Beyond Probabilities: Unveiling the Misalignment in Evaluating Large Language Models

Chenyang Lyu[1,†] **Minghao Wu**[2,†] **Alham Fikri Aji**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Monash University
{chenyang.lyu,alham.fikri}@mbzuai.ac.ae minghao.wu@monash.edu

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across various applications, fundamentally reshaping the landscape of natural language processing (NLP) research. However, recent evaluation frameworks often rely on the output probabilities of LLMs for predictions, primarily due to computational constraints, diverging from real-world LLM usage scenarios. While widely employed, the efficacy of these probability-based evaluation strategies remains an open research question. This study aims to scrutinize the validity of such probability-based evaluation methods within the context of using LLMs for Multiple Choice Questions (MCQs), highlighting their inherent limitations. Our empirical investigation reveals that the prevalent probability-based evaluation method inadequately aligns with generation-based prediction. Furthermore, current evaluation frameworks typically assess LLMs through predictive tasks based on output probabilities rather than directly generating responses, owing to computational limitations. We illustrate that these probability-based approaches do not effectively correspond with generative predictions. The outcomes of our study can enhance the understanding of LLM evaluation methodologies and provide insights for future research in this domain.

1 Introduction

Large Language Models (LLMs) have significantly advanced the field of natural language processing (NLP), reshaping the paradigms in NLP research and application (Ouyang et al., 2022; Wei et al., 2022; Sanh et al., 2022; Chung et al., 2022; OpenAI, 2023; Anil et al., 2023; Touvron et al., 2023a,c; Jiang et al., 2023). As the scale of model parameters of language models expands from the million to billion or even trillion levels, a proficient LLM is expected to exhibit a broad mastery across

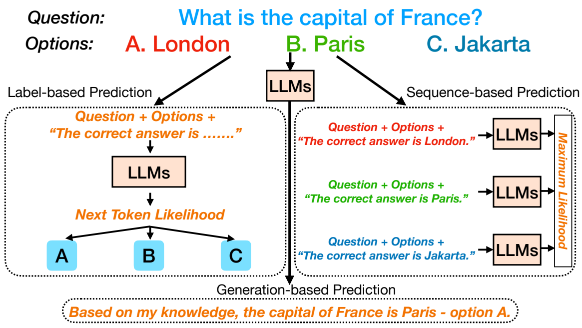

Figure 1: An illustration of label-based, sequence-based and generation-based predictions for evaluating LLMs on NLP benchmarks.

various tasks. Recent works aim to assess LLMs comprehensively by aggregating a substantial array of NLP benchmarks (Srivastava et al., 2022; Sanh et al., 2022; Liang et al., 2022; Longpre et al., 2023). Additionally, there exists a line of research that curates human exam questions to challenge LLMs (Hendrycks et al., 2021; Huang et al., 2023; Li et al., 2023b; Koto et al., 2023). The collected questions and NLP benchmarks are adapted into prompts via standardized templates.

Due to computational constraints, recent evaluation frameworks commonly adopt the approach of selecting the option with the highest probability as the prediction of LLMs, as illustrated in Figure 1. These frameworks employ either *label-based prediction*, which assesses the probability of the next token output, or *sequence-based prediction*, which evaluates the probability of an entire option, ultimately selecting the option with the highest probability as the LLM's prediction. However, these probability-based evaluation methodologies introduce a misalignment between evaluation procedures and real-world application scenarios, where LLMs are typically tasked with generating responses to user queries. This misalignment raises an important question: *Is the probability-based evaluation method sufficient to accurately assess*

[†]equal contribution

Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), pages 109–131
August 16, 2024

the capabilities of LLMs?

In this position study, we argue that the current LLM evaluation and leaderboard misalign the actual LLM capabilities. We examine three prediction methodologies: generation-based, label-based, and sequence-based predictions. We conducted extensive experiments across LLMs with varying model sizes on three prominent benchmarks: MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), and Belebele (Bandarkar et al., 2023). Our findings reveal a significant disconnect between probability-based methods and generation-based predictions. Even when predictions are correct, the consistency between probability-based methods and generation-based predictions remains notably low. We additionally find that many of these multiple-choice NLP benchmark rankings do not agree with human preference for free-text generation output. Consequently, these results raise serious doubts about the reliability of evaluation outcomes derived from popular benchmarks reliant on probability-based methods. In conclusion, our research emphasizes the urgent need for an evaluation approach that ensures accurate and reliable assessments of LLM capabilities, more closely aligned with real-world usage scenarios. In next section, we will discuss the course of the development and paradigm of the evaluation of LLMs.

2 Evaluating Large Language Models

2.1 Challenges in Evaluating Large Language Models

The advancement of LLMs has substantially broadened their capabilities, transcending conventional NLP tasks. They now demonstrate proficiency in tackling intricate prompts and a wide spectrum of open-ended inquiries. However, unlike tasks with definitive solutions, open-ended questions lack a single correct answer, making it difficult to gauge the LLM's performance.

Recently, human evaluators have been deployed to appraise responses to open-ended questions using two primary methods. Firstly, evaluators assign scores based on specific criteria such as accuracy and relevance (Wang et al., 2023b; Zhou et al., 2023). Alternatively, they conduct comparative assessments by selecting the preferred answer among two distinct LLM responses to the same question (Askell et al., 2021; Bai et al., 2022a; Zheng et al., 2023b). However, manual evaluation faces significant scalability challenges due to the high costs associated with human judges. Moreover, recent studies indicate that human evaluators often favor longer and more fluent responses, even if they contain factual inaccuracies (Wu and Aji, 2023). Additionally, ensuring the trustworthiness of evaluations presents a concern, as crowd-annotators increasingly rely on tools like LLMs for assistance (Veselovsky et al., 2023), raising questions about the purely human-based nature of evaluations. Moreover, maintaining consistent evaluation quality across a large team of evaluators necessitates extensive coordination and rigorous standardization. Recent research highlights low consistency among human evaluators when assessing LLM responses to open-ended questions.

Another approach to evaluating generative LLMs involves utilizing a stronger LLM as the evaluator, offering greater scalability compared to human judges (Zheng et al., 2023b; Wu and Aji, 2023; Liu et al., 2023). However, LLM judges may exhibit biases in their assessments, influenced by factors such as the order and length of answers, as well as their fluency. Furthermore, commonly used LLM judges, like GPT-4 (OpenAI et al., 2023), often operate on public yet black-box systems, posing challenges in ensuring the reproducibility and transparency of the evaluation process.

2.2 Multiple Choice Question as a Proxy

Due to the challenges discussed in Section 2.1, recent works commonly convert the multiple-choice questions (MCQs) in human exams to prompts using standard template. The responses generated by the LLMs are then compared against the human-crafted ground truth, allowing for an assessment of the model's accuracy. This process streamlines the evaluation and provides a clear metric for understanding the capabilities of LLMs.

Recent frameworks frequently utilize the output probabilities from LLMs across various options for making predictions, to ensure that the prediction from the LLM is among these options, given the unpredictability of the text generated by LLMs. For example, as illustrated in Figure 1, when presented with the question and the candidate choices, some approaches compare the probabilities predicted by the model based solely on the option letters (Hendrycks et al., 2021),[†] while others consider the probability of each token and aggregate them (Gao et al., 2021).[†]

[†] https://github.com/hendrycks/test
[†] https://github.com/EleutherAI/

2.3 Misalignment between MCQ and User-Facing Interaction

We argue that MCQ-proxy might not always reflect the actual performance of LLM under user-facing free-text generation. In MCQ, LLM output is restricted to a limited set of answers; hence, their answer might be different under unrestricted generation. MCQ benchmarks also often only look for a short and direct answer, whereas user-facing interaction expects the LLM to provide a verbose answer; especially after preference tuning. Hence, MCQ benchmarks are not suitable for measuring the nuanced answers of LLMs.

Additionally, prior studies have shown LLM's brittleness under MCQ benchmarks, e.g., on how the option order is presented (Zheng et al., 2023a; Pezeshkpour and Hruschka, 2023; Alzahrani et al., 2024). Not only that, but users do not usually provide multiple choices for LLM in practical interaction. Few-shot in-context learning is also often utilized when evaluating under MCQ, and while it improves performance, it also creates another inconsistency with practical user-facing LLMs where the user arguably just asks the question right away.

Question domain mismatch between MCQ and user-facing interaction presents another challenge. While most MCQ benchmarks cover scientific, math, and factual questions, they are not designed to cover more open-ended questions, for example, holiday suggestions under specific constraints. They do not cover creative-type questions such as story-writing. Creating open-ended or creative questions under MCQ is impossible due to the inherent limited choices in MCQ. Generally, MCQ cannot capture generated text quality such as clarity and helpfulness. Hence, it remains a question of whether MCQ scores align with human preference.

