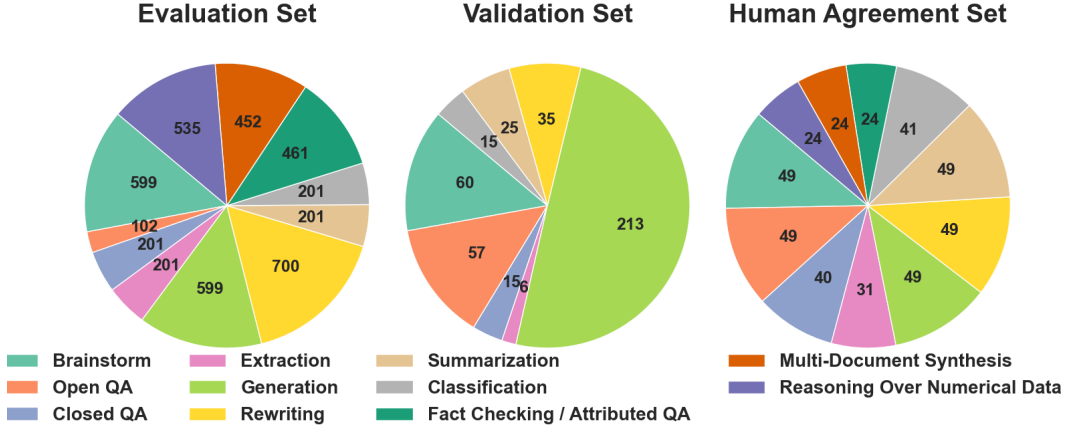


Figure 5: **Task Categorical Distribution of the three subsets in HREF.** Left: evaluation set; Middle: development set; Right: human agreement set.

understanding tasks are exclusive to in the evaluation set of HREF and can be considered held-out tasks. We generate a baseline model response from Llama-3.1-405B-Instruct-FP8 for each instruction. We will later show that the rankings on this set highly correlate with those from the evaluation set in Section 4.3⁵.

4.1 PRIVATE EVALUATION SET

Instruction and Human Response Collection. We hire human experts to write instructions and corresponding responses specifically targetting the taxonomy of tasks shown in Table 1. This results in 4,258 high quality instruction-response pairs. Figure 5 Left shows the resulting distribution of the instructions.

Baseline Response Generation. We generate a baseline response for each instruction to be compared against by a target model using the open model Llama-3.1-405B-Instruct-FP8 using greedy decoding. We compare this model with other choices for baseline models in Section 5.1.

4.2 EVALUATION DETAILS

Pipeline. For a target model, we first generate its response to each instruction to compare against the baseline model response using HREF, and consider it a win if HREF either prefers the target model response or selects a tie. To obtain the final **expected win rate**, we compute the frequency of wins for the target model across all data points.

Method Details. Following the observation from Section 5.1, we use the composite method with Llama-3.1-70B-Instruct as the judge model.

Decoding Strategy. For reproducibility, we choose greedy decoding for these models. We find that this choice does not significantly impact the evaluation results—we find a high correlation (0.98 Spearman and 0.99 Pearson) between the results obtained from using greedy decoding and those obtained from using a temperature of 1.0 on our development set.

⁵Data Link: <https://huggingface.co/datasets/allenai/href>.

Rank	Model	General Category								Science Category			All
		Brn	OQA	CQA	Ext	Gen	Rew	Sum	Cls	FC	MDS	RND	
1	Llama-3.1-70B-Instruct	48.6	84.3	55.2	45.0	45.5	44.6	43.3	54.5	50.3	49.8	57.1	49.8
1	Mistral-Large-Instruct-2407	54.5	58.8	35.1	39.6	51.4	50.5	45.3	44.5	48.5	59.1	26.0	47.6
3	Qwen2.5-72B-Instruct	51.4	71.6	30.2	40.8	47.7	46.0	36.4	43.5	45.6	57.5	23.4	44.8
3	Qwen1.5-110B-Chat	47.7	80.4	31.4	30.4	43.9	39.0	37.1	49.5	38.8	51.7	39.8	43.0
3	Llama-3.1-Tulu-3-70B-DPO	51.3	56.9	35.9	39.1	46.1	43.7	27.7	41.5	43.2	51.5	25.4	42.8
3	Llama-3.1-Tulu-3-70B	50.6	58.3	35.1	38.1	44.4	44.1	28.7	42.8	42.1	50.2	26.0	42.3
3	Mistral-Small-Instruct-2409	47.0	67.6	32.2	36.4	42.4	44.9	40.1	39.3	39.7	54.7	25.8	42.2
8	Llama-3.1-8B-Instruct	46.7	79.4	40.6	32.4	36.3	36.7	32.9	43.0	29.9	35.8	39.3	38.5
9	Yi-1.5-34B-Chat	49.7	71.6	25.2	27.0	36.2	30.1	21.5	35.8	31.8	41.1	28.8	35.4
9	Qwen2-72B-Instruct	42.4	71.6	24.8	32.4	31.7	29.1	15.3	40.0	38.4	26.4	47.2	35.0
9	OLMo-2-1124-13B-Instruct	37.7	59.8	30.9	20.3	34.8	36.0	32.2	44.0	25.9	44.6	28.4	34.7
9	Phi-3-medium-4k-instruct	27.7	81.4	34.9	24.3	26.2	28.1	24.0	45.3	37.2	17.2	60.2	33.3
9	Llama-3.1-Tulu-3-8B-DPO	40.9	63.7	26.0	19.6	36.6	33.7	22.8	33.1	27.9	45.0	19.9	33.2
9	Llama-3.1-Tulu-3-8B	41.6	58.8	23.5	18.3	35.6	34.0	21.0	32.3	28.0	41.6	21.9	32.7
15	OLMo-2-1124-7B-Instruct	38.1	47.5	19.6	16.8	32.3	30.1	17.6	26.9	18.3	36.4	15.8	27.7
15	tulu-2-dpo-70b	17.5	60.8	26.5	21.0	22.9	21.5	11.9	30.8	23.2	18.1	55.6	25.8
17	Llama-2-70b-chat-hf	22.3	61.3	30.0	26.0	19.2	17.1	21.8	34.3	19.1	13.2	48.5	24.5
17	Mistral-7B-Instruct-v0.3	24.1	60.8	21.3	14.1	19.1	18.5	16.8	27.1	17.3	25.5	46.5	24.3
17	Llama-3.1-Tulu-3-70B-SFT	2.8	95.1	37.6	28.2	13.0	11.7	4.0	36.6	35.3	10.5	60.4	23.6
20	WizardLM-13B-V1.2	16.9	63.2	22.8	17.1	15.8	14.7	7.7	33.6	9.7	11.3	43.9	19.6
20	Llama-2-13b-chat-hf	16.8	52.5	21.8	19.8	15.9	14.2	19.8	27.4	11.5	9.1	42.0	19.4
20	tulu-v2.5-ppo-13b	37.7	21.6	6.4	10.9	16.4	15.5	16.1	12.4	11.7	21.9	21.4	19.0
20	tulu-2-dpo-13b	9.3	68.6	17.8	13.6	13.3	15.6	9.2	24.9	10.9	10.2	50.2	18.4
20	vicuna-13b-v1.5	4.1	81.4	26.5	13.9	10.2	11.9	7.9	25.6	10.6	5.5	55.8	17.4
25	Llama-3.1-Tulu-3-8B-SFT	1.5	87.7	31.9	17.6	6.8	7.4	0.7	31.3	18.7	4.0	53.9	17.0
25	Llama-2-7b-chat-hf	16.4	41.7	21.3	12.9	13.3	12.9	9.7	19.2	7.3	7.5	33.5	15.6
27	vicuna-7b-v1.5	2.9	75.5	20.3	10.6	8.2	7.0	2.7	24.1	6.7	4.2	52.8	14.3
27	tulu-2-dpo-7b	4.9	55.9	14.6	9.9	10.0	9.9	6.9	18.4	5.4	6.0	46.5	13.8
29	OLMo-7B-SFT-hf	1.2	80.4	13.4	6.9	6.7	4.4	1.5	22.4	4.7	2.4	50.0	12.1
29	mpt-7b-chat	0.5	73.0	12.1	3.7	5.0	3.9	1.5	21.1	4.1	1.3	47.4	10.8
29	koala-13B-HF	1.0	70.1	15.8	8.4	4.3	5.0	1.0	16.4	5.2	2.2	39.8	10.4
32	gpt4all-13b-snoozy	0.7	90.2	3.7	6.2	6.4	5.7	1.0	11.7	5.1	1.3	35.9	9.9
32	OLMo-7B-0724-Instruct-hf	8.9	36.3	5.7	5.2	5.9	5.1	5.7	8.0	5.0	7.3	22.7	8.8
32	dolly-v2-12b	0.3	75.5	12.4	7.4	2.8	2.0	0.0	6.7	3.3	0.9	39.6	8.6
32	koala-7B-HF	0.4	65.7	8.2	7.9	3.3	3.4	0.0	9.7	3.1	1.3	38.7	8.6
32	dolly-v2-7b	0.2	70.6	9.4	5.0	2.0	1.3	0.0	6.2	3.6	0.2	42.0	8.2
37	oasst-sft-1-pythia-12b	0.2	62.3	2.0	1.5	1.7	0.6	0.0	3.2	2.4	0.9	23.2	5.1

