HumanRankEval: Automatic Evaluation of LMs as Conversational Assistants

Anonymous ACL submission

Abstract

Language models (LMs) as conversational assistants recently became popular tools that help people accomplish a variety of tasks. These typically result from adapting LMs pretrained on general domain text sequences through further instruction-tuning and possibly preference optimisation methods. The evaluation of such LMs would ideally be performed using human judgement, however, this is not scalable. On the other hand, automatic evaluation featuring auxiliary LMs as judges and/or knowledge-based tasks is scalable but struggles with assessing conversational ability and adherence to instructions. To help accelerate the development of LMs as conversational assistants, we propose a novel automatic evaluation task: HumanRankEval (HRE). It consists of a large-scale, diverse and high-quality set of questions, each with several answers authored and scored by humans. To perform evaluation, HRE ranks these answers based on their log-likelihood under the LM's distribution, and subsequently calculates their correlation with the corresponding human rankings. We support HRE's efficacy by investigating how efficiently it separates pretrained and instruction-tuned LMs of various sizes. We show that HRE correlates well with human judgements and is particularly responsive to model changes following instruction-tuning.

1 Introduction

The evaluation of Language Models (LMs) is a challenging problem and a prolific research subject. Many benchmarks have recently been proposed aiming to evaluate the general capabilities of LMs, covering both automatic and human evaluation (Chang et al., 2023). Evaluating LMs' capabilities as conversational assistants, i.e. its adherence to human instructions, is particularly challenging as model inputs and outputs are more unstructured and open-ended. Ideally, human judgement should be employed to evaluate such open-ended outputs, typically either through interactive conversation with LMs (Zheng et al., 2023b) or by presenting participants with outputs from different LMs and collecting their preferences (van der Lee et al., 2021). As this approach is time-consuming and does not scale well, previous work proposed to substitute human judgement with auxiliary large LMs. However, these efforts have so far only been applied with proprietary models, e.g. GPT-4 (Zheng et al., 2023b; Dubois et al., 2023), and with mixed results (Chiang and Lee, 2023). On the other hand, conventional automatic evaluation of LMs on knowledge-based tasks such as multiple-choice question-answering (QA) (Zellers et al., 2019; Clark et al., 2018; Lin et al., 2021), can

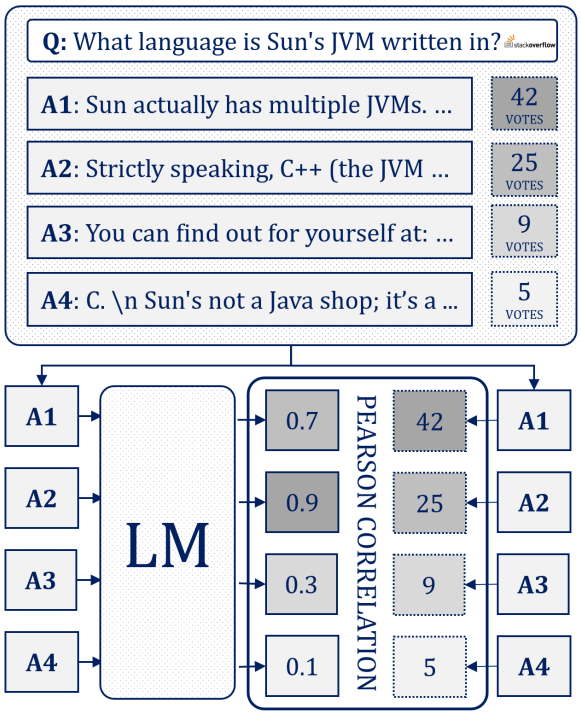

Figure 1: Overview of HumanRankEval: given a question with multiple answers, we correlate human scores of each answer with the log-likelihoods of the LM. The unabridged answers can be found in Figure 2.

Evaluation Type	Ground Truth	Metric(s)
Human	Human judgement	Elo, Win-Rate, Other
Knowledge-based	Human-authored text (e.g. exams, tests)	Accuracy-like
LM-as-a-judge	LMs (e.g. ChatGPT, GPT-4)	Elo, Win Rate, Other
HumanRankEval	**Human-authored text (Ranked QA pairs)**	**Correlation**

Table 1: Human evaluation versus relevant automatic evaluation types and their key features.

measure specific task performance in a scalable manner, but is not necessarily indicative of how an LM would perform these tasks in an open-ended conversational setting (Tunstall et al., 2023).

To this end, we introduce HumanRankEval (HRE), an automatic evaluation task for LMs as conversational assistants that comprises a novel dataset and metric. The core idea behind HRE is to measure an LM's alignment with human preferences (HP). Intuitively, given a question (Q) with multiple available responses ($A1, \ldots, A4$), HRE measures how well an LM's "preference" ranking over those answers aligns with those of humans (see Figure 1). We approximate HP by collecting a set of questions and rated answers from StackOverflow and StackExchange. The HRE dataset covers a diverse collection of 14 topics, each containing 500 information-seeking questions paired with the top-4 answers rated (on average) by 100+ domain experts. To estimate the "preferences" of an LM, we obtain the log-likelihood of each answer under the model's distribution. The HRE metric is calculated as the correlation of the LM's rankings against the corresponding human rankings. We should note that we do not consider HRE as a replacement for human judgement, but rather propose its usage for fast iterations during development.