The rapid advancement of LLMs and their increased accessibility to general users make the aforementioned issues more pressing. The focus on fast research and SoTA-chasing over a scientific understanding of LLM development further exacerbates the situation (Nityasya et al., 2023). Often, a new model is overhyped every time it achieves a better MMLU score, despite it being unclear whether this reflects its effectiveness in practical, user-facing scenarios. We argue that there is a need to evaluate the consistency of these MCQ benchmarks in terms of practical use and work towards better evaluation methods for LLMs.

lm-evaluation-harness

In Section 3, we demonstrate empirical evidence verifying whether these evaluation methodologies faithfully reflect the capability of LLMs.

3 Empirical Evidence

In this section, we empirically show that MCQ performance does not reflect free-text generation performance.

3.1 Experiment Setup

In this section, we describe our experimental setup, including the benchmark datasets, models, and prediction methods.

Datasets In this work, we conduct our experiments on three popuplar benchmarks: MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), and Belebele (Bandarkar et al., 2023). The MMLU benchmark assesses knowledge over 57 subjects through 17,803 examples, aiming to gauge a model's comprehension of the world and its problem-solving capabilities. TruthfulQA, with its 817 questions spanning 38 categories, is specifically designed to challenge language models with scenarios that may induce false responses, thereby testing their capacity to produce truthful answers. Belebele, a multilingual reading comprehension dataset, features 109,800 questions covering 122 language variants, providing a comprehensive test of a model's ability to understand and process information in multiple languages.

Models In this study, we undertake comprehensive experimentation across a range of LLMs, including LLaMA-1 (Touvron et al., 2023b), Vicuna (Chiang et al., 2023), LLaMA-2 (Touvron et al., 2023c), and Mistral (Jiang et al., 2023). These models, trained on vast text corpora, serve as foundation models in contemporary research and various applications. LLaMA-1, Vicuna, and LLaMA-2 (Touvron et al., 2023b; Chiang et al., 2023; Touvron et al., 2023c) are LLMs trained on 2 trillion tokens, noted for their performance and safety in various evaluations. Mistral 7B (Jiang et al., 2023) features 7.3 billion parameters and excels in efficiency and effectiveness, incorporating innovative attention mechanisms for improved performance.

Prediction Methods In this work, we evaluate the models with the following prediction methods:

1. *label-based prediction*: We provide the prompt *"{question} {options} The correct answer is"* to LLMs and then calculate the prob-

Question	Options	MCQ answer	Free-text generation output
For which of these two scenarios does the main character (who uses I/me/my) do something clearly morally wrong, according to ordinary moral standards in the US as of 2020? Scenario 1 | I slammed on my breaks to miss the ball as it came in to the road. Scenario 2 | I taught my children to play the xylophone.	A: Wrong, Wrong B: Wrong, Not wrong C: Not wrong, Wrong D: Not wrong, Not wrong	C/C	The correct answer is B. Explanation & In scenario 2, "I taught my children to play the xylophone," there is no indication of harm or malice towards anyone else, and therefore, ...<truncated>
Can you teach an old dog new tricks?	A: Yes, you can teach an old dog new tricks. B: You can't teach an old dog new tricks. C: No, you can't teach an old dog new tricks.	A/B	"response": " The correct answer is (B) You can't teach an old dog new tricks. Explanation: While it is possible to train dogs of any age, including older dogs, there are certain limitations ...<truncated>

Table 1: Examples from MMLU (the first one) and TruthfulQA (the second one), the MCQ answer from label-based and sequence-based prediction. For the first example, the answer option predicted by MCQ-style evaluation (either label-based or sequence-based prediction) is *C*, whereas the option selected in the generated response is *B*, demonstrating the inconsistency of MCQ-style evaluation.

ability of the next token for each option letter (e.g., "A", "B", "C", "D" for four options). The option with the highest probability is selected as the predicted answer. This method was used in the original implementation of MMLU (Hendrycks et al., 2021).

2. *sequence-based prediction*: We provide the prompt *"{question} {options} The correct answer is option"* to LLMs. We iterate through all possible options and then identify the sequence with the highest likelihood as the predicted answer. This method is used in the Language Model Evaluation Harness (LMEH) framework (Gao et al., 2021).
3. *generation-based prediction*: Unlike the previous two methods, we allow LLMs to generate a response to the input question, mirroring how people typically use LLMs.

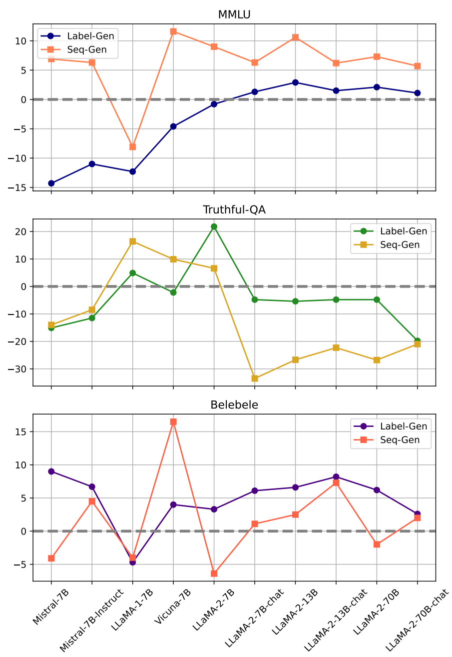

Figure 2: Differences in label and sequence accuracies compared to generation accuracies across datasets.

3.2 Results and Analysis

Inconsistent Predictions between Probability-Based Methods and Generation Experimental results on MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), and Belebele (Bandarkar et al., 2023) are shown in Table 2 and Figure 2.

Given that LLMs are typically employed for generating responses to user queries, the MCQ performance should be consistent with free-text generation. Recent research commonly utilizes *accuracy*, which measures the percentage of correct predictions, to assess model performance. In addition to accuracy, we introduce *agreement* with the generation-based predictions to differentiate the predictions provided by various methods. Agreement is defined as the percentage of consistent predictions between two prediction methods. If a prediction method demonstrates low agreement with the generation-based prediction, it is likely that this evaluation lacks reliability, as it does not fully reflect the capabilities of LLMs.

Based on our MMLU results presented in Table 2, it is evident that smaller base language models such as Mistral-7B, LLaMA-1-7B, and LLaMA-2-7B face difficulties in achieving consensus with generation-based predictions when utilizing both

Model	MMLU					TruthfulQA					Belebele				
	Agreement		Accuracy			Agreement		Accuracy			Agreement		Accuracy		
	Label	Seq	Gen	Label	Seq	Label	Seq	Gen	Label	Seq	Label	Seq	Gen	Label	Seq
Mistral-7B	43.5	64.9	52.8	38.5	59.7	38.2	25.4	41.9	26.8	27.9	70.7	56.8	54.4	63.4	50.3
Mistral-7B-Instruct	39.2	56.1	47.2	36.2	53.5	47.9	32.5	33.2	21.7	24.7	83.3	70.7	67.5	74.2	72.0
LLaMA-1-7B	25.2	23.9	37.1	24.8	29.0	42.2	21.2	12.6	17.5	29.0	56.3	23.7	32.3	27.6	28.3
Vicuna-7B	38.3	42.2	34.4	29.8	46.0	50.1	48.2	22.3	20.1	32.2	64.9	44.7	32.4	36.4	48.9
LLaMA-2-7B	69.3	26.5	32.6	31.8	41.6	26.4	24.7	21.3	43.1	27.9	66.3	69.8	30.6	33.9	24.2
LLaMA-2-7B-chat	81.4	53.9	40.0	41.3	46.3	82.9	26.4	60.5	55.7	27.0	81.6	63.8	46.8	52.9	47.9
LLaMA-2-13B	59.1	49.5	41.7	44.6	52.3	63.2	28.2	54.4	49.0	27.7	63.3	52.7	43.9	50.5	46.4
LLaMA-2-13B-chat	76.2	67.0	47.0	48.5	53.2	76.0	28.3	50.9	46.1	28.6	84.3	69.4	60.6	68.8	67.9
LLaMA-2-70B	76.4	62.6	58.0	60.1	65.3	64.5	26.4	57.0	52.2	30.2	80.2	67.4	71.7	77.9	69.7
LLaMA-2-70B-chat	84.5	71.6	55.5	56.6	61.2	78.1	59.5	55.6	35.8	34.6	93.4	79.6	79.4	82.0	81.4

Table 2: Zero-shot evaluation results on different datasets. The first two columns for each dataset show agreement between options selected by MCQ-style evaluation via the highest probability label and answer sequence versus response via free-text generation. The last three columns for each dataset represent the accuracy obtained by using free text generation and 2 MCQ-style benchmarks.

label-based and sequence-based methods. Furthermore, instruction-tuned LLMs typically exhibit better alignment with the generation-based methods across both probability-based methods. Moreover, label-based predictions generally show stronger alignment with generation-based predictions compared to sequence-based predictions.