Table 5: **expected win rates of all 37 starting models evaluated on the evaluation set of HREF.** All numbers are in %. (i) indicates the ranking. Brn → Brainstorm; OQA → Open QA; CQA → Closed QA; Ext → Extraction; Gen → Generation; Rew → Rewriting; Sum → Summarization; Cls → Classification; FC → Fact Checking / Attributed QA; MDS → Multi-Document Synthesis; RND → Reasoning Over Numerical Data;

Prompt Template. To reduce the difference between the model judge and human annotations in terms of their annotation criteria, we adopt the prompt template given to the human annotators (See Appendix D.2) and carefully modify it for LLM prompting (See Appendix C). We compare this with other choices of prompts in Section 5.2.

Expected Win Rate. Because we allow tie in LLM-as-a-Judge both with and without human response, we define **expected win rate** as the sum of the frequency that our composite method prefers the target model over the baseline model and **half** the frequency that our composite method selects a tie, over all samples.

Note that we keep the option of judging ties and consider it as a win.

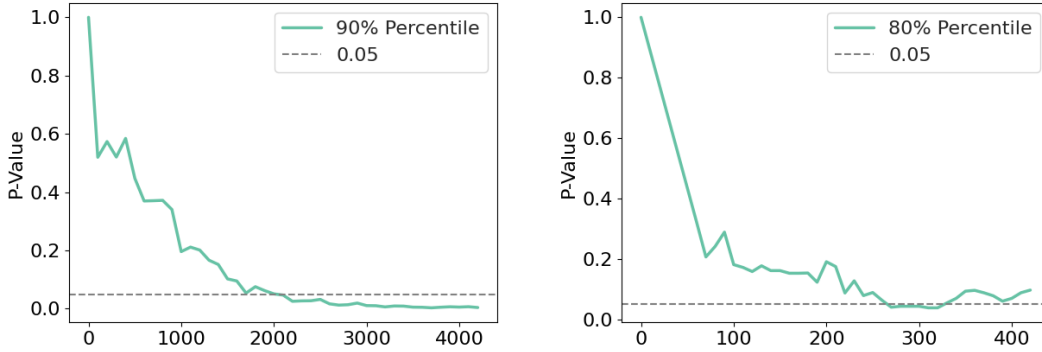


Figure 6: **P-values of paired T-test on annotations across 13 models on the evaluation and development set.** Evaluation set on the left; Development set on the right. We show the average and 90th and 80th quantile p-values from doing paired t-test among all model pairs among 13 models with different numbers of annotated samples used.

4.3 RESULTS ON CURRENT LLMs

We evaluate 37 LLMs with a variety of model families and sizes on HREF as the initial benchmark. Table 5 presents the results ranked by their total expected win rates, along with their expected win rates in each of the 11 categories. See the full table in Appendix A.

In general, LLMs with larger sizes display higher expected win rates, and such trends holds consistently within the same model family. For example, Llama-2-70b-chat-hf holds a higher expected win rate than Llama-2-13b-chat-hf on average. Also note that model expected win rates vary across different categories. For example, while Mistral-Large-Instruct-2407 a high average expected win rate among the models that we evaluate, it performs poorly in Open QA. This demonstrates the importance of focusing into the evaluation on individual task and underscores the advantage of HREF in providing task-centric evaluation.

Correlation with evaluation on the development set. We also evaluate the same group of LLMs on our development set with 8 categories (See Section 4.1), additionally with several GPT models. See the full results in Appendix A. We observe similar trends to those seen in the test dataset. To validate that model developers can expect a reasonable transfer of their results from the public development set to the private evaluation set, we calculate the correlation of the expected win rates between these two sets and observe high correlations: a Spearman correlation of 0.98 and a Pearson correlation of 0.99.

4.4 STATISTICAL SIGNIFICANCE

To ensure the reliability of our evaluation set in distinguishing between models, we evaluate HREF’s capability of statistically distinguishing among a diverse set of models of reasonable size. Specifically, we sample from a pool of 13 models following Li et al. (2023) but use a set of more recent and diverse models ⁶. For each pair of models, we apply a paired t-test to evaluate the null hypothesis that the preference predictions from the pair of models have identical expected values, and we measure the resulting p-values. We perform this analysis on both of the evaluation set and development set.

Capacity of the development and test sets. Figure 6 Left shows that with fewer than 2000 samples in the evaluation set, the p-values at 90th quantile falls below 0.05, which suggests that our evaluation set is able to statistically significantly distinguish between 90% of the model pairs. Similarly, Figure 6 right suggests that our development set is able to statistically significantly distinguish between 80% of the model pairs.

