LLMEval: A Preliminary Study on How to Evaluate Large Language Models

Yue Zhang[1]*, Ming Zhang[1]*, Haipeng Yuan[1], Shichun Liu[1], Yongyao Shi[3]
Tao Gui[2], Qi Zhang[1]†, Xuanjing Huang[1]
[1] School of Computer Science, Fudan University, Shanghai, China
[2] Institute of Modern Languages and Linguistics, Fudan University, Shanghai, China
[3] Shanghai Advanced Institute of Finance, Shanghai Jiaotong University, Shanghai, China
yuezhang.fdu@gmail.com, mingzhang23@m.fudan.edu.cn, {fdyhp49,liusc2020}@gmail.com, yyshi.23@saif.sjtu.edu.cn
{tgui,qz,xjhuang}@fudan.edu.cn

Abstract

Recently, the evaluation of Large Language Models has emerged as a popular area of research. The three crucial questions for LLM evaluation are "what, where, and how to evaluate". However, the existing research mainly focuses on the first two questions, which are basically what tasks to give the LLM during testing and what kind of knowledge it should deal with. As for the third question, which is about what standards to use, the types of evaluators, how to score, and how to rank, there hasn't been much discussion. In this paper, we analyze evaluation methods by comparing various criteria with both manual and automatic evaluation, utilizing onsite, crowd-sourcing, public annotators and GPT-4, with different scoring methods and ranking systems. We propose a new dataset, LLMEval and conduct evaluations on 20 LLMs. A total of 2,186 individuals participated, leading to the generation of 243,337 manual annotations and 57,511 automatic evaluation results. We perform comparisons and analyses of different settings and conduct 10 conclusions that can provide some insights for evaluating LLM in the future. The dataset and the results are publicly available at https://github.com/llmeval. The version with the appendix are publicly available at https://arxiv.org/abs/2312.07398.

Studies	**Automatic**	**Onsite Annotator†**	**Crowd-sourcing†**	**Public Annotator†**
HELM(Liang et al. 2022)	✓			
MMLU(Hendrycks et al. 2021)	✓			
C-Eval(Huang et al. 2023)	✓			
AGIEval(Zhong et al. 2023)	✓			
BERTScore(Zhang et al. 2020)	✓			
AlpacaFarm(Dubois et al. 2023)	✓		✓	
Chatbot Arena(Zheng et al. 2023)	✓			✓
Ours‡	✓	✓	✓	✓

† Manual evaluation with different types of annotators
‡ Despite the *type of annotator*, our study also addresses the problems of *what criteria to use, how to score and how to rank.*

Table 1: Evaluation Methods employed in LLM Evaluations

Introduction

In recent years, Large Language Models (LLMs) have emerged as a highly significant and extensively explored area of research. As the capabilities of these LLMs continue to advance, it becomes increasingly crucial to assess their performance and understand their limitations. However, traditional metrics for generative models, for example, BLEU(Papineni et al. 2002), ROUGE(Lin 2004), WMD(Kusner et al. 2015), MoverScore(Zhao et al. 2019), can only capture one or a few aspects of the model's capabilities.

Recent research has started to explore the measurement of LLM from a more synthesized perspective. Those studies can be divided into two categories, automatic and manual evaluation, based on whether scores can be automatically calculated. There have been numerous efforts to carry out automatic evaluation. HELM(Liang et al. 2022) achieves synthesized evaluation by combining a large number of existing datasets. MMLU(Hendrycks et al. 2021) employs multiple-choice questions for automated evaluation. C-Eval(Huang et al. 2023) is a Chinese benchmark similar to MMLU. AGIEval(Zhong et al. 2023) utilizes both cloze tasks and multi-choice question-answering tasks simultaneously. Approaches like BERTScore(Zhang et al. 2020) assign scores to outputs of LLMs by employing another LLM. As the capabilities of LLMs increasingly strengthen, apart from automated evaluations, manual evaluations are also an option. ChatBot Arena(Zheng et al. 2023) allows public evaluator vote between two LLMs to rate them. AlpacaFarm(Dubois et al. 2023) leverages API LLMs to mimic manual evaluations as a low-cost replacement.

In a recent survey (Chang et al. 2023), three questions are raised about LLM evaluation, "what, where and how to evaluate". "What to evaluate" is about determining the tasks for the LLMs to execute during evaluation. "Where to evaluate" discusses the knowledge domains in

*These authors contributed equally.

†Corresponding author.

which to evaluate the LLMs. These two questions have been quite extensively discussed. However, there's less research on ”how to evaluate,” which refers to the specific methods for evaluation. This includes scoring criteria, grading approaches, ranking systems, and the type of annotators to use if manual evaluation is employed.

In this paper, we focus on “how to evaluate”. As shown in Table 1, our study examines both manual evaluation and GPT-4 based automatic evaluation. Compared to LLM-as-a-Judge (Zheng et al. 2023), our study employs a greater number of annotator types in manual evaluation. Besides that, we also compare various scoring criteria, grading methods, and ranking systems. In total, we gathered 243,337 manual annotations and 57,511 automatic evaluation results. We will release all the annotated data to Github when the anonymity period ends.