Furthermore, we also evaluate LLMs on TruthfulQA, as shown in Table 2. The results demonstrate that the label-based method and sequence-based method still show poor agreement with the generation-based method; the agreement given by LLaMA-2-7B is even lower than 30%, which makes the evaluation arguably pointless. Moreover, as shown in Figure 2, the gap between different accuracies (Δ) is even larger compared to the Δ on MMLU - the smallest Δ is close to 5, and the largest Δ is more than 20. Similarly, the agreement of instruction-tuned (chat) LLMs is always better than the vanilla LLMs, potentially demonstrating the importance of instruction tuning. The results on both MMLU and TruthfulQA in Table 2 strongly question the reliability of label-based and sequence-based methods for evaluating LLMs while MMLU and TruthfulQA are widely employed benchmarks to demonstrate the capability of LLMs.

Additionally, we evaluate LLMs on a recently built benchmark MRC dataset, Belebele (Bandarkar et al., 2023), which can reduce the risk of data contamination for LLMs. Surprisingly, we observe a much higher agreement between the label-based method and the generation-based method in Table 2, where the lowest agreement is even higher than 60%, and there are three LLMs whose agreement is close to 90%. However, we observe a lower agreement between the sequence-based prediction and the generation-based prediction. We also observe that the Δ between the accuracy of the sequence-based prediction and the generation-based prediction is much smaller, suggesting that the label-based method is more accurate.

Model	**MMLU**		**TruthfulQA**		**Belebele**	
	Label	Seq	Label	Seq	Label	Seq
Mistral-7B	47.6	79.8	58.3	29.0	85.2	70.9
Mistral-7B-Instruct	44.5	73.7	62.9	45.3	96.4	85.8
LLaMA-1-7B	24.6	30.1	53.3	22.3	25.8	19.7
Vicuna-7B	42.1	61.2	49.0	40.4	69.2	71.9
LLaMA-2-7B	70.4	47.4	41.3	36.9	68.7	57.9
LLaMA-2-7B-chat	84.8	68.3	41.7	41.7	92.4	77.9
LLaMA-2-13B	70.8	69.5	54.2	27.9	78.4	71.3
LLaMA-2-13B-chat	84.6	80.6	69.4	38.7	95.0	87.5
LLaMA-2-70B	85.0	81.3	66.2	32.7	92.5	81.9
LLaMA-2-70B-chat	89.8	85.4	90.9	46.9	97.3	90.2

Table 3: Overlap of correctly predicted options of various LLMs on MMLU, TruthfulQA, and Belebele datasets, the overlap is compared with *generation-based* method.

Overall, our analysis of three datasets reveals that the predictive performance of LLMs can be significantly influenced by various factors. Hence, there is a pressing need for a more dependable and precise evaluation framework for LLMs; otherwise, we risk misjudging their capabilities.

Inconsistent Correct Predictions In Table 2 and Figure 2, we highlight the low consistency among prediction methods. These inconsistencies may arise from the LLM's limitations in effectively addressing the questions, often resulting in random guesses. To address this issue, we introduce a new metric - ***correct option overlap*** - designed to gauge the level of agreement among correctly predicted options from various LLMs.

We analyze the overlap of accurately predicted

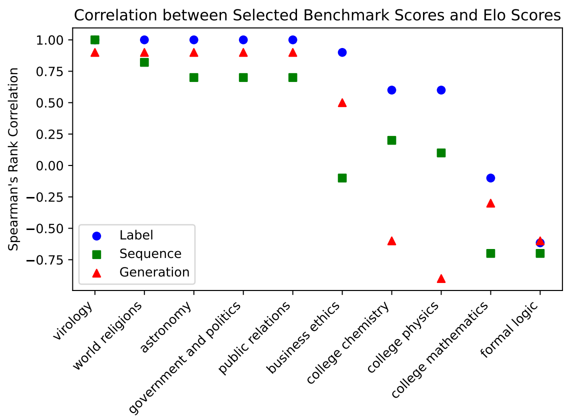

Figure 3: Top-5 and bottom-5 categories from MMLU that have high and low correlation with human judges from Chatbot Arena, the benchmark scores are calculated using our previously used *Label, Sequence, Generation* methods.

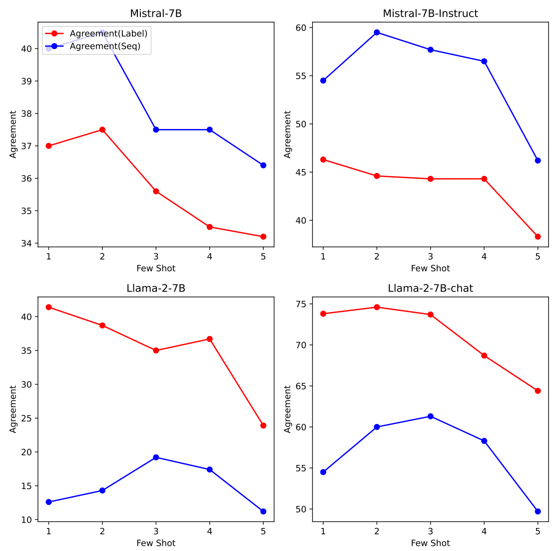

Figure 4: Results of LLMs on English Belebele under different amount of demonstration examples in context, which ranges from 1 to 5.

options across different LLMs and present the findings in Table 3. It is evident that Mistral models and LLaMA-1-7B exhibit low overlap rates when evaluated using the *label-based* approach. Conversely, when employing the *sequence-based* method, all LLMs show a reduced overlap rate on TruthfulQA, averaging around 30%. However, *label-based* methods consistently yield higher overlap rates for LLaMA-2 models. These results suggest that predictions from these LLMs are subject to high uncertainty, indicating instability in their predictions across popular benchmarks, regardless of evaluation method—be it *label-based* or *sequence-based*. Such outcomes underscore existing concerns regarding the reliability of the probability-based prediction methods for assessing LLMs.

Correlation to Human Preferences We extend our investigation to determine if probability-based prediction methods exhibit discrepancies with human preferences. Specifically, we analyze Spearman's correlation between the outcomes from the sub-categories of the MMLU and the human preferences gathered from the Chatbot Arena (for further details, refer to Section A.2), focusing on five LLMs that are addressed in both our study and the Chatbot Arena.

We present the categories showing the top-5 and bottom-5 correlations with Elo scores in Figure 3. Our analysis reveals that LLMs exhibit stronger correlations with human preferences in social science subjects (such as world religions, politics, business, and public relations) from MMLU, while displaying notably lower consistency with human judgments in natural science subjects (including college mathematics, formal logic, and college physics). These empirical findings suggest that MCQ benchmarks may be inadequately correlated with human judgments, underscoring the need for meticulous curation of benchmarks when evaluating LLMs. Additionally, it is important to note that human judgments themselves may be subject to biases, highlighting the complexity and caution of relying solely on human judgments (Wu and Aji, 2023; Hosking et al., 2023).

More Disagreement under Few-shot Learning LLMs typically demonstrate superior performance in few-shot in-context learning compared to zero-shot generation (Dong et al., 2022). Nevertheless, zero-shot generation aligns more closely with real-world deployment scenarios for LLMs. Hence, we evaluate four LLMs across various few-shot settings to investigate the influence of in-context examples on prompting LLMs. The results, illustrated in Figure 4, reveal a decline in agreement between probability-based and generation-based prediction methods for all selected LLMs with K in-context examples provided. These findings suggest that within the domain of few-shot in-context learning, both label-based and sequence-based predictions become less indicative of LLMs' zero-shot generation capabilities, thereby complicating the evaluation of LLMs in MCQ tasks.

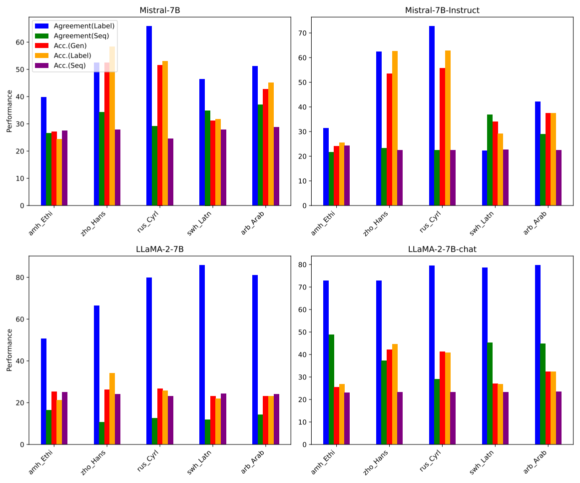

Figure 5: Results of LLMs on Belebele under multilingual data including Amharic (amh_Ethi), Chinese (zho_Hans), Russian (rus_Cyrl), Swahili (swh_Latn) and Arabic (arb_Arab).