⁶Qwen1.5-110B-Chat, Mistral-Large-Instruct-2407, Yi-1.5-34B-Chat, tulu-2-dpo-70b, vicuna-13b-v1.5, Qwen2-72B-Instruct, mpt-7b-chat, koala-7B-HF, OLMo-7B-SFT-hf, dolly-v2-12b, Llama-2-7b-chat-hf, oasst-sft-1-pythia-12b, gpt4all-13b-snoozy-t=0.0

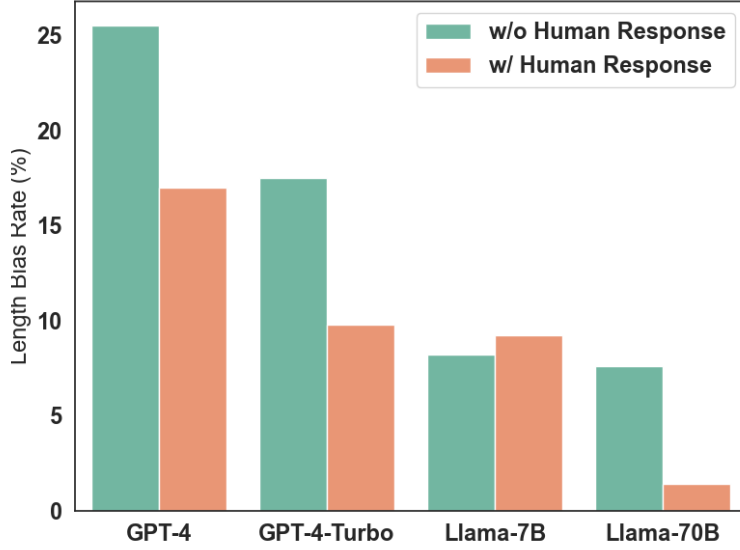


Figure 7: **Length Bias Rate of Different LLM Judges.** It is clear that Llama-3.1-70B Instruct has the least length bias, and such bias is further reduced when using human-written responses as additional context.

Relevance of HREF. As the size of the model pool and the strength of the models in the pool increase, the chance that a model pair will be indistinguishable (t-test with a p-value less than 0.05) will also increase. In other words, a larger evaluation set will be needed to distinguish more and stronger models. Hence, as the community keeps developing stronger models, we expect HREF, with the largest evaluation set among similar benchmarks, to remain relevant for longer.

5 DISCUSSION ON DESIGN CHOICES

In this section, we discuss the specific design choices and the advantages they bring to HREF, including the choice of the judge model for LLM-as-a-Judge, and choice of the baseline model, and the choice of the prompt template.

5.1 CHOICE OF THE JUDGE AND BASELINE MODELS

Unlike prior work (Li et al., 2023; Chiang et al., 2024; Zheng et al., 2023; Li et al., 2024; Lin et al., 2024), we choose Llama-3.1-70B-Instruct as the LLM judge and Llama-3.1-405B-Instruct-FP8 as our baseline model instead of GPT models. In this section, we discuss the rationale behind such choices.

High Human Agreement Rate with the Judge Model. Llama-3.1-70B agrees with human judgments the most on HREF as discussed in Section 3.

A Less Length-biased Judge Model. Previous work (Dubois et al., 2024; Lin et al., 2024; Li et al., 2024) has observed that the judge LLMs strongly prefer longer responses and has adopted length normalization methods to account for such bias. We quantify the length bias of various judge models on our human agreement set, by measuring the difference between each judge’s frequency of preferring longer responses versus the frequency of preferring shorter responses. We refer to this difference as the *length bias rate*. Since we explicitly control for response length while sampling responses in the human agreement set (see Section 2.1.1), we expect a model with no length bias to have a length bias rate close to 0% on our dataset. Figure 7 shows that Llama-3.1-70B-Instruct has the lowest length bias rate among all the four judge models that we experiment with. The use of human written responses further lowers its length bias rate to 1.4%. As a result, we chose not to add any length debiasing controls.

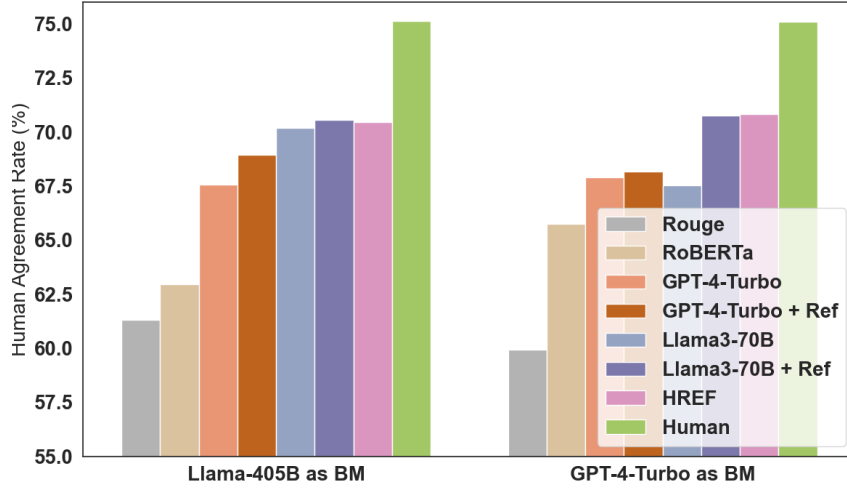


Figure 8: **Impact of Changing Baseline Model.** The average human agreement rates of various evaluators using two different baseline models. We observe very similar trends when using Llama-3.1-405B-Instruct-FP8 and GPT-4-Turbo as the baseline model.

A Less Costly Judge Model. Referring to the price from Lepton AI⁷, Llama-3.1-70B-Instruct is at least 12.5 times cheaper than GPT-4 Turbo, and 37.5 times cheaper than GPT-4. To minimize the computational requirements of evaluating new models, we restrict the evaluator size to 70B, which requires at most 4 A100 GPUs to run.

Invariant to a Baseline Model Change. To analyze the impact of the choice of the baseline model, we conduct the same experiments as in Section 3 but with GPT-4-Turbo as the baseline model on a subset of 1100 samples (275 instruction) of the human agreement set. Figure 8 compares the average human agreement rates of various evaluators using Llama-3.1-405B-Instruct-FP8 and GPT-4-Turbo as the baseline models. We observe similar trends with both the baseline models, indicating that the reliability of the evaluation setup is unaffected by using the open-weight Llama model instead of the closed GPT-4-Turbo model.

Reproducible Evaluations. Closed API models can be modified internally causing their outputs to change over time or eventually put to retirement, all of which makes evaluations relying on them irreproducible. In contrast, using an open-weight model like Llama-3.1-70B-Instruct, renders HREF more transparent and reproducible.

Feasibility of a Private Test Set. Using API models as judges requires sharing the instructions and responses with those models, meaning that the test set cannot remain truly private. Moreover, the common practice of synthesizing training datasets from such API models can potentially lead to test set contamination. By using an open-weight models running locally, we can keep our test data truly private to all models.

5.2 CHOICE OF THE PROMPT TEMPLATE

Unlike prior work such as AlpacaEval, we directly transform the guidelines we provide to human annotators into the prompt we provide to the judge LLMs, and we show the reasoning behind such choice here. We structure each prompt template into two components: a guideline and a list of demonstration examples. We interchange these components with those from AlpacaEval and compare the 4 resulting prompt templates using Llama-3.1-70B-Instruct with human-written responses on our development set as shown in Table 6. Table 6 shows that using a different set of examples (Prompt B), dropping the examples (Prompt C), or completely changing the prompt (Prompt D) negatively impacts agreement with human annotators compared to aligning the model prompt with the guidelines provided to human annotators (Prompt A). These results imply that **ensuring the consistency between the guidelines given to human annotators and the prompts for LLMs**

⁷<https://www.lepton.ai/pricing>.