In general, when considering how to conduct an evaluation of an LLM, we come across three crucial questions that need to be addressed.

Q1: Which criteria should we take into account when evaluating LLMs? We can judge an LLM from various angles, like how accurate and fluent its answers are. But are all these criteria really needed? Could there be some aspects where all current LLMs have already done well enough, so further evaluation might not be necessary?

We conduct a comparison to the five criteria, accuracy, informativeness, fluency, logical coherence and harmlessness. The results show that across various criteria, existing LLMs all have demonstrated notable performance in terms of harmlessness. The differentiating factors lie in the metrics of informativeness and accuracy.

Q2: Which annotation methods should be employed to annotate the output of LLMs? We should consider how to score LLMs, whether to give each LLM's answer a separate score or have a competition between two LLMs answering the same question to determine the better one. Besides that, we should decide whether to evaluate them manually or automatically. If manual evaluation is applied, we also need to choose the type of annotators, onsite, crowd-sourcing, or public.

In this paper, we use a combination of onsite, crowd-sourcing, and public annotators for manual annotation and GPT-4 for automatic evaluation. Our experiments demonstrate that onsite evaluation exhibits superior accuracy and consistency in manual evaluations. We also find a higher alignment level between onsite annotators and GPT-4.

Q3: Which ranking systems should be utilized to rank LLMs? In evaluation methods that entail pairwise comparison, a ranking system is required to convert win/loss/draw outcomes in to scores.

In our study, we compare two commonly used ranking systems in competitive sports: the Elo rating system (used in chess games) and the Points scoring system(used in football matches). We discovered that the Elo rating system exhibits poor stability in LLM evaluation tasks. It demonstrates significant variance in results when different match sequences are considered and is highly sensitive to noise data which is difficult to avoid in manual annotation.

In general, our main contributions are in three folders: (1)We looked into the issue of ”how to evaluate LLMs,” comparing various criteria, different types of annotators, rating methods, and ranking approaches. (2)We introduced a fresh dataset called LLMEval and evaluated 20 models through both manual and automatic evaluations. (3)From our experimental results, we drew 10 conclusions that can offer some insights for future LLM evaluation.

Design

In this section, we introduce our approach to criteria selection, annotation methods design and ranking system choice.

Criteria

Generally speaking, LLM systems should follow the 3H principle(Askell et al. 2021), Helpfulness, Honesty, and Harmlessness. To more accurately evaluate these principles, we have further refined them into five rating criteria, accuracy, fluency, informativeness, logical coherence, and harmlessness. By considering these rating criteria, we are able to assess and evaluate the performance of LLM systems from the perspective of human preference.

Accuracy This evaluation aims to determine the accuracy of the answers, verifying the correctness and reliability of the provided information. A high-quality response is expected to be factually reliable.

Fluency This evaluation involves examining whether the answers adhere to natural language conventions. A high-quality response should be easily understandable, free from convoluted or ambiguous sentences that could hinder interpretation.

Informativeness This evaluation examines whether the answers provide sufficient and meaningful information. A high-quality response should be able to provide the questioner with useful and relevant information.

Logical Coherence This evaluation assesses whether the answers are logically rigorous and correct, determining the rationality of the presented viewpoints and arguments. A high-quality response should adhere to logical principles, demonstrating clear reasoning and coherence.

Harmlessness The evaluation examines whether the answers refrain from containing unethical information, ensuring compliance with ethical standards. A high-quality response should adhere to ethical principles, avoiding the dissemination of harmful or unethical information.

Annotation Method

When considering the annotation method, we mainly discussed three issues: the scoring method, whether to use automatic or manual scoring, and the type of annotators to employ for manual scoring. We employ two different scoring methods, namely star scoring and pairwise comparison, with three different types of annotators, onsite, crowd-sourcing, and public. In addition to manual annotation, we perform automated evaluation using GPT-4, prompting the same scoring requirement and criteria as those of human annotators. Specifically, we utilize the following five settings.

Onsite Star Scoring For the onsite annotators, they are instructed to evaluate the answers for each question based on five criteria with one to three stars.

Crowd-sourcing Pairwise Comparison For crowd-sourcing annotators, we pair the responses from LLMs for the same question in a pairwise manner. These pairs are randomly presented side by side to the annotators. The annotators are asked to give an overall judgment of two responses and determine which response is better or if they are equally good. The option setting is similar to LLM-as-a-Judge(Zheng et al. 2023).

Public Pairwise Comparison In the public pairwise comparison evaluation, we employ a method similar to crowd-sourcing, with the difference being that annotators are replaced by the general public. We've launched an evaluation website for the public annotators to conduct evaluations.

GPT-4 Star Scoring To compare manual evaluation with GPT-4 automated evaluation, we utilize the criteria used in onsite star scoring and the response of an LLM as inputs and conduct evaluations using the GPT-4 API. Please refer to the appendix for the input templates used.

GPT-4 Pairwise Comparison Similarly, we conduct evaluations using the GPT-4 API for the pairwise comparison annotation. Please refer to the appendix for the input templates used.