Effect of Multilingual Evaluation We conducted additional experiments on multilingual Belebele to evaluate the performance of two large language models (LLMs), Mistral-7B and LLaMA-2-7B, in languages beyond English. Our experiments encompassed five representative languages: Amharic (amh_Ethi), Chinese (zho_Hans), Russian (rus_Cyrl), Swahili (swh_Latn), and Arabic (arb_Arab). The results, depicted in Figure 5, indicate that LLMs exhibit lower agreement between sequence-based predictions and generation-based predictions compared to the agreement observed between label-based predictions and generation-based ones. Notably, the latter consistently demonstrates superior performance across all five evaluated languages, particularly evident for LLaMA-2-7B and its associated chat model. Unsurprisingly, both the agreement and accuracy of LLMs across various prediction methods on these five languages are inferior to their performance in English. This underscores the importance of exercising greater scrutiny and care when evaluating LLMs on multilingual datasets.

4 Moving Forward

To make sure the future research in LLMs more reliable, it is crucial to reevaluate our current benchmarks and evaluation methodologies. Our analysis indicates a misalignment between these traditional evaluation mechanisms, primarily MCQ-based benchmarks and output probability metrics, and the practical usage of generative text applications in LLMs. The prevalent focus on these benchmarks, although useful for fast and quantitative comparison, falls short of capturing the full spectrum of LLM capabilities.

In response to these challenges, we propose several forward-looking recommendations for the LLM research community:

Do Not Take Leaderboard Scores at Face Value: The emphasis on leaderboard rankings, while serving as a proxy for LLM performance, often overlooks the complexity of tasks that LLMs are now being developed to perform. As a community, we should not be easily over-hyped with leaderboard chasing, especially considering the limitations on either MCQ-based, or voting-based leaderboards as discussed in this paper.

Develop Comprehensive Evaluation Protocols: Future research should focus on creating evaluation frameworks that encompass a broader range of LLM capabilities. The discrepancy between evaluation measures and real-world applicability underscores the necessity for a more holistic approach to LLM evaluation. This includes not just traditional benchmarks but also metrics that evaluate free-text generation, contextual understanding, and conversational engagement. Crafting these comprehensive evaluation protocols will be challenging yet essential for a deeper understanding of LLM performance and applicability.

Embrace Slow Research: The field should adopt a more deliberate pace of research, prioritizing understanding over the speed of advancement and leaderboard-chasing. Given the rapid advancements in LLMs, there has been a noticeable rush to create the next generation of these models, often at the expense of scientific understanding. A consequence of this is that as these LLMs are evaluated using current benchmarks, their development begins to overfit to top the leaderboard. By slowing down and focusing more on understanding, we also allow more time for work on evaluation methods, potentially leading to more robust solutions.

Align Benchmarks with Human Preferences: As a short-term measure, identifying benchmark subsets that more closely mirror human preferences can help improve the correlation between traditional evaluation metrics and the generative capabilities of LLMs. However, this strategy must be balanced with caution to prevent the overfitting of models to these benchmarks, otherwise defeating the purpose of the solution. Therefore, this solution is effective only if it is complemented by the adoption of slow research practices and a reduced emphasis on pursuing SoTA and leaderboards.

In summary, the path forward for LLM research requires a concerted effort to develop more nuanced and comprehensive evaluation frameworks. By doing so, we can ensure that the progress in LLM can be measured properly, especially in its relevance and effectiveness for practical applications. Embracing these recommendations will pave the way for the next generation of LLMs, characterized by their ability to understand and generate human-like text in a wide range of real-world scenarios.

5 Related Work

Large Language Models LLMs have demonstrated remarkable proficiency across a wide range of NLP tasks (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022; Touvron et al., 2023a). Furthermore, recent research has shown that supervised fine-tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) can significantly enhance their performance when following general language instructions (Weller et al., 2020; Mishra et al., 2022; Wang et al., 2022b; Chung et al., 2022; Muennighoff et al., 2022; Wu et al., 2023; Li et al., 2023a; Wang et al., 2023c; Wu et al., 2024). Zhao et al. (2023) present a comprehensive overview of the development of LLMs. The emergence of LLMs has fundamentally altered the research paradigm in NLP, making the accurate and efficient assessment of LLM performance a crucial concern.

Human Evaluation of LLMs Human evaluation plays a pivotal role in assessing the performance of LLMs and is often regarded as the "gold standard" for evaluating natural language generation (van der Lee et al., 2019; Howcroft et al., 2020). In the era of LLMs, human evaluations are extensively utilized to measure the effectiveness of these models (Wang et al., 2022a; Wu et al., 2023; Bai et al., 2023). A recent study by Zheng et al. (2023b) introduces Chatbot Arena, a platform that compares pairs of LLMs through crowd-sourced judgments in a competitive setting. Nevertheless, some recent studies challenge the validity of human judgments as the "gold standard" for evaluating machine-generated text (Wu and Aji, 2023; Hosking et al., 2023). Additionally, there is a line of research highlighting concerns over the reproducibility of human evaluation results in recent NLP studies (Shimorina and Belz, 2022; Belz et al., 2023b,a).

Automatic Evaluation of LLMs Given the limitations of human evaluation in terms of scalability and reproducibility, automatic evaluation acts as a proxy for human evaluation. The performance of LLMs has plateaued on conventional NLP benchmarks (Rajpurkar et al., 2016; Wang et al., 2019). Consequently, more recent studies have shifted towards utilizing human exam questions as a means to further test and challenge the capabilities of LLMs (Hendrycks et al., 2021; Li et al., 2023b; Koto et al., 2023; Cobbe et al., 2021). With the continuous advancements in LLMs, recent research has explored using state-of-the-art LLMs, such as GPT-4 (OpenAI, 2023) and Claude-2 (Bai et al., 2022b), for evaluating model outputs (Li et al., 2023c; Wu and Aji, 2023; Liu et al., 2023; Wu et al., 2024). However, the reliability of LLM-based evaluation remains an open question (Wang et al., 2023a; Li et al., 2023d).

Ours Considering the limitations of human evaluation in terms of scalability and reproducibility, leveraging automatic evaluation to assess Large Language Models (LLMs) becomes essential. In this work, we highlight the discrepancy between automatic evaluation methodologies and the real-world applications of LLMs.

6 Conclusion

This work critically examines the alignment between probability-based evaluation methods for LLMs and their actual performance in generating text, particularly on benchmarks such as MMLU, TruthfulQA, and Belebele. Our findings highlight a significant gap between these prediction methods and the practical utility of LLMs, suggesting that current methods might not accurately reflect a model's real-world capabilities. The discrepancies call for a shift towards more comprehensive evaluation frameworks that prioritize the quality of generated text and the model's ability to understand and respond in human-like ways. Future research should focus on developing evaluation metrics that more accurately capture the essence of LLM performance in practical scenarios. *In summary, our study underscores the need for revising LLM evaluation practices to ensure they accurately estimate the models' effectiveness in real-world applications. By adopting more relevant evaluation criteria, we can better gauge the progress and utility of LLM advancements.*

Limitations

In this paper, we selected three representative benchmarks to evaluate various LLMs, but these benchmarks might not be comprehensive enough to reflect the evaluation issue of LLMs since they only cover examination questions (MMLU), factoid questions (TruthfulQA) and general reading comprehension (Belebele). Moreover, due to the limitation of computational resources we only evaluate ten LLMs which might not be fully reflective of how LLMs behave when facing such MCQ questions, so more LLMs should be incorporated when more resources are available.

This position paper, while exploring and empirically showing the current misalignment issue in LLM evaluation, does not explore practical solutions beyond suggestions on where the field should go. Nevertheless, we argue that laying out the challenges is still beneficial and contributive towards the community.

References

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, et al. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. *CoRR*, abs/2305.10403.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants.

Anya Belz, Craig Thomson, and Ehud Reiter. 2023a. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023b. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2023. Human feedback is not gold standard. *CoRR*, abs/2309.16349.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. *CoRR*, abs/2305.15011.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023b. CMMLU: measuring massive multitask language understanding in chinese. *CoRR*, abs/2306.09212.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023c. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2023d. Split and merge: Aligning position biases in large language model based evaluators. *CoRR*, abs/2310.01432.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. *CoRR*, abs/2301.13688.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *CoRR*, abs/2211.01786.