	Guideline	Examples	Human Agreement Rate	r with Prompt D
A	HREF	HREF	68.4%	–
B	HREF	AlpacaEval	66.2%	0.98
C	HREF	None	65.6%	0.98
D	AlpacaEval	None	66.3%	0.95

Table 6: **Prompt Template Comparision.** A overview and comparison among four prompt templates on their guideline, examples, human agreement rate, and correlation with the prompt we use on the development set (Prompt A). Note that the prompt that AlpacaEval uses for LLM does not contain examples, and we adopt the examples they give to human annotators for Prompt A.

effectively improves the agreements between the human annotators and the judge LLMs, as they are encouraged to judge based on the same criteria.

With these four prompts, we evaluate 33 models on our development set and calculate the Pearson correlation on the resulting scores. As shown in Table 6, the strong correlation between our prompt (Prompt D) and AlpacaEval’s prompt (Prompt A) shows that our prompt reasonably aligns with the prompt used in prior work, and the strong correlation between our prompt and the alternative examples (Prompt B and C) shows that our prompt is not overly dependent or biased towards the specific examples that we select.

6 RELATED WORK

To evaluate the capability of post-trained LLMs in instruction-following, prior work has constructs benchmarks in several ways.

Instruction Source. Prior work have chosen to source instructions from real-world users. ChatbotArena (Chiang et al., 2024) is a benchmark that constantly collects instructions from the online community users by directly prompting for the user’s inputs. ArenaHard (Li et al., 2024) automatically curates instructions from those collected by Chatbot Area. These benchmarks possess sets of instructions that closely matches human’s common interest in terms of instruction categories, but they are also heavily skewed towards OpenQA and Generation as a result. Another widely recognized benchmark is AlpacaEval (Li et al., 2023; Dubois et al., 2024), which is consist of synthetically generated instructions generated using human-written template (Wang et al., 2022). WildBench (Lin et al., 2024) also collect instructions from the user in the wild. MT-Bench, with task-specific instructions created by human experts, is the most similar to our work, but it is restricted by the small size of the instruction size. Our work have collected instructions covering a wider range of tasks with a much larger evaluation set.

Evaluating Instruction-Following Models. When evaluating a LLM’ responses to a instruction, prior work either directly grade the response with a score, or perform a pairwise comparison with the response form another LLM (Zheng et al., 2023). Chatbot Arena (Chiang et al., 2024) prompts the same user who creates the instruction to also do a pairwise comparison between responses from two models (i.e., selecting the better response), and the benchmark’s evaluation results are treated as ground-truth and compared against by several other benchmarks (Li et al., 2023; Lin et al., 2024). However, such evaluation requires extensive human feedback, which is expensive to collect for majority of the benchmarks. LLM-as-a-judge, acting as a proxy for human annotators, has been widely adopted by many benchmarks in both single response grading and pairwise comparison. However, prior work use closed API models, which lacks transparency and consistency in their judgment. Our work is uses LLM-as-a-judge with public models and shows the benefits that brings.

Reference Guided Evaluation Comparing text embeddings to a human-written reference answer is widely used in traditional NLP tasks, especially summarization (Zhang et al., 2019; Lin, 2004; Papineni et al., 2002; Banerjee & Lavie, 2005), but it is less clear how to properly utilize the reference answer to evaluation more open-ended instruction-following. AlpacaEval (Li et al., 2023) has found that including model-generated responses in the prompt when using LLM-as-a-Judge is beneficial in following instruction related to math. Our work adopt an combination of comparing text embeddings to human-written responses and using human-written responses with LLM-as-a-

Judge depending on the task categories. Additionally, we provide insights about when and how these responses are beneficial.

Risk of Contamination When the test data of the prior work are public released, they are at a high risks of being contaminated. They can potentially lose the robustness and credibility in their evaluation when the evaluated LLMs are trained on the test data. To migrate such risk, WildBench (Lin et al., 2024) keeps their test set private and only release a development set. However, another implicit source of potential contamination remains unsolved when prompting the closed API models with the test data either when using them as the baseline model or as judges. Although not by directly training on the test data, LLMs can still gain knowledge about the them through either distillation from closed API models or training on synthetic data generated by these models (Dubey et al., 2024; Wang et al., 2022; Zhou et al., 2024; Peng et al., 2023; Xu et al., 2024; Zhao et al., 2024). HREF migrates such risks by using local public open-weight models for both the baseline model and the judge.

LIMITATIONS

Multi-turn Evaluation. Multi-turn evaluation is not the focus of work, and HREF is only suitable for single-turn instruction following evaluation. We suggest using benchmarks like WildBench for multi-turn evaluation.

Absolute Rating. Our work focuses solely on improving pairwise evaluation, which requires the use of a baseline model. We recognize that there might be circumstances where an independent absolute score can be useful, and we leave the topic of improving the accuracy of absolute rating of an LLM in instruction-following for future work.

REFERENCES

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.
- Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

- Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world’s first truly open instruction-tuned llm. *Company Blog of Databricks*, 2023.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpaca-eval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.
- Prolific First. Prolific first, 2014. URL <https://www.prolific.com/>.
- Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. Koala: A dialogue model for academic research. Blog post, April 2023. URL <https://bair.berkeley.edu/blog/2023/04/03/koala/>.
- Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpaca-eval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.
- Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*, 2024.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL <http://arxiv.org/abs/1907.11692>.
- Lester James V Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. ai feedback. *arXiv preprint arXiv:2410.19133*, 2024.
- 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. In *NeurIPS*, 2022.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

- Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. No robots. https://huggingface.co/datasets/HuggingFaceH4/no_robots, 2023.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=CfXh93NDgH>.
- Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Rank	Model	General Category								All
		Brn	OQA	CQA	Ext	Gen	Rew	Sum	Cls	
1	Mistral-Large-Instruct-2407	67.5	70.7	23.3	35.7	60.3	58.6	44.0	71.9	60.3
1	gpt-4-turbo-2024-04-09	59.2	74.1	26.7	21.4	54.4	64.3	44.0	43.8	56.0
1	gpt-4o-2024-05-13	58.3	82.8	23.3	7.1	54.7	55.7	34.0	37.5	55.3
1	gpt-4-1106-preview	58.3	75.9	20.0	21.4	53.7	47.1	44.0	50.0	54.4
5	Llama-3.1-70B-Instruct	52.5	65.5	63.3	14.3	49.8	55.7	66.0	53.1	53.7
6	Qwen2.5-72B-Instruct	63.3	72.4	30.0	21.4	49.1	47.1	50.0	53.1	53.1
6	Mistral-Small-Instruct-2409	59.2	72.4	26.7	21.4	46.3	50.0	44.0	50.0	50.8
8	Qwen1.5-110B-Chat	53.3	81.9	20.0	35.7	42.5	44.3	34.0	56.2	48.6
9	Llama-3.1-8B-Instruct	53.3	58.6	40.0	14.3	37.4	41.4	44.0	34.4	42.8
9	Qwen2-72B-Instruct	50.0	77.6	20.0	50.0	30.8	24.3	34.0	31.2	39.4
9	Yi-1.5-34B-Chat	51.7	60.3	30.0	57.1	32.0	32.9	20.0	46.9	38.8
9	Phi-3-medium-4k-instruct	29.2	77.6	40.0	35.7	28.0	21.4	24.0	46.9	35.3
13	gpt-3.5-turbo	19.2	81.0	40.0	7.1	22.0	10.0	12.0	43.8	29.2
13	tulu-2-dpo-70b	21.7	79.3	26.7	14.3	19.2	21.4	16.0	31.2	28.3
13	Mistral-7B-Instruct-v0.3	22.5	67.2	26.7	7.1	21.0	25.7	14.0	37.5	28.0
16	Llama-2-70b-chat-hf	20.0	60.3	53.3	21.4	17.5	21.4	14.0	46.9	26.2
16	WizardLM-13B-V1.2	16.7	69.0	26.7	14.3	14.5	7.1	8.0	34.4	22.3
16	tulu-v2.5-ppo-13b	45.0	34.5	0.0	7.1	14.3	17.1	14.0	15.6	20.9
16	vicuna-13b-v1.5	10.0	77.6	16.7	21.4	8.9	10.0	20.0	43.8	20.8
16	tulu-2-dpo-13b	6.7	65.5	10.0	21.4	12.4	18.6	10.0	31.2	19.9
21	Llama-2-13b-chat-hf	19.2	48.3	10.0	7.1	13.3	7.1	16.0	31.2	19.0
21	Llama-2-7b-chat-hf	17.5	41.4	26.7	21.4	12.4	10.0	10.0	21.9	17.7
21	tulu-2-dpo-7b	3.3	67.2	16.7	21.4	8.4	4.3	10.0	21.9	16.4
21	vicuna-7b-v1.5	3.3	75.9	23.3	14.3	5.6	4.3	4.0	25.0	16.0
21	gpt4all-13b-snoozy	1.7	77.6	23.3	7.1	3.3	2.9	0.0	12.5	14.0
26	OLMo-7B-SFT-hf	0.0	70.7	13.3	7.1	3.3	2.9	4.0	25.0	13.1
26	OLMo-7B-0724-Instruct-hf	15.0	46.6	0.0	0.0	7.0	5.7	4.0	9.4	12.9
26	dolly-v2-7b	1.7	67.2	16.7	7.1	1.9	1.4	0.0	6.2	11.3
26	koala-13B-HF	0.0	62.1	10.0	14.3	2.3	0.0	4.0	21.9	11.2
26	mpt-7b-chat	0.0	56.9	13.3	0.0	3.3	0.0	0.0	28.1	10.8
26	dolly-v2-12b	0.0	62.1	13.3	0.0	0.9	2.9	0.0	12.5	10.0
32	koala-7B-HF	0.0	55.2	16.7	7.1	0.0	0.0	0.0	18.8	8.8
32	oasst-sft-1-pythia-12b	0.0	63.8	0.0	0.0	0.0	0.0	0.0	0.0	8.6

Table 7: **expected win rates of all 33 starting models evaluated on the validation set of HREF.** All numbers are in %. (*i*) indicates the ranking. Brn \rightarrow Brainstorm; OQA \rightarrow Open QA; CQA \rightarrow Closed QA; Ext \rightarrow Extraction; Gen \rightarrow Generation; Rew \rightarrow Rewriting; Sum \rightarrow Summarization; Cls \rightarrow Classification; FC \rightarrow Fact Checking / Attributed QA; MDS \rightarrow Multi-Document Synthesis; RND \rightarrow Reasoning Over Numerical Data;

A FULL VALIDATION SET RESULTS

See the expected win rate of all of the starting 29 models evaluated on validation set of HREF in Table 7.

B FORMULATION

We formally define the research problem and our proposed evaluation method, HREF.

B.1 PROBLEM DEFINITION

We denote HREF’s evaluation dataset as D , with each element being (in, o_B, o_H) , denoting the instruction, the baseline model response, and the human written reference response respectively.

Given a target LLM \mathcal{T} , HREF aims to estimate the rate that human would consider the responses from \mathcal{T} are at least as good as the baseline model \mathcal{B} in following instructions, which we formally

Model Family	Model Name
Dolly	dolly-v2-12b dolly-v2-7b
Koala	koala-7B-HF koala-13B-HF
Llama-2	Llama-2-7b-chat-hf Llama-2-13b-chat-hf Llama-2-70b-chat-hf
Llama-3	Meta-Llama-3-8B-Instruct Meta-Llama-3.1-8B-Instruct Meta-Llama-3-70B-Instruct Meta-Llama-3.1-70B-Instruct
Mistral	Mistral-7B-Instruct-v0.3 Mistral-Small-Instruct-2409 Mistral-Large-Instruct-2407
MPT	mpt-7b-chat
OpenAssistant	oasst-sft-1-pythia-12b
OLMo	OLMo-7B-SFT-hf OLMo-7B-0724-Instruct-hf
Phi	Phi-3-medium-4k-instruct
Qwen	Qwen2-72B-Instruct Qwen1.5-110B-Chat
Vicuna	vicuna-7b-v1.5 vicuna-13b-v1.5
WizardLM	WizardLM-13B-V1.2
Yi	Yi-1.5-34B-Chat
GPT-3	gpt-3.5-turbo
GPT-4	gpt-4-1106-preview gpt-4-turbo-2024-04-09 gpt-4o-2024-05-13
O-1	o1-mini o1-preview

Table 8: Full list of model family and names that we use to construct the model pool where we sample the responses for the human agreement set.

defined as:

$$expectedwinrate(\mathcal{T}, \mathcal{B}) = \frac{1}{|D|} \sum_{(in, o_{\mathcal{B}}, o_{\mathcal{H}}) \in D} p(in, o_{\mathcal{T}}, o_{\mathcal{B}}, o_{\mathcal{H}})$$

where $o_{\mathcal{T}} = \mathcal{T}(in)$ represents the response of \mathcal{T} given the instruction as the input, and $p(in, o_{\mathcal{T}}, o_{\mathcal{B}}, o_{\mathcal{H}})$ is a binary function representing the pairwise preference (0 if the baseline model is preferred and 1 otherwise).

B.2 LLM-AS-A-JUDGE WITH OPTIONAL HUMAN REFERENCE

We propose the evaluation method, LLM-as-a-judge with human reference, as one of the methods to estimate $p(in, o_{\mathcal{T}}, o_{\mathcal{B}})$. Specifically, we embed $in, o_{\mathcal{T}}, o_{\mathcal{B}}, o_{\mathcal{H}}$ into a prompt template as the input to a separate judge model \mathcal{J} formally:

$$p(in, o_{\mathcal{T}}, o_{\mathcal{B}}, o_{\mathcal{H}}) = \mathcal{J}(in, o_{\mathcal{T}}, o_{\mathcal{B}}, o_{\mathcal{H}})$$

Note that when not using a reference, the definition is the same except that $o_{\mathcal{H}}$ will not be an input to \mathcal{J} .

B.3 ROBERTA EMBEDDING: COMPARING TEXT EMBEDDINGS WITH HUMAN REFERENCE

We also propose to compare the cosine similarity between the text embeddings of $o_{\mathcal{T}}$ and $o_{\mathcal{H}}$ against $o_{\mathcal{B}}$ and $o_{\mathcal{H}}$. Formally,

$$p(in, o_{\mathcal{T}}, o_{\mathcal{B}}, o_{\mathcal{H}}) = \begin{cases} 0 & \text{if } \text{sim}(o_{\mathcal{T}}, o_{\mathcal{H}}) < \text{sim}(o_{\mathcal{B}}, o_{\mathcal{H}}) \\ 1 & \text{otherwise.} \end{cases}$$

where

$$\text{sim}(o_{\mathcal{X}}, o_{\mathcal{Y}}) = \frac{\text{Embed}(o_{\mathcal{X}}) \cdot \text{Embed}(o_{\mathcal{Y}})}{\|\text{Embed}(o_{\mathcal{X}})\| \|\text{Embed}(o_{\mathcal{Y}})\|}$$

with $\text{Embed}(o_{\mathcal{Y}})$ represents some embeddings of $o_{\mathcal{Y}}$.