We support HRE's efficacy by investigating how effectively it separates pretrained and instruction-tuned LMs of various sizes. We then compare our results against those of other evaluation frameworks, showing that HRE correlates well with human evaluation of LMs and provides unique insights. Specifically, relative to OpenLLM, a highly popular automatic evaluation leaderboard (Beeching et al., 2023), HRE is able to more effectively differentiate pretrained and instruction-tuned LMs. Our contributions are threefold: 1) we create a large-scale, high-quality, diverse QA dataset to capture/approximate human preferences, 2) introduce an efficient automatic method to evaluate LMs as conversational assistants by measuring the correlation of LM and human preferences, and 3) perform analysis that shows HRE correlates well with human judgement and provides unique insights.[1]

2 Related Work

The evaluation of LMs is a highly active research topic, exemplified by a recent survey (Chang et al., 2023) that tracks over 250 papers, with over 100 of those published in just the last 12 months. There are additional surveys focused on alignment (Wang et al., 2023c), trustworthiness (Liu et al., 2023b), morals (Scherrer et al., 2023) and fairness (Li et al., 2023c) as well as multiple benchmarks with leaderboards covering a wide variety of LM behaviours (Zhong et al., 2023; Wang et al., 2023a; Srivastava et al., 2022; Chia et al., 2023; Ye et al., 2023; Liang et al., 2022; Dubois et al., 2023; Liu et al., 2023a; Yuan et al., 2023; Sun et al., 2023; Ziyu et al., 2023), to list just a few. Therefore, we focus on methods relevant to evaluating LMs as conversational assistants to differentiate from prior work.

2.1 Human Evaluation

Due to the open-ended nature of the output, human judgement is considered the gold standard for evaluating LMs as conversational assistants (Ji et al., 2023; Song et al., 2023; Rafailov et al., 2023), however, such evaluation is costly and can be biased (Wu and Aji, 2023). These issues are more prevalent in crowd-sourcing settings where participants need to be vetted to ensure their expertise and reliability, especially given that the motivations at play (e.g. to complete as many assessments as fast as possible) may run counter to the purposes of the evaluation (van der Lee et al., 2021). Evaluation is often set up as an interactive dialog with each LM where participants are asked to rate its performance in various metrics (van der Lee et al., 2021; Ji et al., 2022) or by contrasting multiple LM outputs (produced by the same input/prompt) and voting for the one that is preferred (Bai et al., 2022). The latter preferences can be converted into Elo ratings to

[1] Data and code will be released on acceptance.

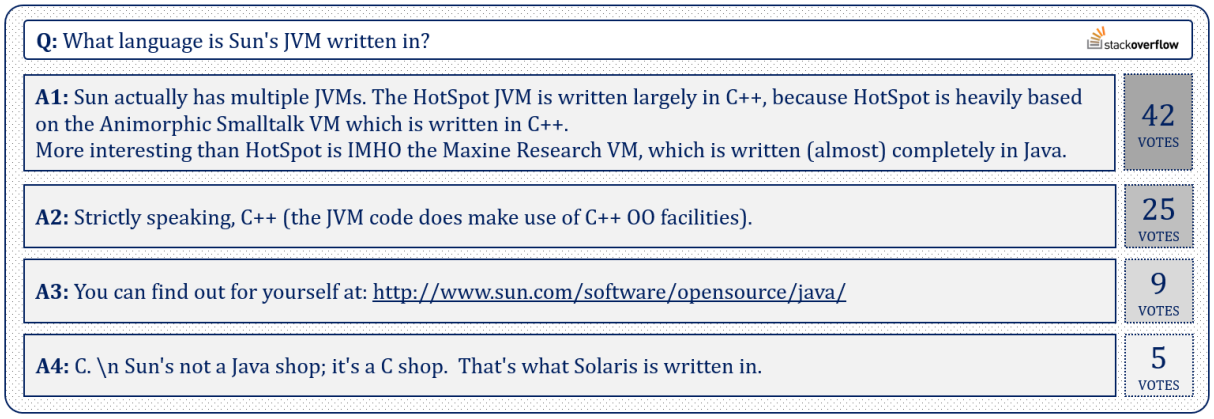

Figure 2: HumanRankEval example from StackOverflow (Java topic).

obtain LM rankings (Zheng et al., 2023b; Wu and Aji, 2023). A public leaderboard that maintains such rankings is Chatbot Arena[2]. Its game-like environment encourages users to guess the identity of two LMs at the end of an anonymous interaction. These multi-turn conversations are unstructured and depend on the interests of participants.

2.2 Automatic Evaluation

2.2.1 Knowledge-based Evaluation

A subset of automatic evaluation focuses on knowledge-based tasks with strictly-defined inputs and outputs, to enable the easy application of automatic metrics and measure performance. This is in contrast to how conversation assistants operate, where input and output is more open-ended. For LMs as conversational assistants, the focus of knowledge-based evaluation is to measure the general capabilities of the model, rather than particular performance on downstream tasks. As such, evaluation is usually applied through zero-shot or few-shot/prompt settings, without fine-tuning LMs on task-specific data. Examples include multiple-choice QA (Liu et al., 2020), code generation (Chen et al., 2021), Tool/API usage (Liu et al., 2023a), general and advanced knowledge tests (Hendrycks et al., 2020; Liu et al., 2020; Cobbe et al., 2021; Zellers et al., 2019; Clark et al., 2018; Lin et al., 2021), complex logical reasoning (Cobbe et al., 2021), school admission tests (Zhong et al., 2023) and fine-grained "skill sets" evaluation (Ye et al., 2023). Individual benchmarks are often aggregated into high-profile public rankings such as the OpenLLM Leaderboard[3], which we reference throughout. Importantly, such tasks and metrics do not accurately estimate how an LM may perform on them within a conversational context, as they were not designed for this purpose.