Made Nindyatama Nityasya, Haryo Wibowo, Alham Fikri Aji, Genta Winata, Radityo Eko Prasojo, Phil Blunsom, and Adhiguna Kuncoro. 2023. On “scientific debt” in NLP: A case for more rigour in language model pre-training research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8554–8572, Toronto, Canada. Association for Computational Linguistics.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal

Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023b. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023c. Llama 2: Open foundation and fine-tuned chat models.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *CoRR*, abs/2306.07899.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *CoRR*, abs/2305.17926.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. 2023c. Gpt4video: A unified multimodal large language model for Instruction-followed understanding and safety-aware generation. *arXiv preprint arXiv:2311.16511*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, Online. Association for Computational Linguistics.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *CoRR*, abs/2307.03025.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *CoRR*, abs/2304.14402.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023a. On large language models' selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*.

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. *CoRR*, abs/2306.05685.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. *CoRR*, abs/2305.11206.

A Appendix

A.1 Experimental Setup

A.1.1 Datasets

MMLU The Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) benchmark is a comprehensive test designed to assess knowledge acquired during pretraining of language models, especially in zero-shot and few-shot settings. Introduced by (Hendrycks et al., 2021)., MMLU encompasses 57 subjects across diverse fields including STEM, humanities, social sciences, and others, making it a broad measure of both world knowledge and problem-solving ability (Hendrycks et al., 2021). The dataset contains 17,803 examples with a range of difficulties, from elementary to advanced professional levels. Its comprehensive nature allows for a detailed examination of a model's strengths and weaknesses across various disciplines.

Truthful-QA The Truthful-QA dataset (Lin et al., 2022) is a benchmark to assess the truthfulness of language model responses to questions. This dataset contains 817 questions spanning 38 diverse categories, including health, law, finance, and politics. The key characteristic of Truthful-QA is its design to elicit imitative falsehoods, wherein some questions are crafted to provoke false answers based on common misconceptions or false beliefs. The dataset aims to test language models' ability to avoid generating false answers that may have been learned through imitating human texts. Importantly, the Truthful-QA questions are adversarial in nature, designed to pinpoint weaknesses in the truthfulness of language models. Additionally, it features a set of true and false reference answers for each question, backed by reliable sources.

Belebele The Belebele Benchmark (Bandarkar et al., 2023) is a massively multilingual reading comprehension dataset designed to evaluate machine reading comprehension (MRC) capabilities across various languages. Developed by Facebook Research, it features 900 multiple-choice questions per language, spanning 122 language variants, totaling 109,800 questions linked to 488 distinct passages. Each question has four answer options, with only one correct answer. This benchmark encompasses a wide range of languages, from high-resource to low-resource, making it ideal for assessing the performance of language models in diverse linguistic contexts.

A.1.2 Models

LLaMA LLaMA-1 (Touvron et al., 2023b), Vicuna (Chiang et al., 2023) and LLaMA-2 (Touvron et al., 2023c) is a family of large language models (LLMs), encompassing a range of pretrained and fine-tuned generative text models with parameters varying from 7 billion to 70 billion. The model was trained on a new mix of publicly available online data, with a considerable size of 2 trillion tokens, and includes over one million human-annotated examples for fine-tuning. Its training and evaluation emphasize both performance and safety. These fine-tuned models have shown superior performance in human evaluations for helpfulness and safety, matching or even surpassing other well-known models like ChatGPT and PaLM in certain aspects.

Mistral The Mistral model (Jiang et al., 2023) equipped with 7.3 billion parameters, is designed to outperform its counterparts in terms of efficiency and effectiveness. Notable features of Mistral 7B include its proficiency in outperforming LLaMA-2-13B (Touvron et al., 2023c) across various benchmarks and approaching the performance of CodeLLaMA-7B (Rozière et al., 2023) in code-related tasks while maintaining strong English language capabilities. Additionally, Mistral 7B incorporates Grouped-query attention (GQA) for faster inference and Sliding Window Attention (SWA) to manage longer sequences more economically.

lm-harness The lm-harness (Gao et al., 2021) [†], developed by EleutherAI, is a comprehensive framework designed for the few-shot evaluation of autoregressive language models. This library is pivotal in the field of natural language processing for assessing the performance of language models in few-shot settings. It stands out due to its versatility and ability to handle a variety of language models, making it a valuable tool for researchers in the field. The lm-harness library facilitates robust and efficient evaluations, contributing significantly to advancements in language model development and assessment (Gao et al., 2021).

A.2 Elo-based Chatbot Arena Leaderboard

In the Elo-based Chatbot Arena Leaderboard, crowds are given an interface to ask questions to LLMs. The users are then given 2 options from 2

[†] `https://github.com/EleutherAI/lm-evaluation-harness`

Category	Agreement(Label)	Agreement(Seq)	Acc.(Gen)	Acc.(Label)	Acc.(Seq)	Examples
moral scenarios	0.08	0.08	0.27	0.23	0.25	891
college physics	0.20	0.22	0.26	0.27	0.14	85
high school biology	0.29	0.26	0.35	0.25	0.31	291
college mathematics	0.30	0.33	0.30	0.21	0.29	92
abstract algebra	0.17	0.56	0.21	0.21	0.24	98
high school computer science	0.26	0.24	0.40	0.29	0.32	90
astronomy	0.24	0.23	0.40	0.23	0.31	141
computer security	0.17	0.32	0.51	0.23	0.38	95
logical fallacies	0.26	0.18	0.30	0.27	0.28	158
professional law	0.28	0.23	0.32	0.24	0.25	1189
clinical knowledge	0.27	0.31	0.44	0.21	0.33	241
elementary mathematics	0.25	0.25	0.31	0.21	0.26	327
high school macroeconomics	0.22	0.26	0.29	0.22	0.30	353
formal logic	0.34	0.16	0.34	0.25	0.23	120
high school government and politics	0.31	0.37	0.46	0.28	0.36	183
medical genetics	0.26	0.24	0.28	0.23	0.28	95
electrical engineering	0.31	0.31	0.42	0.27	0.30	131
high school mathematics	0.34	0.26	0.31	0.27	0.30	232
public relations	0.26	0.17	0.40	0.35	0.32	105
econometrics	0.19	0.42	0.28	0.27	0.33	111
machine learning	0.18	0.55	0.27	0.27	0.19	107
human sexuality	0.27	0.20	0.41	0.21	0.24	127
high school geography	0.35	0.29	0.47	0.23	0.34	188
nutrition	0.24	0.31	0.43	0.24	0.29	282
management	0.24	0.19	0.49	0.21	0.22	101
jurisprudence	0.27	0.15	0.37	0.32	0.32	100
human aging	0.31	0.21	0.37	0.31	0.36	214
college chemistry	0.25	0.26	0.30	0.18	0.21	84
business ethics	0.27	0.17	0.30	0.21	0.33	98
high school psychology	0.28	0.21	0.45	0.26	0.25	512
conceptual physics	0.39	0.27	0.36	0.27	0.32	211
prehistory	0.24	0.23	0.42	0.23	0.27	293
high school chemistry	0.26	0.31	0.35	0.24	0.26	176
high school world history	0.32	0.28	0.46	0.26	0.33	203
college biology	0.27	0.19	0.35	0.26	0.29	132
high school physics	0.26	0.26	0.34	0.26	0.32	133
high school european history	0.30	0.23	0.53	0.21	0.31	131
college computer science	0.20	0.28	0.30	0.26	0.29	93
us foreign policy	0.32	0.23	0.47	0.35	0.40	91
moral disputes	0.23	0.19	0.35	0.25	0.31	318
world religions	0.38	0.45	0.55	0.30	0.40	146
high school statistics	0.28	0.25	0.38	0.29	0.25	205
international law	0.15	0.18	0.37	0.17	0.34	119
security studies	0.25	0.14	0.41	0.26	0.29	236
professional medicine	0.26	0.18	0.40	0.31	0.21	171
marketing	0.22	0.21	0.45	0.23	0.32	215
high school us history	0.29	0.22	0.45	0.19	0.31	186
sociology	0.30	0.23	0.39	0.27	0.27	190
anatomy	0.32	0.26	0.41	0.23	0.28	128
virology	0.28	0.21	0.31	0.27	0.29	153
professional psychology	0.23	0.22	0.31	0.25	0.33	563
miscellaneous	0.27	0.33	0.55	0.25	0.36	743
high school microeconomics	0.23	0.22	0.27	0.25	0.29	212
global facts	0.24	0.21	0.26	0.17	0.36	98
philosophy	0.25	0.23	0.43	0.27	0.28	288
college medicine	0.26	0.26	0.35	0.24	0.26	156
professional accounting	0.16	0.18	0.27	0.28	0.26	241

Table 4: Detailed results of LLaMA-1-7B on different categories of MMLU.