C LLM-AS-A-JUDGE PROMPT AND PARSING

Figure 9 shows the prompt template for LLM-as-a-Judge where we embed the instruction, the target and reference model responses, and the human written reference into to construct the final prompt for the judge LLM as mentioned in Section 2.2. Figure 10 shows the one without including human reference. We design the template to match the guideline we give to human annotators in Section D.2, resulting in a 2-shot prompting. Note that we randomly swap the target and reference model response to avoid potential label bias.

During parsing, we strip and normalized the generated output, and map the exact match of "a" into 0, and "b" or "tie" into 1. We optionally reverse the preference if the embedded responses are swapped. Note that when the parsing fails, we ignore the current data point in the calculate of the expected win rates.

Figure 11, Figure 12, and Figure 13 shows the other modified version of prompt templates that we compare our prompt template against in Section 5.2.

D HUMAN AGREEMENT ANALYSIS DETAILS

D.1 MODEL POOL

The full model pool from which we sample the responses to construct our human agreement dataset in Section 2.1.1 and Section 4.1 includes Dolly (Conover et al., 2023), Koala (Geng et al., 2023), Llama-2 (Touvron et al., 2023), Llama-3.1 (Dubey et al., 2024), Mistral (Jiang et al., 2023), MPT (Dubey et al., 2024), Pythia (Biderman et al., 2023), OLMo (Groeneveld et al., 2024), Phi (Abdin et al., 2024), Qwen (Bai et al., 2023), Vicuna (Chiang et al., 2023), WizardLM (Xu et al., 2024), Yi (Young et al., 2024), GPT-3 (Brown, 2020), GPT-4 (Achiam et al., 2023), and O1⁸. See Table 8 for the full list of model names.

D.2 HUMAN ANNOTATION GUIDELINE

Figure 14 shows the full guideline we provide to the annotators during preference collection. We adopt the guideline from Li et al. (2023) with some modifications.

D.3 ANNOTATION WEBSITE

See Figure 15 for an overview of the website that we direct our human annotator to. We ask them to spend time in getting familiar with website before annotations.

D.4 LEAVE-ONE-OUT AGREEMENT RATE CALCULATION

Algorithm 1 provides a detailed overview of the metric Leave-One-Out Agreement Rate used in human agreement analysis (Section 2).

⁸<https://openai.com/o1/>

```

<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful assistant that helps us rate an AI model's responses to instructions.<|eot_id|><|start_header_id|>user<|end_header_id|>
Decide which response from the AI system following the instruction is better, considering the following questions:

1. Does the response precisely follow the instruction? For example, a response that includes unrelated information or does not fulfill the task is not precisely following the instruction. Compare each response with the provided human response to decide if a response faithfully follows the instruction, especially when the instruction asks for expected word count or format.
2. Is the response helpful? For example, if the instruction asks for a recipe for healthy food, and the response is a useful recipe, then you can consider it helpful.
3. Is the response language natural? For example, AI responses are often verbose or repetitive, which is not natural. Compare with the provided human response to decide whether a response is natural.
4. Is the response factual/accurate? AI responses often make up new information. For example, if the response claims that Donald Trump is the current U.S. president, then you should consider it inaccurate. Compare with the provided human response to verify whether a response is factual and accurate, especially with numbers.
5. Based on your aesthetics, which one do you prefer? For example, you might prefer one poem over another poem.

Select the response A or B that you prefer, or select tie if the two responses are similarly good or bad. Your answer should ONLY contain: A or B or tie. Here are two examples:

# Example:
## Instruction:
Paraphrase the following question: "What is the capital of France?"

## Response A:
The capital of France is Paris.

## Response B:
Can you tell me the name of the city that serves as the capital of France?

## Human Response:
What's the name of the France's capital

In this example, B paraphrases the question as asked by the instruction. In contrast, A does not follow instruction as it answers the question instead. Human Response also paraphrases the question in a different way.
## Which is best, A, B, or tie?
B

# Example:
## Instruction:
Change the first verb to a synonym:
I love surfing

## Response A:
I hate surfing

## Response B:
and am planning on going surfing tomorrow morning.

## Human Response:
I like surfing

In this example, Neither output is correct. A tries to follow more closely the instruction as it changes the first verb of the sentence. Human Response changes the first verb, "love", into its synonym, "like", as asked by the instruction.
## Which is best, A, B, or tie?
B

# Example:
## Instruction:
Change the first verb to a synonym:
I love surfing

## Response A:
I hate surfing

## Response B:
I don't like surfing

## Human Response:
I like surfing

In this example, neither output is correct and the responses are very similar. Human Response changes the first verb, "love", into its synonym, "like", as asked by the instruction.
## Which is best, A, B, or tie?
tie

Now is the real task, just select among: A, B, or tie.
# Task:
## Instruction:
{instruction}

## Response A:
{output_1}

## Response B:
{output_2}

## Human Response:
{output_human}

## Which is best, A, B, or tie?<|eot_id|><|start_header_id|>assistant<|end_header_id|>

```

Figure 9: Prompt Template For LLM-as-a-Judge with Human Response. The prompt template we use to prompt our judge model Llama-3.1-70B-Instruct to give the preference between two model responses along with human reference. Note that we intentionally transform the guidelines we give to the human annotators into this prompt to maximize the fairness in comparison.

```

<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful assistant that helps us rate an AI model's responses to
instructions.<|eot_id|><|start_header_id|>user<|end_header_id|>
Decide which response from the AI system following the instruction is better, considering the following questions:

1. Does the response precisely follow the instruction? For example, a response that includes unrelated information or
does not fulfill the task is not precisely following the instruction.
2. Is the response helpful? For example, if the instruction asks for a recipe for healthy food, and the response is a useful
recipe, then you can consider it helpful.
3. Is the response language natural? For example, AI responses are often verbose or repetitive, which is not natural.
4. Is the response factual/accurate? AI responses often make up new information. For example, if the response claims
that Donald Trump is the current U.S. president, then you should consider it inaccurate.
5. Based on your aesthetics, which one do you prefer? For example, you might prefer one poem over another poem.

Select the response A or B that you prefer, or select tie if the two responses are similarly good or bad. Your answer
should ONLY contain: A or B or tie. Here are two examples:

# Example:
## Instruction:
What type of Rifle is an AK47?

## Response A:
An AK47 is a type of assault rifle.

## Response B:
The AK47 is a type of rifle that is commonly associated with Soviet and Russian military forces. It is a semi-automatic
assault rifle that is chambered in 7.62x39mm and is considered to be one of the most popular and widely used firearms in
the world. The AK47 is known for its durability, reliability, and accuracy, and it is able to fire up to 45 rounds per minute. It
is also popular among civilians due to its low cost and ease of use. The AK47 is an iconic weapon that has been featured
in numerous movies, video games, and other pop culture references.

Here A directly and concisely answers the type of AK47 as asked by the instruction. In contrast, B is too verbose and
talks about AK47's background, advantages, and influence, which are unrelated to the instruction.
## Which is best, A, B, or tie?
A

# Example:
## Instruction:
Given a table of data, compute the average
Table of data:
1, 3, 5, 6, 9

## Response A:
The average is 5.

## Response B:
4.75

Here B is more accurate than A because 4.75 from B is closer to the correct average 4.8 than 5 from A.
## Which is best, A, B, or tie?
B

Now is the real task, just select among: A, B, or tie.
# Task:
## Instruction:
{instruction}

## Response A:
{output_1}

## Response B:
{output_2}

## Which is best, A, B, or tie?<|eot_id|><|start_header_id|>assistant<|end_header_id|>