2.2.2 LM-as-a-judge

A faster alternative to human evaluation has been proposed recently, i.e. to use LMs as judges (typically larger than the LMs being judged). The most popular examples include MT-Bench (Zheng et al., 2023b) and AlpacaEval[4] (Dubois et al., 2023). MT-Bench prompts GPT-4 to score the quality of the candidate LM on a 10-point scale over 80 two-turn conversations. AlpacaEval instructs GPT-4 to vote whether the output of the candidate LM or ChatGPT is better, resulting in a win-rate % against GPT-3.5, using 805 manually selected prompts. However, these models are known to have biases (Wu and Aji, 2023) and their appropriateness for LM evaluation is frequently being questioned (Aiyappa et al., 2023; Chiang and Lee, 2023; Li et al., 2023b). At the time of this writing, such approaches have only been explored in connection with proprietary models, with concerns regarding data privacy, API costs and a lack of control making them less amenable to open research.

2.2.3 A Note on Multi-Turn Evaluation

Even though the goal is to evaluate LMs as conversational assistants, most automatic evaluation methods (including HRE) are limited to evaluating single turn conversations. This is due to the difficulty of integrating LM interaction within an auto-

[2] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

[3] https://hf.co/spaces/HuggingFaceH4/open_LM_leaderboard

[4] https://tatsu-lab.github.io/alpaca_eval/

matic task. MT-Bench contains two-turn prompts, but assumes no interaction either, with the second-turn prompt attending on a reference answer.

3 HumanRankEval

We now introduce HumanRankEval, an automatic evaluation task (comprising a novel dataset and metric) for LMs as conversational assistants. As mentioned earlier, the core idea behind HRE is to evaluate LMs by observing how an LM's "preference" ranking (derived from the model's log-probabilities over several answers) aligns with human-obtained rankings. To achieve this, we gather open-ended, information-seeking questions from popular online communities to capture HP. Each question comes with several answers ranked by domain enthusiasts (see Figure 2 for an example), indicating the order of responses (most to least preferable). Our data sources consist of StackExchange and StackOverflow. As both contain a plethora of topics, some of which may be considered subjective, we endeavoured to select the more objective/quantitative topics that we would expect to have a high degree of consensus among users, i.e. most people would agree on "good" answers.

3.1 StackExchange

StackExchange is a trusted site for communities of experts answering questions on various subjects. The data dumps were sourced from the Internet Archive[5] and processed with Eleuther's scripts[6]. Due to limited data availability (after filtering for quality), we set the number of questions to 500 for each topic for a uniform distribution over all domains. We selected questions from popular discussion topics: Unix-based OS, English Language, Physics, LaTeX, Software Engineering, Maths and Statistics. We also created three "mixed topics" (500 questions each) from somewhat less popular subsets that did not individually yield enough questions after filtering: *CS+DB* (CodeReview, Computer Science, Data Science and Databases), *App+Andr* (Apple and Android) and *Lang+Sci* (Latin, Chinese, French, German, Japanese, Spanish plus Engineering, Chemistry, Biology, Earth Science and Astronomy).

[5] https://archive.org/download/stackexchange

[6] https://github.com/EleutherAI/stackexchange-dataset

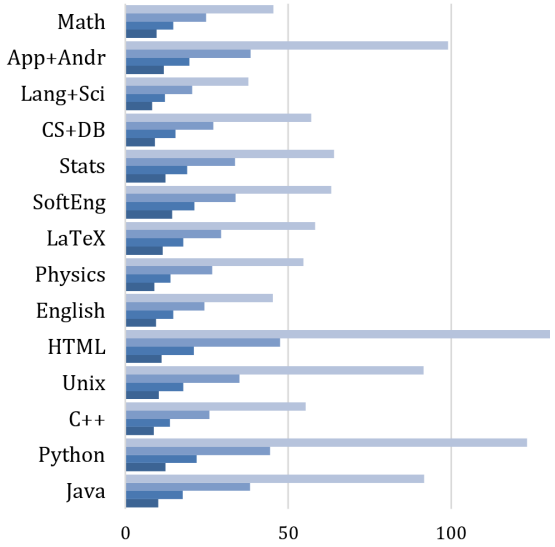

Figure 3: Average votes per answer/topic. Each answer has approximately double the votes of the next answer. More details can be found in Figure 11 in the Appendix.

3.2 StackOverflow

StackOverflow is a highly popular website and a leading community of people who contribute their expertise on a plethora of technical topics. In order to prevent HRE from being dominated by programming languages, i.e provide a balance against the more general topics of StackExchange, we selected questions from each of the following popular topics: Python, Java, HTML (includes CSS, JavaScript) and C++. The dataset was contributed by Li et al. (2023d)[7]. Once again, we set the number of questions to 500 per topic for a balanced dataset.