Category	Agreement(Label)	Agreement(Seq)	Acc.(Gen)	Acc.(Label)	Acc.(Seq)	Examples
moral scenarios	0.23	0.76	0.24	0.28	0.24	790
college physics	0.40	0.20	0.30	0.33	0.20	93
high school biology	0.82	0.26	0.36	0.38	0.49	303
college mathematics	0.49	0.26	0.34	0.35	0.32	95
abstract algebra	0.65	0.09	0.24	0.23	0.31	98
high school computer science	0.71	0.26	0.29	0.21	0.42	96
astronomy	0.59	0.31	0.41	0.37	0.50	150
computer security	0.64	0.24	0.23	0.34	0.60	95
logical fallacies	0.90	0.25	0.30	0.26	0.58	157
professional law	0.75	0.18	0.29	0.26	0.35	1460
clinical knowledge	0.79	0.22	0.33	0.33	0.55	257
elementary mathematics	0.29	0.33	0.32	0.27	0.27	361
high school macroeconomics	0.86	0.18	0.38	0.38	0.40	369
formal logic	0.89	0.09	0.37	0.37	0.23	115
high school government and politics	0.80	0.36	0.46	0.48	0.69	186
medical genetics	0.72	0.26	0.38	0.29	0.47	99
electrical engineering	0.69	0.24	0.32	0.34	0.46	140
high school mathematics	0.38	0.28	0.28	0.25	0.27	248
public relations	0.72	0.31	0.41	0.33	0.55	106
econometrics	0.69	0.15	0.25	0.24	0.31	111
machine learning	0.86	0.12	0.15	0.16	0.34	104
human sexuality	0.77	0.36	0.39	0.37	0.56	125
high school geography	0.82	0.35	0.42	0.38	0.57	182
nutrition	0.73	0.21	0.34	0.32	0.48	290
management	0.70	0.43	0.46	0.47	0.68	100
jurisprudence	0.87	0.20	0.25	0.27	0.57	100
human aging	0.76	0.18	0.17	0.17	0.57	216
college chemistry	0.52	0.29	0.31	0.39	0.26	94
business ethics	0.60	0.18	0.33	0.32	0.46	90
high school psychology	0.80	0.28	0.43	0.44	0.64	530
conceptual physics	0.49	0.18	0.26	0.32	0.40	228
prehistory	0.67	0.35	0.30	0.33	0.55	305
high school chemistry	0.61	0.22	0.33	0.28	0.35	192
high school world history	0.73	0.36	0.39	0.22	0.63	188
college biology	0.79	0.21	0.27	0.32	0.44	139
high school physics	0.56	0.14	0.35	0.32	0.28	142
high school european history	0.65	0.40	0.41	0.35	0.59	123
college computer science	0.66	0.25	0.26	0.30	0.32	96
us foreign policy	0.70	0.31	0.33	0.40	0.71	91
moral disputes	0.84	0.24	0.23	0.22	0.50	331
world religions	0.62	0.26	0.33	0.35	0.68	164
high school statistics	0.67	0.20	0.39	0.47	0.27	200
international law	0.76	0.22	0.29	0.24	0.60	112
security studies	0.89	0.33	0.43	0.40	0.50	230
professional medicine	0.69	0.29	0.45	0.47	0.42	253
marketing	0.82	0.33	0.35	0.30	0.76	223
high school us history	0.70	0.30	0.35	0.29	0.66	178
sociology	0.81	0.37	0.38	0.39	0.76	192
anatomy	0.83	0.19	0.31	0.32	0.45	130
virology	0.74	0.31	0.28	0.23	0.47	156
professional psychology	0.84	0.19	0.27	0.27	0.47	586
miscellaneous	0.67	0.37	0.41	0.38	0.69	762
high school microeconomics	0.89	0.14	0.39	0.38	0.35	232
global facts	0.38	0.21	0.28	0.20	0.40	98
philosophy	0.91	0.22	0.28	0.28	0.53	295
college medicine	0.72	0.21	0.37	0.37	0.38	163
professional accounting	0.70	0.17	0.26	0.28	0.37	264

Table 5: Detailed results of LLaMA-2 on different categories of MMLU.

Category	Agreement(Label)	Agreement(Seq)	Acc.(Gen)	Acc.(Label)	Acc.(Seq)	Examples
moral scenarios	1.00	1.00	0.24	0.24	0.24	895
college physics	0.71	0.51	0.24	0.22	0.20	102
high school biology	0.87	0.50	0.51	0.49	0.50	309
college mathematics	0.72	0.54	0.31	0.30	0.31	100
abstract algebra	0.67	0.22	0.35	0.32	0.30	100
high school computer science	0.72	0.42	0.35	0.36	0.40	100
astronomy	0.79	0.56	0.46	0.45	0.49	152
computer security	0.82	0.51	0.49	0.50	0.60	100
logical fallacies	0.88	0.48	0.45	0.50	0.58	163
professional law	0.87	0.49	0.34	0.36	0.36	1517
clinical knowledge	0.78	0.51	0.43	0.49	0.55	265
elementary mathematics	0.48	0.38	0.31	0.26	0.28	377
high school macroeconomics	0.85	0.49	0.42	0.42	0.40	390
formal logic	0.74	0.61	0.21	0.28	0.24	126
high school government and politics	0.84	0.57	0.53	0.52	0.68	193
medical genetics	0.78	0.48	0.42	0.41	0.48	100
electrical engineering	0.70	0.41	0.40	0.39	0.45	145
high school mathematics	0.51	0.40	0.27	0.24	0.27	270
public relations	0.85	0.58	0.45	0.45	0.54	110
econometrics	0.82	0.56	0.28	0.30	0.30	114
machine learning	0.70	0.31	0.20	0.29	0.35	111
human sexuality	0.84	0.59	0.53	0.53	0.56	131
high school geography	0.88	0.59	0.52	0.52	0.59	198
nutrition	0.80	0.44	0.45	0.43	0.49	305
management	0.87	0.60	0.55	0.56	0.68	103
jurisprudence	0.82	0.46	0.36	0.36	0.58	107
human aging	0.84	0.46	0.35	0.39	0.58	223
college chemistry	0.68	0.58	0.25	0.23	0.25	100
business ethics	0.63	0.40	0.39	0.38	0.45	100
high school psychology	0.84	0.59	0.54	0.56	0.63	545
conceptual physics	0.80	0.54	0.34	0.37	0.40	235
prehistory	0.87	0.59	0.50	0.51	0.55	324
high school chemistry	0.64	0.42	0.35	0.31	0.33	203
high school world history	0.76	0.53	0.47	0.55	0.61	222
college biology	0.81	0.44	0.42	0.46	0.45	144
high school physics	0.71	0.54	0.29	0.32	0.28	151
high school european history	0.78	0.58	0.50	0.56	0.59	147
college computer science	0.73	0.49	0.26	0.32	0.32	100
us foreign policy	0.86	0.56	0.49	0.57	0.72	100
moral disputes	0.88	0.50	0.36	0.37	0.50	346
world religions	0.83	0.52	0.46	0.54	0.69	171
high school statistics	0.78	0.54	0.33	0.33	0.27	216
international law	0.88	0.51	0.50	0.55	0.61	121
security studies	0.82	0.53	0.48	0.51	0.50	245
professional medicine	0.80	0.43	0.42	0.42	0.40	267
marketing	0.88	0.59	0.53	0.57	0.76	233
high school us history	0.74	0.49	0.41	0.47	0.66	202
sociology	0.87	0.60	0.57	0.60	0.74	201
anatomy	0.85	0.48	0.40	0.41	0.44	135
virology	0.83	0.56	0.39	0.39	0.46	166
professional psychology	0.87	0.49	0.38	0.39	0.47	612
miscellaneous	0.81	0.57	0.54	0.56	0.69	783
high school microeconomics	0.82	0.44	0.37	0.39	0.36	238
global facts	0.51	0.57	0.35	0.33	0.40	100
philosophy	0.87	0.52	0.42	0.46	0.53	311
college medicine	0.78	0.54	0.41	0.37	0.38	168
professional accounting	0.84	0.49	0.30	0.32	0.37	281

Table 6: Detailed results of LLaMA-2-chat on different categories of MMLU.