```

Figure 10: **Prompt Template For LLM-as-a-Judge.** The prompt template we use to prompt our judge model Llama-3.1-70B-Instruct to give the preference between two model responses without a reference. Note that we intentionally transform the guidelines we give to the human annotators into this prompt to maximize the fairness in comparison.


```

<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful assistant that helps us rate an AI model's responses to instructions.<|eot_id|><|start_header_id|>user<|end_header_id|>
Decide which response from the AI system following the instruction is better, considering the following questions:

1. Does the response precisely follow the instruction? For example, a response that includes unrelated information or does not fulfill the task is not precisely following the instruction. Compare each response with the provided human response to decide if a response faithfully follows the instruction, especially when the instruction asks for expected word count or format.
2. Is the response helpful? For example, if the instruction asks for a recipe for healthy food, and the response is a useful recipe, then you can consider it helpful.
3. Is the response language natural? For example, AI responses are often verbose or repetitive, which is not natural. Compare with the provided human response to decide whether a response is natural.
4. Is the response factual/accurate? AI responses often make up new information. For example, if the response claims that Donald Trump is the current U.S. president, then you should consider it inaccurate. Compare with the provided human response to verify whether a response is factual and accurate, especially with numbers.
5. Based on your aesthetics, which one do you prefer? For example, you might prefer one poem over another poem.

Select the response A or B that you prefer, or select tie if the two responses are similarly good or bad. Your answer should ONLY contain: A or B or tie. Here are two examples:

# Example:
## Instruction:
Paraphrase the following question: "What is the capital of France?"

## Response A:
The capital of France is Paris.

## Response B:
Can you tell me the name of the city that serves as the capital of France?

## Human Response:
What's the name of the France's capital

In this example, B paraphrases the question as asked by the instruction. In contrast, A does not follow instruction as it answers the question instead. Human Response also paraphrases the question in a different way.
## Which is best, A, B, or tie?
B

# Example:
## Instruction:
Change the first verb to a synonym:
I love surfing

## Response A:
I hate surfing

## Response B:
and am planning on going surfing tomorrow morning.

## Human Response:
I like surfing

In this example, Neither output is correct. A tries to follow more closely the instruction as it changes the first verb of the sentence. Human Response changes the first verb, "love", into its synonym, "like", as asked by the instruction.
## Which is best, A, B, or tie?
B

# Example:
## Instruction:
Change the first verb to a synonym:
I love surfing

## Response A:
I hate surfing

## Response B:
I don't like surfing

## Human Response:
I like surfing

In this example, neither output is correct and the responses are very similar. Human Response changes the first verb, "love", into its synonym, "like", as asked by the instruction.
## Which is best, A, B, or tie?
tie

Now is the real task, just select among: A, B, or tie.
# Task:
## Instruction:
{instruction}

## Response A:
{output_1}

## Response B:
{output_2}

## Human Response:
{output_human}

## Which is best, A, B, or tie?<|eot_id|><|start_header_id|>assistant<|end_header_id|>

```

Figure 11: **Prompt Template with demonstration examples replaced.** A modified version of the prompt template we use to prompt our judge model Llama-3.1-70B-Instruct to give the preference between two model responses with a reference. We replace the demonstrations examples with the ones adopted from the examples given to the human annotators by AlpacaEval.

```

<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful assistant that helps us rate an AI model's responses to
instructions.<|eot_id|><|start_header_id|>user<|end_header_id|>
Decide which response from the AI system following the instruction is better, considering the following questions:

1. Does the response precisely follow the instruction? For example, a response that includes unrelated information or does not fulfill
the task is not precisely following the instruction. Compare each response with the provided human response to decide if a
response faithfully follows the instruction, especially when the instruction asks for expected word count or format.
2. Is the response helpful? For example, if the instruction asks for a recipe for healthy food, and the response is a useful recipe,
then you can consider it helpful.
3. Is the response language natural? For example, AI responses are often verbose or repetitive, which is not natural. Compare with
the provided human response to decide whether a response is natural.
4. Is the response factual/accurate? AI responses often make up new information. For example, if the response claims that Donald
Trump is the current U.S. president, then you should consider it inaccurate. Compare with the provided human response to verify
whether a response is factual and accurate, especially with numbers.
5. Based on your aesthetics, which one do you prefer? For example, you might prefer one poem over another poem.

Select the response A or B that you prefer, or select tie if the two responses are similarly good or bad. Your answer should ONLY
contain: A or B or tie.

Now is the real task, just select among: A, B, or tie.
# Task:
## Instruction:
{instruction}

## Response A:
{output_1}

## Response B:
{output_2}

## Human Response:
{output_human}

## Which is best, A, B, or tie?<|eot_id|><|start_header_id|>assistant<|end_header_id|>

```

Figure 12: **Prompt Template with demonstration examples removed.** A modified version of the prompt template we use to prompt our judge model Llama-3.1-70B-Instruct to give the preference between two model responses with a reference. We removes the demonstration examples.

```

<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful assistant, that ranks models by the quality of their answers.<|eot_id|><|start_header_id|>user<|end_header_id|>
I want you to create a leaderboard of different of large-language models. To do so, I will give you the instructions (prompts) given to
the models, the responses of two models, and a reference response written by human expert. Please rank the models based on
which responses would be preferred by humans. All inputs and outputs should be python dictionaries.

Here is the prompt:
{
  "instruction": ""#{instruction}""
}

Here is the human-written response to be used as your reference:
{
  "reference": ""#{output_human}""
}

Here are the outputs of the models:
[
  {
    "model": "model_1",
    "answer": ""#{output_1}""
  },
  {
    "model": "model_2",
    "answer": ""#{output_2}""
  }
]

Now please rank the models by the quality of their answers, so that the model with rank 1 has the best output. Then return a list of
the model names and ranks, i.e., produce the following output:
[
  {"model": <model-name>, 'rank': <model-rank>},
  {"model": <model-name>, 'rank': <model-rank>}
]

Your response must be a valid Python dictionary and should contain nothing else because we will directly execute it in Python.
Please provide the ranking that the majority of humans would give.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

```

Figure 13: **Prompt Template from AlpacaEval.** A modified version of the prompt template we use to prompt our judge model Llama-3.1-70B-Instruct to give the preference between two model responses with a reference. We adopt the exactly prompt that AlpacaEval uses for their judge LLMs.

Welcome! Thank you for participating in this study! We are researchers at the Allen Institute for AI (AI2) and the University of Washington working on building language models that you can interact with (i.e., chatbots like ChatGPT). To improve this technology, we would like to understand what kinds of responses from chatbots you prefer. The purpose of this platform is to collect these preferences. Please read the information below before you start.