3.3 Data Filtering

HRE includes QA pairs that meet the following criteria: i) the question has at least 4 answers (keep the top 4) to ensure a meaningful ranking, ii) the answers are scored by at least 40 people (10 per answer, on average) to ensure a minimum annotator pool size for each question thus giving a more reliable agreement on the rankings, iii) each answer has at least 5 votes to ensure a minimum annotator pool for each answer hence avoiding low quality responses, iv) the maximum length of each QA pair is 4,000 characters to evaluate models with shorter context windows without truncation, v) answers with identical votes are discarded (we keep the first answer with N votes) and vi) duplicate QA pairs are discarded to ensure unique QA pairs

[7] https://huggingface.co/datasets/suriyagunasekar/stackoverflow-with-meta-data

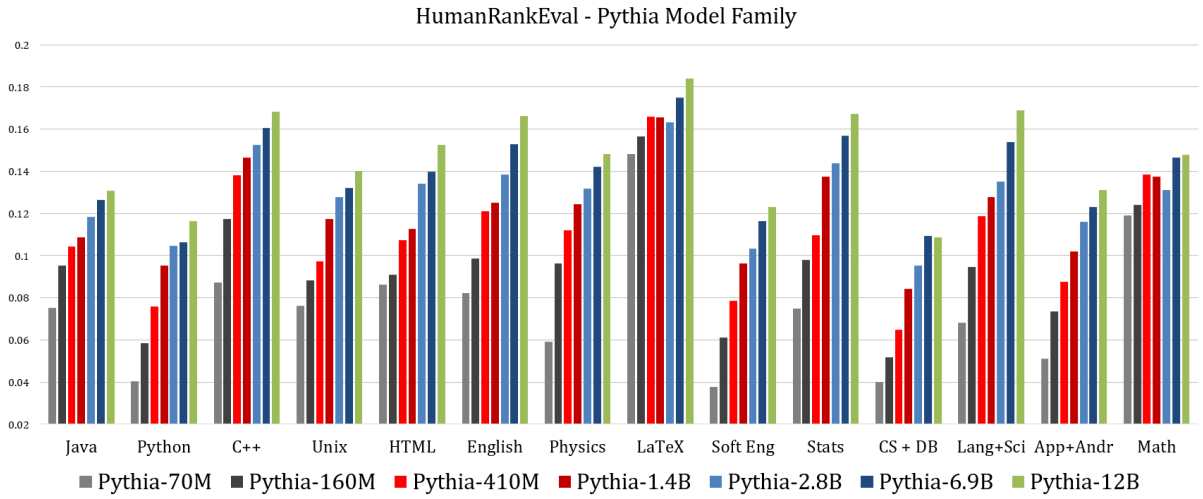

Figure 4: HumanRankEval (per-topic) scores for Pythia LMs.

for each topic. This resulted in 7K questions (28K answers) spanning 14 topics, shown in Figure 3. The QA pairs collectively received over 700k votes (7k questions, 100+ votes per question on average) from more than 100K domain experts and enthusiasts, assuming a ∼20% proportion of unique users, as in LMSYS-Chat-1M (Zheng et al., 2023a).

3.4 HumanRankEval Score

The HumanRankEval metric is based on the assumption that an LM's conversational quality can be estimated by whether the sequences it produces more frequently are more preferable to humans than the infrequent ones. Sequence generation tasks such as WMT (Barrault et al., 2020), HumanEval (Chen et al., 2021) and GSM8K (Cobbe et al., 2021) provide the LM with a prompt (e.g. problem description), generate the output token-by-token, possibly extract the answer from the returned text, then compute the score. Alternatively, we can provide the questions as prompts to the LM, and assuming direct access to the logits of the LM being evaluated, determine the log-likelihood of the HRE human-authored answers under that model's distribution. More formally, we compute the log-likelihood of answer tokens T_a using model p (normalised by character length C_a) conditioned on the question, to obtain log-likelihoods ll for each answer $a \in A$, as shown in Equation 1.

$$ll = \left[\frac{1}{C_a} \sum \log\left(\frac{e^{p(t)}}{\sum_{t=1}^{T_a} e^{p(t)}} \right) \right] \forall a \in A \quad (1)$$

Note that C_a and T_a are obtained only from answer tokens, a standard implementation.[8] Subsequently, the log-likelihoods ll are correlated with human rankings using Pearson (Freedman et al., 2007) correlation. A discussion about the reasons for choosing Pearson over Spearman Rank (Zar, 2005) coefficient follows in section 5.1. Finally, the correlation coefficients are micro-averaged across all 7K questions to compute the HumanRankEval score.

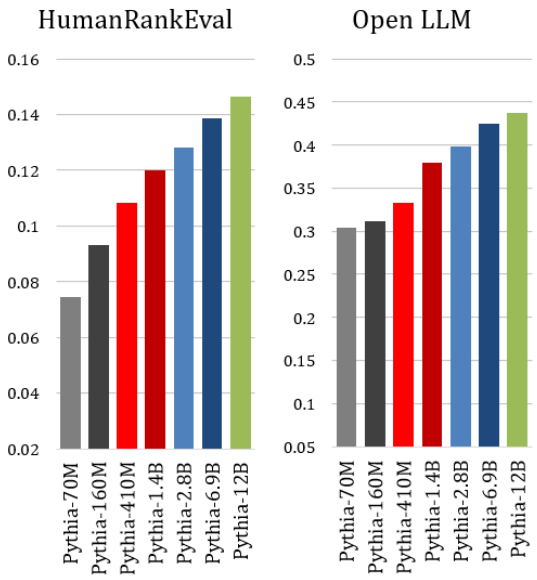

Figure 5: HumanRankEval (avg) scores for Pythia LMs.

4 Results

4.1 Experimental Settings

We benchmark a broad selection of open-source LMs (pretrained and instruction-tuned) available

[8] We follow Eleuther's tokenizer-agnostic method of character (rather than token) length normalisation.