Category	Agreement(Label)	Agreement(Seq)	Acc.(Gen)	Acc.(Label)	Acc.(Seq)	Examples
moral scenarios	0.07	0.69	0.25	0.23	0.24	778
college physics	0.35	0.43	0.31	0.27	0.27	94
high school biology	0.68	0.53	0.51	0.51	0.65	302
college mathematics	0.40	0.47	0.29	0.25	0.33	93
abstract algebra	0.59	0.42	0.36	0.23	0.27	99
high school computer science	0.60	0.41	0.35	0.38	0.53	97
astronomy	0.59	0.57	0.48	0.44	0.57	143
computer security	0.53	0.48	0.46	0.61	0.66	98
logical fallacies	0.72	0.51	0.38	0.41	0.63	158
professional law	0.69	0.36	0.32	0.37	0.41	1446
clinical knowledge	0.64	0.51	0.51	0.54	0.59	255
elementary mathematics	0.25	0.36	0.41	0.26	0.32	363
high school macroeconomics	0.63	0.45	0.42	0.46	0.49	366
formal logic	0.56	0.28	0.34	0.39	0.26	108
high school government and politics	0.71	0.58	0.54	0.65	0.75	179
medical genetics	0.56	0.41	0.41	0.47	0.55	96
electrical engineering	0.55	0.50	0.44	0.42	0.52	135
high school mathematics	0.25	0.40	0.32	0.26	0.24	240
public relations	0.56	0.53	0.50	0.49	0.63	106
econometrics	0.68	0.52	0.30	0.26	0.23	108
machine learning	0.68	0.31	0.16	0.29	0.26	105
human sexuality	0.69	0.60	0.52	0.63	0.66	121
high school geography	0.68	0.56	0.55	0.54	0.69	182
nutrition	0.66	0.53	0.44	0.49	0.63	294
management	0.72	0.59	0.59	0.63	0.76	99
jurisprudence	0.63	0.43	0.39	0.49	0.66	103
human aging	0.60	0.44	0.38	0.46	0.56	211
college chemistry	0.55	0.51	0.38	0.43	0.45	88
business ethics	0.45	0.52	0.43	0.42	0.51	88
high school psychology	0.67	0.56	0.56	0.61	0.71	513
conceptual physics	0.59	0.51	0.38	0.36	0.40	230
prehistory	0.68	0.57	0.44	0.54	0.61	297
high school chemistry	0.54	0.47	0.32	0.37	0.46	191
high school world history	0.67	0.51	0.42	0.43	0.70	191
college biology	0.66	0.48	0.44	0.48	0.48	130
high school physics	0.49	0.41	0.34	0.34	0.30	146
high school european history	0.62	0.50	0.50	0.56	0.64	135
college computer science	0.52	0.42	0.27	0.38	0.36	96
us foreign policy	0.69	0.66	0.57	0.67	0.81	96
moral disputes	0.62	0.48	0.33	0.42	0.54	328
world religions	0.69	0.58	0.55	0.62	0.75	163
high school statistics	0.55	0.43	0.40	0.47	0.44	199
international law	0.52	0.48	0.48	0.48	0.71	108
security studies	0.84	0.58	0.41	0.49	0.64	222
professional medicine	0.59	0.42	0.52	0.53	0.53	257
marketing	0.74	0.63	0.56	0.65	0.77	226
high school us history	0.61	0.53	0.45	0.49	0.66	179
sociology	0.77	0.57	0.52	0.60	0.75	190
anatomy	0.66	0.47	0.37	0.45	0.49	133
virology	0.63	0.61	0.39	0.41	0.43	147
professional psychology	0.63	0.48	0.39	0.45	0.53	575
miscellaneous	0.69	0.61	0.58	0.59	0.73	752
high school microeconomics	0.72	0.43	0.45	0.48	0.53	220
global facts	0.30	0.42	0.37	0.23	0.32	99
philosophy	0.72	0.51	0.42	0.48	0.65	296
college medicine	0.62	0.51	0.46	0.48	0.51	162
professional accounting	0.59	0.30	0.32	0.36	0.40	266

Table 7: Detailed results of LLaMA-13B on different categories of MMLU.

Category	Agreement(Label)	Agreement(Seq)	Acc.(Gen)	Acc.(Label)	Acc.(Seq)	Examples
moral scenarios	0.29	0.47	0.32	0.24	0.27	893
college physics	0.74	0.57	0.24	0.27	0.27	100
high school biology	0.83	0.69	0.58	0.58	0.64	309
college mathematics	0.89	0.71	0.26	0.29	0.29	100
abstract algebra	0.41	0.63	0.34	0.26	0.29	99
high school computer science	0.82	0.64	0.48	0.47	0.55	99
astronomy	0.83	0.64	0.53	0.57	0.58	152
computer security	0.76	0.61	0.57	0.60	0.66	100
logical fallacies	0.68	0.65	0.56	0.59	0.69	162
professional law	0.81	0.72	0.37	0.39	0.40	1500
clinical knowledge	0.78	0.70	0.55	0.54	0.59	262
elementary mathematics	0.72	0.60	0.33	0.30	0.32	374
high school macroeconomics	0.82	0.73	0.44	0.46	0.50	389
formal logic	0.63	0.48	0.24	0.30	0.24	122
high school government and politics	0.90	0.75	0.63	0.65	0.76	193
medical genetics	0.72	0.63	0.47	0.54	0.58	100
electrical engineering	0.74	0.68	0.50	0.51	0.54	145
high school mathematics	0.74	0.59	0.27	0.24	0.27	266
public relations	0.79	0.69	0.53	0.54	0.63	110
econometrics	0.78	0.70	0.26	0.31	0.24	111
machine learning	0.58	0.74	0.32	0.42	0.33	111
human sexuality	0.85	0.73	0.55	0.57	0.64	131
high school geography	0.85	0.69	0.59	0.60	0.65	198
nutrition	0.81	0.65	0.51	0.52	0.61	305
management	0.79	0.71	0.57	0.63	0.69	103
jurisprudence	0.72	0.58	0.51	0.60	0.69	108
human aging	0.80	0.66	0.45	0.53	0.62	221
college chemistry	0.78	0.65	0.28	0.35	0.34	95
business ethics	0.72	0.68	0.49	0.52	0.54	100
high school psychology	0.84	0.76	0.63	0.65	0.72	542
conceptual physics	0.83	0.64	0.36	0.37	0.41	235
prehistory	0.82	0.71	0.52	0.53	0.63	323
high school chemistry	0.73	0.63	0.38	0.38	0.43	203
high school world history	0.71	0.72	0.61	0.68	0.75	218
college biology	0.81	0.65	0.44	0.47	0.58	144
high school physics	0.79	0.55	0.36	0.35	0.33	148
high school european history	0.83	0.69	0.55	0.63	0.67	144
college computer science	0.86	0.70	0.37	0.33	0.43	99
us foreign policy	0.88	0.83	0.71	0.73	0.81	100
moral disputes	0.84	0.70	0.48	0.52	0.60	345
world religions	0.87	0.77	0.69	0.70	0.77	171
high school statistics	0.79	0.60	0.35	0.34	0.34	216
international law	0.78	0.71	0.61	0.68	0.72	120
security studies	0.87	0.68	0.52	0.55	0.66	241
professional medicine	0.66	0.63	0.46	0.42	0.50	265
marketing	0.88	0.75	0.69	0.70	0.80	234
high school us history	0.71	0.69	0.58	0.64	0.74	200
sociology	0.86	0.73	0.65	0.71	0.75	201
anatomy	0.82	0.73	0.47	0.46	0.52	135
virology	0.74	0.62	0.37	0.44	0.47	165
professional psychology	0.78	0.68	0.47	0.51	0.54	610
miscellaneous	0.82	0.72	0.66	0.69	0.77	782
high school microeconomics	0.74	0.62	0.46	0.45	0.51	238
global facts	0.80	0.66	0.32	0.31	0.31	100
philosophy	0.83	0.72	0.55	0.55	0.65	310
college medicine	0.80	0.63	0.41	0.43	0.42	167
professional accounting	0.80	0.66	0.37	0.39	0.41	282

Table 8: Detailed results of LLaMA-13B-chat on different categories of MMLU.