Please take your time to read the user requests and the chatbot responses carefully before making your decisions. Some requests and responses are much longer than others. So we expect you will spend 5-10 minutes (or even slightly more) on the longer ones, and less than a minute on the shorter ones. Note that we are looking for quality annotations and you will be paid by the time you spend on the annotation, rather than the number of annotations you complete. So please do not rush through the annotations.

Your task is to decide which response is better (A and B). There are several dimensions that you can think along. Consider the following questions:

1. Does the response precisely follow the instruction? For example, a response that includes unrelated information or does not fulfill the task is not precisely following the instruction.
2. Is the response helpful? For example, if the instruction asks for a recipe for healthy food, and the response is a useful recipe, then you can consider it helpful.
3. Is the response language natural? For example, AI responses are often verbose or repetitive, which is not natural.
4. Is the response factual/accurate? AI responses often make up new information. For example, if the response claims that Donald Trump is the current U.S. president, then you should consider it inaccurate.
5. Based on your aesthetics, which one do you prefer? For example, you might prefer one poem over another poem.

To help you understand the annotation task, we provide two examples below:

1. What type of Rifle is an AK47?

Response A: An AK47 is a type of assault rifle.

Response B: The AK47 is a type of rifle that is commonly associated with Soviet and Russian military forces. It is a semi-automatic assault rifle that is chambered in 7.62x39mm and is considered to be one of the most popular and widely used firearms in the world. The AK47 is known for its durability, reliability, and accuracy, and it is able to fire up to 45 rounds per minute. It is also popular among civilians due to its low cost and ease of use. The AK47 is an iconic weapon that has been featured in numerous movies, video games, and other pop culture references.

In this example, response 1 directly and concisely answers the type of AK47 as asked by the instruction. In contrast, response 2 is too verbose and talks about AK47's background, advantages, and influence, which are unrelated to the instruction. You should select 1.

2. Given a table of data, compute the average. Table of data: 1, 3, 5, 6, 9.

Response A: The average is 5.

Response B: 4.75

In this example, response 1 is more accurate than response 2 because 4.75 from response 1 is closer to the correct average 4.8 than 5 from response 2. You should select 2.

Figure 14: **Guideline for Human Annotator.** The guideline we provide for the human annotators. A modified version from Li et al. (2023).

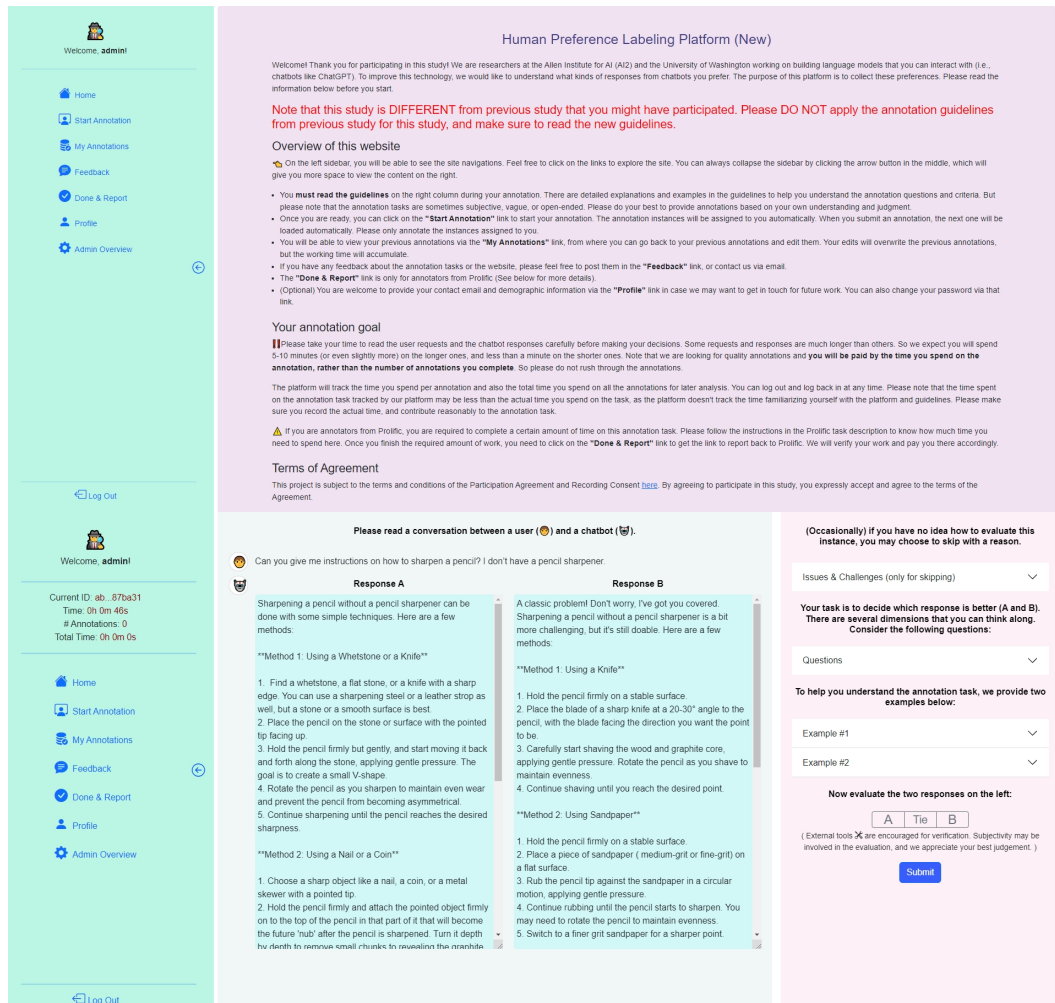


Figure 15: **Annotation Website.** The main pages of the website we build for collecting human annotations. The website framework is adopted from Miranda et al. (2024).

Algorithm 1 Algorithms to calculate Leave-One-Out (LOO) agreement rate either within the set of annotation of annotations (inner) or against a evaluator prediction (outer).

```
function GET_MODE(annotations)
  modes  $\leftarrow$  list of annotations with highest occurrence frequency
  if length of modes > 1 then
    return randomly chosen annotation from modes
  else
    return modes[0]
  end if
end function

function LEAVE_ONE_OUT_AGREEMENT_INNER(annotations)
  n_annotations  $\leftarrow$  length of annotations
  n_correct_predictions  $\leftarrow$  0
  for each  $i$  from 1 to n_annotations do
    target_annotations  $\leftarrow$  annotations without  $i$ -th element
    mode  $\leftarrow$  get_mode(target_annotations)
    if annotations[ $i$ ] = mode then
      correct_predictions  $\leftarrow$  correct_predictions + 1
    end if
  end for
  return n_correct_predictions / n_annotations
end function

function LEAVE_ONE_OUT_AGREEMENT_OUTER(annotations, prediction)
  n_annotations  $\leftarrow$  length of annotations
  n_correct_predictions  $\leftarrow$  0
  for each  $i$  from 1 to n_annotations do
    target_annotations  $\leftarrow$  annotations without  $i$ -th element
    mode  $\leftarrow$  get_mode(target_annotations)
    if prediction = mode then
      correct_predictions  $\leftarrow$  correct_predictions + 1
    end if
  end for
  return n_correct_predictions / n_annotations
end function
```