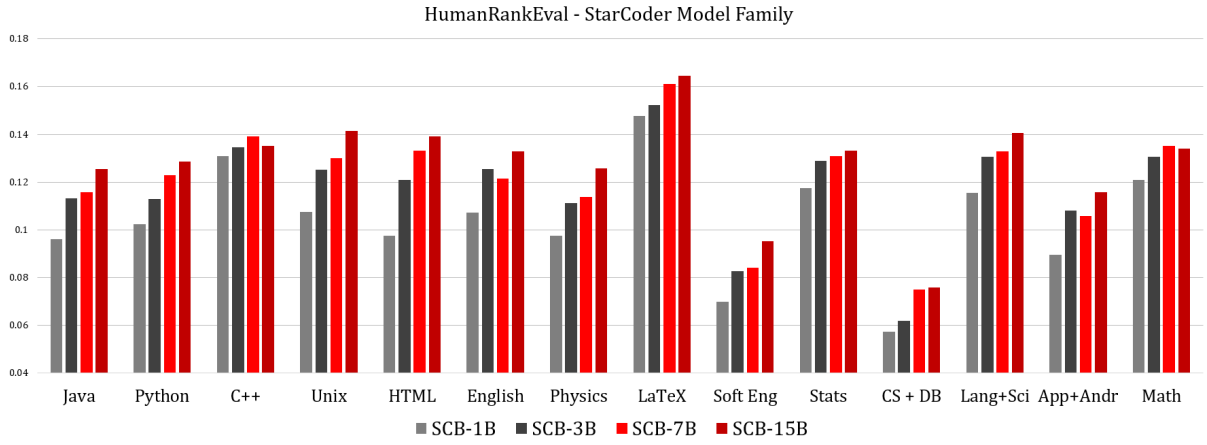

Figure 6: HumanRankEval (per-topic) scores for the StarCoderBase (SCB) LMs.

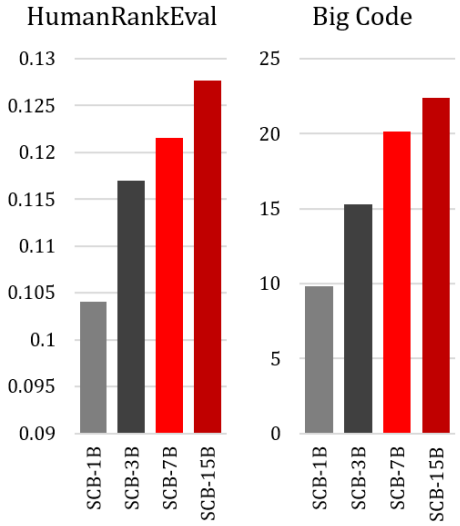

Figure 7: HumanRankEval (avg) scores for the StarCoderBase (SCB) LMs versus Big Code Leaderboard.

from the Huggingface repository (Wolf et al., 2019). LMs with AutoModel[9] and deepspeed inference[10] support (tensor parallel), LM-Eval harness (Gao et al., 2021) compatibility, up to 16B parameters in size were selected for efficient iteration and accessible research. This includes some of the most popular and frequently used LMs such as Llama2, Llama2-Chat (Touvron et al., 2023) (7B + 13B), CodeLlama, CodeLlama-Instruct (Rozière et al., 2023) (7B + 13B), Palmyra (Writer, 2023b) and Camel (Writer, 2023a) (5B each, Camel is instruction-tuned), Pythia-Instruct (1.4B) from LambdaLabs[11], Vicuna (7B + 13B, both instruction-tuned) from LMSYS[12], four StarCoder (Li et al., 2023a) and seven Pythia (Biderman et al., 2023) models (from 70M to 15.5B parameters), MPT-Chat (7B) (MosaicML, 2023), Zephyr (7B, instruction-tuned) Alpha + Beta (Tunstall et al., 2023), WizardLM (Xu et al., 2023) (13B) and Koala (13B) (Geng et al., 2023), both instruction-tuned. Proprietary LMs were excluded as HRE needs access to the logits to compute scores.

4.2 Increasing Model Sizes

In this section, we verify the consistency of HRE scores by observing how they increase as the size of pretrained models (code and natural language) from the same families increases. This expectation is based on the assumption that the learning capacity and general capabilities of LMs increase with the number of trainable parameters (keeping the data constant), and is supported by their performance in OpenLLM. Figures 4 and 6 show the per-topic scores while Figures 5 and 7 show the overall (micro-average) scores for seven Pythia models (70M - 12B) and four StarCoderBase models (1B - 15.5B), respectively. The Pythia models were specifically trained to study LM behavior across different sizes. As expected, different models are cleanly separated by HumanRankEval. Using a single factor ANOVA, the differences were significant between the Pythia (p=8.32-e12) and the StarCoderBase models (p=0.048).

4.3 Correlation with Human Evaluation

In order to support HRE as a reliable proxy for human judgement, we show how its scores correlate with the human-obtained Chatbot Arena ratings.