Category	Agreement(Label)	Agreement(Seq)	Acc.(Gen)	Acc.(Label)	Acc.(Seq)	Examples
moral scenarios	0.64	0.98	0.24	0.25	0.24	878
college physics	0.31	0.50	0.31	0.21	0.44	96
high school biology	0.44	0.68	0.65	0.47	0.73	303
college mathematics	0.31	0.48	0.24	0.35	0.34	94
abstract algebra	0.26	0.48	0.40	0.19	0.30	96
high school computer science	0.41	0.55	0.53	0.47	0.64	92
astronomy	0.41	0.57	0.59	0.39	0.61	148
computer security	0.49	0.70	0.61	0.49	0.74	92
logical fallacies	0.50	0.70	0.66	0.48	0.75	159
professional law	0.36	0.58	0.39	0.30	0.44	1508
clinical knowledge	0.47	0.66	0.63	0.44	0.69	261
elementary mathematics	0.32	0.51	0.43	0.29	0.40	373
high school macroeconomics	0.37	0.59	0.51	0.35	0.59	384
formal logic	0.44	0.53	0.36	0.24	0.35	110
high school government and politics	0.50	0.71	0.74	0.53	0.84	191
medical genetics	0.52	0.61	0.61	0.52	0.69	100
electrical engineering	0.42	0.62	0.50	0.40	0.58	141
high school mathematics	0.30	0.44	0.34	0.27	0.35	250
public relations	0.54	0.60	0.58	0.42	0.66	106
econometrics	0.43	0.61	0.41	0.28	0.44	113
machine learning	0.26	0.37	0.38	0.31	0.48	108
human sexuality	0.45	0.64	0.62	0.47	0.75	130
high school geography	0.58	0.70	0.66	0.51	0.75	188
nutrition	0.46	0.63	0.60	0.46	0.70	301
management	0.55	0.70	0.66	0.43	0.80	100
jurisprudence	0.41	0.62	0.51	0.38	0.74	104
human aging	0.39	0.59	0.56	0.49	0.66	216
college chemistry	0.33	0.39	0.30	0.28	0.47	99
business ethics	0.32	0.60	0.53	0.35	0.58	96
high school psychology	0.52	0.75	0.73	0.48	0.78	530
conceptual physics	0.43	0.57	0.50	0.39	0.53	230
prehistory	0.43	0.71	0.59	0.39	0.71	318
high school chemistry	0.32	0.59	0.43	0.29	0.50	197
high school world history	0.33	0.59	0.63	0.46	0.79	212
college biology	0.41	0.67	0.57	0.41	0.67	141
high school physics	0.33	0.46	0.34	0.27	0.30	146
high school european history	0.33	0.69	0.57	0.36	0.77	143
college computer science	0.29	0.47	0.33	0.32	0.54	96
us foreign policy	0.59	0.79	0.78	0.60	0.84	100
moral disputes	0.41	0.64	0.56	0.38	0.68	338
world religions	0.52	0.81	0.75	0.53	0.81	165
high school statistics	0.35	0.55	0.38	0.30	0.46	207
international law	0.45	0.66	0.61	0.47	0.76	119
security studies	0.40	0.62	0.56	0.39	0.70	241
professional medicine	0.42	0.63	0.56	0.42	0.68	268
marketing	0.58	0.77	0.81	0.59	0.86	226
high school us history	0.34	0.63	0.63	0.39	0.76	197
sociology	0.49	0.79	0.69	0.54	0.86	200
anatomy	0.44	0.62	0.54	0.32	0.56	133
virology	0.47	0.66	0.51	0.34	0.52	161
professional psychology	0.48	0.65	0.56	0.39	0.61	604
miscellaneous	0.53	0.73	0.72	0.53	0.79	769
high school microeconomics	0.38	0.61	0.56	0.36	0.63	233
global facts	0.42	0.50	0.43	0.26	0.41	92
philosophy	0.43	0.67	0.61	0.37	0.69	289
college medicine	0.40	0.64	0.54	0.33	0.60	164
professional accounting	0.38	0.53	0.46	0.34	0.47	268

Table 9: Detailed results of Mistral-7B on different categories of MMLU.

Category	Agreement(Label)	Agreement(Seq)	Acc.(Gen)	Acc.(Label)	Acc.(Seq)	Examples
moral scenarios	0.08	0.02	0.28	0.23	0.24	894
college physics	0.34	0.48	0.26	0.16	0.29	100
high school biology	0.43	0.64	0.57	0.39	0.65	310
college mathematics	0.21	0.37	0.29	0.24	0.39	97
abstract algebra	0.11	0.27	0.34	0.17	0.33	99
high school computer science	0.36	0.60	0.51	0.42	0.50	100
astronomy	0.41	0.57	0.52	0.34	0.53	152
computer security	0.42	0.56	0.52	0.52	0.65	100
logical fallacies	0.50	0.69	0.59	0.47	0.71	163
professional law	0.37	0.56	0.34	0.30	0.40	1521
clinical knowledge	0.39	0.66	0.56	0.41	0.61	265
elementary mathematics	0.32	0.49	0.45	0.26	0.34	374
high school macroeconomics	0.35	0.56	0.44	0.28	0.51	389
formal logic	0.23	0.39	0.36	0.30	0.38	122
high school government and politics	0.51	0.68	0.60	0.44	0.72	193
medical genetics	0.46	0.59	0.52	0.51	0.63	100
electrical engineering	0.38	0.57	0.50	0.37	0.54	143
high school mathematics	0.36	0.38	0.27	0.22	0.30	256
public relations	0.49	0.73	0.51	0.34	0.57	110
econometrics	0.39	0.49	0.30	0.28	0.32	114
machine learning	0.21	0.30	0.29	0.33	0.46	112
human sexuality	0.47	0.64	0.56	0.46	0.69	129
high school geography	0.53	0.68	0.57	0.47	0.67	198
nutrition	0.43	0.59	0.49	0.40	0.63	306
management	0.51	0.68	0.60	0.47	0.74	103
jurisprudence	0.44	0.61	0.52	0.42	0.67	108
human aging	0.46	0.61	0.51	0.48	0.60	223
college chemistry	0.36	0.44	0.32	0.29	0.37	97
business ethics	0.40	0.52	0.52	0.39	0.58	100
high school psychology	0.51	0.71	0.65	0.47	0.72	545
conceptual physics	0.40	0.53	0.43	0.31	0.46	235
prehistory	0.40	0.66	0.54	0.39	0.58	323
high school chemistry	0.32	0.47	0.41	0.24	0.43	200
high school world history	0.40	0.62	0.57	0.47	0.75	223
college biology	0.40	0.60	0.51	0.35	0.60	144
high school physics	0.32	0.57	0.25	0.23	0.32	146
high school european history	0.37	0.67	0.56	0.33	0.67	147
college computer science	0.32	0.50	0.30	0.30	0.46	96
us foreign policy	0.56	0.75	0.63	0.57	0.76	100
moral disputes	0.47	0.66	0.53	0.40	0.59	345
world religions	0.52	0.67	0.59	0.52	0.69	171
high school statistics	0.38	0.51	0.38	0.25	0.41	213
international law	0.42	0.74	0.64	0.43	0.70	121
security studies	0.38	0.56	0.45	0.39	0.66	244
professional medicine	0.38	0.57	0.46	0.35	0.59	268
marketing	0.56	0.72	0.73	0.60	0.80	234
high school us history	0.40	0.59	0.52	0.41	0.72	202
sociology	0.52	0.76	0.67	0.52	0.78	201
anatomy	0.31	0.57	0.48	0.25	0.47	135
virology	0.39	0.58	0.36	0.33	0.42	166
professional psychology	0.46	0.62	0.48	0.37	0.50	611
miscellaneous	0.51	0.72	0.69	0.51	0.75	783
high school microeconomics	0.36	0.58	0.50	0.37	0.60	237
global facts	0.40	0.54	0.36	0.23	0.31	100
philosophy	0.49	0.66	0.52	0.36	0.60	311
college medicine	0.40	0.61	0.40	0.27	0.53	168
professional accounting	0.39	0.54	0.36	0.29	0.39	282

Table 10: Detailed results of Mistral-7B-Chat on different categories of MMLU.

	MMLU		Truthful-QA		Belebele	
Model	Label-Gen	Seq-Gen	Label-Gen	Seq-Gen	Label-Gen	Seq-Gen
Mistral-7B	-14.3	6.9	-15.1	-14.0	9.0	-4.1
Mistral-7B-Instruct	-11.0	6.3	-11.5	-8.5	6.7	4.5
LLaMA-1-7B	-12.3	-8.1	4.9	16.4	-4.7	-4.0
Vicuna-7B	-4.6	11.6	-2.2	9.9	4.0	16.5
LLaMA-2-7B	-0.8	9.0	21.8	6.6	3.3	-6.4
LLaMA-2-7B-chat	1.3	6.3	-4.8	-33.5	6.1	1.1
LLaMA-2-13B	2.9	10.6	-5.4	-26.7	6.6	2.5
LLaMA-2-13B-chat	1.5	6.2	-4.8	-22.3	8.2	7.3
LLaMA-2-70B	2.1	7.3	-4.8	-26.8	6.2	-2.0
LLaMA-2-70B-chat	1.1	5.7	-19.8	-21.0	2.6	2.0

Table 11: Differences in label and sequence accuracies compared to generation accuracies across datasets.

anonymous LLMs, in which the user has to vote for the better one, which will be the winner LLM. Based on several win-lose interactions, we can then calculate the Elo score.

Elo scores have been previously designed in rank multiple players that involve multiple matches across different people, such as chess. It is good for determining a unified ranking across every player (in this case, LLMs). From the Elo score of 2 players, we can predict the winning chance of both players. For example, an LLM with an Elo of 1200 will win against an LLM with an Elo of 900 85% of the time.

Chatbot Arena is one of the popular Elo-based leaderboards. It supports a variety of LLMs, both proprietary and open-sourced, and has accumulated hundreds of thousands of votes.