[9] https://huggingface.co/docs/transformers/model_doc/auto

[10] https://www.deepspeed.ai/inference/

[11] https://huggingface.co/lambdalabs

[12] https://hf.co/lmsys/vicuna-13b-v1.5

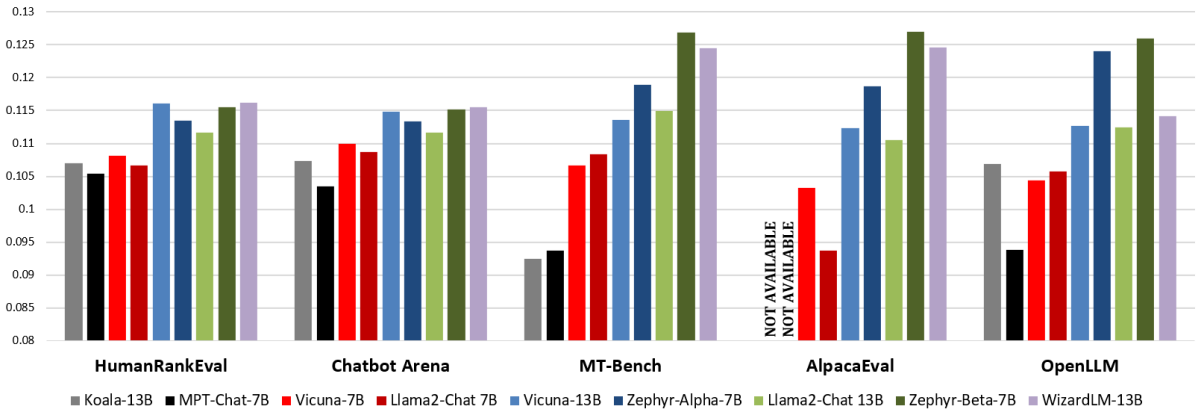

Figure 8: LM rankings (normalised scores) by HumanRankEval, Chatbot Arena, AlpacaEval, MT-Bench and OpenLLM. Koala-13B and MPT-Chat-7B were not available on the AlpacaEval leaderboard at the time of writing.

To this end, in Figure 8, we plot the scores for various instruction-tuned models of different sizes and families. We use the latest Chatbot Arena ratings (as of writing this; 1st Nov. 2023) that were computed from ∼90k user votes. We observe that HRE and Chatbot Arena rankings are the most similar, while there is an obvious misalignment between rankings produced by other popular leaderboards, i.e. MT-Bench (LM-as-a-judge), AlpacaEval (LM-as-a-judge) and OpenLLM (knowledge-based). Figure 9 shows the Pearson correlations between the various rankings. HRE shows the best correlation (0.96) with the human judgements of Chatbot Arena across existing leaderboards, which is to be expected as it was specifically designed for evaluating LMs as conversational assistants. OpenLLM correlates the least (0.85) with human ratings, perhaps unsurprisingly as it consists of knowledge-based automatic tasks. MT-bench's correlation (0.92) indicates that using LM-as-a-judge does offer estimations closer to human judgement than knowledge-based automatic tasks. AlpacaEval is excluded from Figure 9 as rankings were unavailable for some models.

Figure 9: Pearson correlations between HRE, OpenLLM (OLL), MT-Bench (MT) and Chatbot Arena (CBA) model rankings. MT-Bench and OpenLLM have the lowest average agreements.

	CBA	HRE	MT	OLL
CBA	1.00	0.96	0.92	0.85
HRE	0.96	1.00	0.87	0.80
MT	0.92	0.87	1.00	0.83
OLL	0.85	0.80	0.83	1.00

4.4 Instruction Tuning

HRE was developed specifically for evaluating LMs as conversational assistants, i.e. to assess the benefits of methods like instruction tuning and/or preference optimization. MT-Bench and AlpacaEval benchmark only instruction-tuned LMs thus we cannot observe how sensitive they may be to differences between pretrained and instruction-tuned models. As OpenLLM includes both types, we plot OpenLLM and HRE rankings for popular pretrained LMs and their instruction-tuned variants in Figure 10. We observe a divergence between the two leaderboard rankings, further indicated by the low correlation between them (0.3, Pearson). OpenLLM scores underestimate the impact of instruction tuning and/or preference optimization on models' ability to follow human instructions. This is not surprising since most of its tasks assess specific types of knowledge, but this is not necessarily indicative of how an LM would

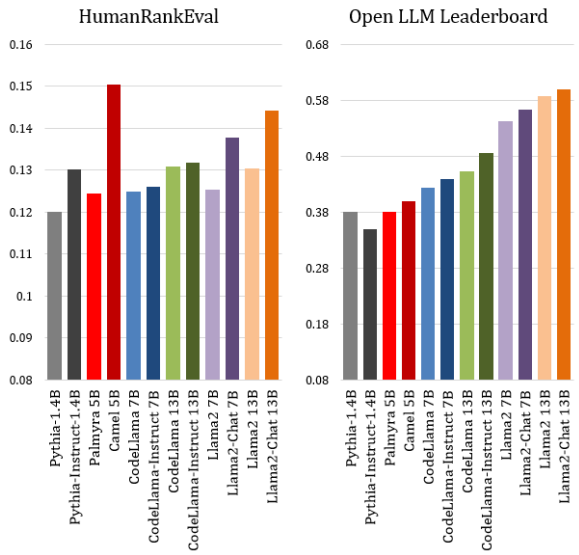

Figure 10: HumanRankEval and OpenLLM scores for a selection of pretrained and instruction-tuned LMs.

perform these tasks in an open-ended conversational setting. On the other hand, we can see that Camel-5B shows a large improvement in HRE after instruction-tuning compared to its base model, Palmyra-5B (no contamination suspected, see Section 5.2). Similarly, Llama2-13B obtains a higher score on the OpenLLM leaderboard than Llama2-Chat-7B, however, HRE is able to correctly detect the superior instruction-following ability of the smaller model, confirmed by the Chatbot Arena ratings in Figure 8. Another example (from same figure) shows an equal or higher preference by humans for Vicuna (13B) over Zephyr (7B) models despite the latter showing a significantly higher score on the OpenLLM Leaderboard. Overall, LMs fine-tuned with instruction data tend to show a noticeable improvement in HRE scores over their "vanilla" pretrained counterparts, however, there are exceptions. We hypothesise that including Self-Instruct (Wang et al., 2022) data (LM-generated, automatically filtered outputs used for fine-tuning code LMs) in CodeLlama-Instruct training may be causing the weak improvement as Wang et al. (2023b) have shown that training with such data adversely affected performance across factual, multilingual and reasoning tasks.

5 Discussion

5.1 Pearson over Spearman

The choice between Pearson and Spearman coefficients for HRE is based on whether correlating the likelihoods/votes themselves or the derived rankings, results in higher agreement with human ratings in Chatbot Arena (see Figure 9). Correlation would be lower for Spearman (0.85) than Pearson (0.96), suggesting a better fit for the latter metric. This is most likely due to the votes being non-uniformly distributed, and Pearson correlation is more suitable for such distributions. Empirically, we also observed that using Pearson results in a clearer separation of models. Figures 4 and 6 show the individual topic scores while Figures 5 and 7 show the averages over 14 topics. We can observe that Pearson correlation monotonically increases for 11 out of 14 topics for Pythia models and 10 out of 14 for the StarCoderBase models. On the other hand, Spearman correlation leads to a less clear separation of models, with only 5 out of 14 topics showing a monotonic increase for Pythia and 3 out of 14 for StarCoderBase respectively (see Figures 12 and 13 in the Appendix).

5.2 Data Contamination

Training LMs on content sourced from StackOverflow and/or StackExchange is not uncommon, e.g. the training data of reward models for Llama2-Chat includes StackExchange data while 2% of Llama1 (Touvron et al., 2023) pretraining data comes from StackExchange. We posit that instruction-tuning on *QA pairs* that overlap with HRE would be the most likely cause of overestimated scores, rather than pretraining on raw web pages. According to their model cards, the benchmarked LMs such as Camel, Vicuna, Koala, MPT-Chat, Pythia-Instruct, Zephyr and Llama-Chat were not instruction-tuned with our data sources yet they show a strong improvement in HRE scores. However, not all LMs provide detailed training information hence the risk of contamination would in those cases be difficult to determine. The most appropriate future-proof action may be deduplicating training data against HumanRankEval to mitigate risks of contamination and accidental score inflation.

6 Conclusions

Multiple benchmarks have been proposed for evaluation of LMs as conversational assistants. However, these are either not specifically designed for this purpose, rely on large (usually proprietary) LMs as the ground truth, or are difficult to scale in terms of sourcing reliable human judges. We have therefore introduced HumanRankEval, a novel automatic evaluation task that comprises a dataset of human-authored questions and answers coupled with a metric. The votes for each question were obtained from over 100 participating domain experts (on average), resulting in high-quality human preferences. HRE performs evaluation by measuring how well the LM's "preferences", estimated as log-likelihoods of answers, correlate with human ratings. To validate HRE, we demonstrated that it cleanly separates pretrained and instruction-tuned LMs of various sizes, and showed that its scores correlate well with human ratings. Relative to knowledge-based evaluation, HRE is particularly adept at detecting changes to LMs' behaviour introduced by instruction-tuning and/or preference optimization. While knowledge-based automatic evaluation can test for specific skills, undesirable biases and essential world knowledge, we expect HRE to accelerate the development of LMs as conversational assistants by providing unique insights.

7 Limitations

Human preferences for our purposes were treated as a composite attribute, and no individual components such as helpfulness, factual correctness, timeliness, safety and so on can be estimated individually by HumanRankEval. LMs scoring higher on HRE are not necessarily more factually correct, less biased or more safe hence researchers are advised to conduct separate evaluation(s) to explicitly test for such behaviours. We acknowledge that, unlike knowledge-based evaluation, the ground truth of human preferences cannot be obtained with the same level of exactitude. HumanRankEval is a new addition to the current consensus and it is possible that the ground truth of human preferences may not be adequately described by any single metric or benchmark. While HRE covers a diverse collection of topics, there are specialist domains that may not be included, but are desired by some researchers. In those cases, we recommend to follow our methodology to extend HRE coverage to new domains that may be of interest. This applies to additional languages as HumanRankEval is overwhelmingly composed of English language content. Neither the StackOverflow nor the StackExchange data have specified any licence information, instructions for intended use or the presence of undesirable content. We subsample the data as is, relying on the corresponding creators of the archives for following appropriate steps. Lastly, we advise that researchers do not solely rely on HRE to verify that a model can be released for public use, and we recommend that human judgement is consulted instead.

References

Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2023. Can we trust the evaluation on chatgpt? *arXiv preprint arXiv:2303.12767*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open LLM Leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

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. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias

Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. *Blog post, April*, 1.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *arXiv preprint arXiv:2307.04657*.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023a. Starcoder: may the source be with you!

Ruosen Li, Teerth Patel, and Xinya Du. 2023b. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023c. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023d. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *CoRR*, abs/2007.08124.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023a. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023b. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

MosaicML. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Nino Scherrer, Claudia Shi, Amir Feder, and David M Blei. 2023. Evaluating the moral beliefs encoded in llms. *arXiv preprint arXiv:2307.14324*.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2023. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Chris van der Lee, Albert Gatt, Emiel Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech and Language*, 67:101151.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023a. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023c. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Writer. 2023a. Camel-5B InstructGPT. https://dev.writer.com.

Writer. 2023b. Palmyra-base Parameter Autoregressive Language Model. https://dev.writer.com.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*.

Zhiqiang Yuan, Junwei Liu, Qiancheng Zi, Mingwei Liu, Xin Peng, and Yiling Lou. 2023. Evaluating instruction-tuned large language models on code comprehension and generation. *arXiv preprint arXiv:2308.01240*.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National*

Conference on Computational Linguistics (Volume 2: Frontier Forum), pages 88–109, Harbin, China. Chinese Information Processing Society of China.

A Appendix

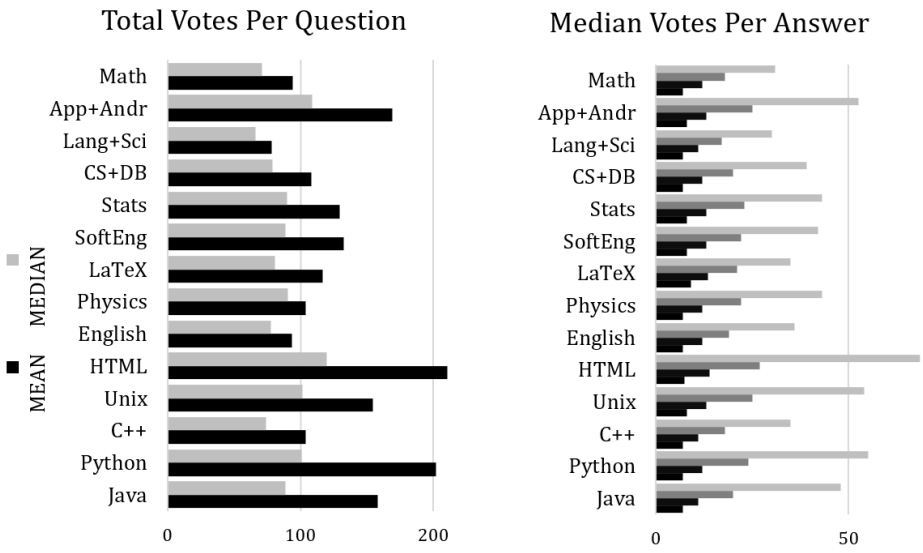

Figure 11: Total votes received per question (median, mean) and median votes received per answer.

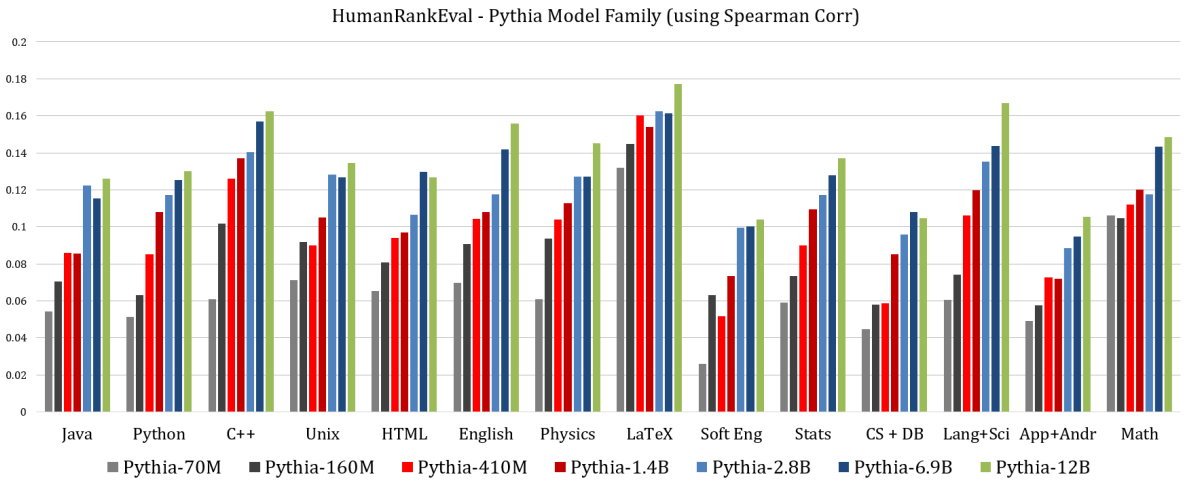

Figure 12: HRE using Spearman correlation (per-topic and overall scores) for the Pythia LMs.

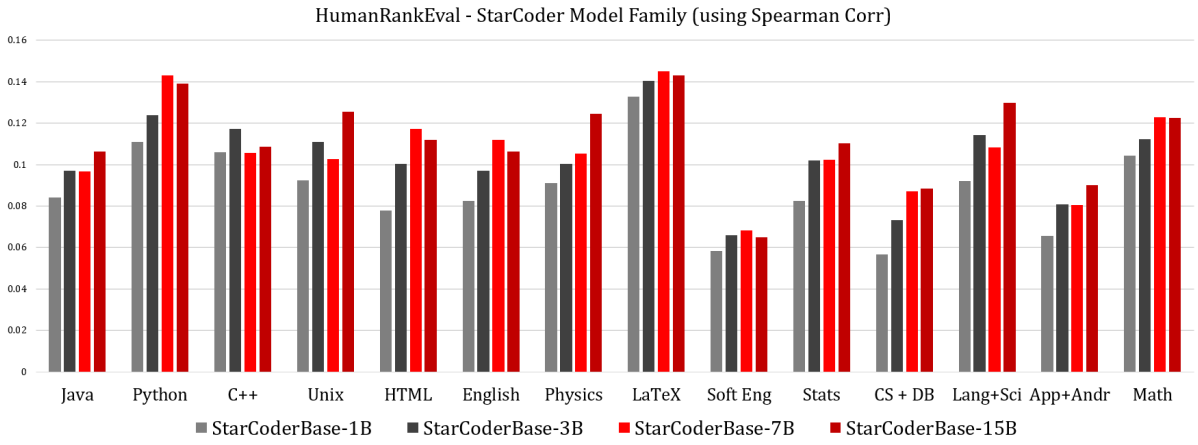

Figure 13: HRE using Spearman correlation (per-topic and overall scores) for the StarCoderBase LMs.