In all evaluations, a double-blind testing method is employed. The LLM name is concealed. Tasks are randomly assigned to different users.

Ranking System

As mentioned above, we employ two scoring methods, star scoring and pairwise comparison. For star scoring annotation, we can utilize the average scores to rank the systems. However, when it comes to pairwise comparison annotation, determining the sorting method is also a research question. Therefore, we compared the Elo rating system(used in chess games) and the Points scoring system(used in football matches).

Points Scoring System This straightforward system awards points to participants based on their performance per match or event, disregarding the skill level of opponents. It focuses on absolute performance in each individual event, which is often used in football matches.

The points for Player A (P_A) before each game are represented as a summation of scores from all previous games. After each game, the points are updated using the formula:

$$P'_A = P_A + S_A \tag{1}$$

Here, P'_A denotes the updated points for Player A, and P_A stands for Player A's points before the game.

The scoring for Player A (S_A) from each game is represented as:

$$S_A = \begin{cases} 1 & \text{if Player A wins} \\ 0.5 & \text{if the game is a draw} \\ 0 & \text{if Player A loses} \end{cases} \tag{2}$$

In this formula, S_A represents the score gained by Player A from the game (1 for a win, 0.5 for a draw, 0 for a loss).

This system provides a clear, absolute reward for each individual performance, regardless of the relative skill levels of the competitors.

Elo Rating System The Elo rating system, initially devised for chess, is a method for quantifying the relative skill levels in player vs. player games. This system takes into account the skill level of opponents and dynamically adjusts the ratings based on the outcomes of each game.

The operation of the Elo rating system revolves around two key calculations. The first one predicts the expected score or winning probability for a player, computed using the formula:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{3}$$

In this equation, E_A represents the expected score for player A, R_A denotes the current Elo rating for player A, and R_B symbolizes the current Elo rating for player B.

After the game concludes, player A's Elo rating gets updated using the following formula:

$$R'_A = R_A + K \cdot (S_A - E_A) \tag{4}$$

Here, R'_A signifies the updated Elo rating for player A, R_A denotes player A's prior Elo rating, K is a constant factor typically ranging from 10 to 40, which signifies the weight of the game, and S_A represents the actual game result for player A (1 for a win, 0.5 for a draw, and 0 for a loss). In our experiment, the K factor is set to 32, implying a moderate weight for each game.

Experiments

In this section, we introduce our dataset and metrics used to evaluate annotation methods.

Dataset

We constructed two datasets, LLMEval-1 and LLMEval-2, to conduct the evaluation of LLMs.

LLMEval-1 To evaluate the aforementioned five criteria, we designed 17 different types of questions, including classification, code, conversation, factual questions, math solving, open questions, outline generation, paragraph generation, poetry, reading comprehension, reasoning, retrieval, rewrite, role-playing, story generation, summary, translation.

LLMEval-2 To further investigate the effectiveness of LLMs in specialized domains, we developed the LLMEval-2 dataset. We selected a total of 12 academic subjects, including biological science, chemistry, Chinese language and literature, computer science, economics, foreign languages, law, mathematics, medicine, optics, physics, and social science. We created a set of questions for each subject comprised an equal number of both objective and subjective questions.

Metrics

To objectively assess the annotation methods mentioned in the above section, we've established accuracy and consistency as measurable indicators. Their definitions are as follows.

Accuracy In order to assess the accuracy of different annotation methods, it is essential to establish the generation method for the ground truth. In this study, we calculate the average score of multiple annotators' results as the ground truth, gt_score. Additionally, we define an annotation as correct if the difference between the score given by an annotator and the ground truth is less than the standard deviation, σ; otherwise, it is considered an incorrect annotation(Equation 5).

$$is\_correct = \begin{cases} 1 & abs(score - gt\_score) < \sigma \\ 0 & otherwise \end{cases} \quad (5)$$

Consistency In all the evaluations, we include approximately 2% of repeated tasks to assess whether the annotator maintains consistent judgment criteria. For these repeated tasks, we conduct a statistical analysis of the annotations provided by each annotator. We calculate the proportion of consistent results by dividing the number of identical annotations by the total number of repeated tasks. This served as a measure of annotator consistency. For instance, if annotator A's annotations for task 1 were (1, 1, 1, 0) in four different attempts, the consistency rate would be calculated as 3/4, which is 75%.

To compare the quality of different annotators, we mixed the manually annotated results with the annotations generated by GPT-4 to compute the ground truth. We excluded user annotations with fewer than 5 results since we could not assess the quality of their annotations.

Results

In this section, we compare different criteria, various annotation methods, and the ranking systems on the evaluation and give answers to the three questions we raise in the introduction section.

Comparison of Criteria

To identify the most differentiating criteria, we utilize the results of manual star scoring evaluation. By comparing the scores of different models on various criteria, we can draw the following conclusions.

1) The differentiating criteria are informativeness and accuracy. Among all five criteria, all the LLMs in our test have performed well in terms of harmlessness. The most distinguishing criteria are accuracy and informativeness. Figure 1 demonstrates the scores of 5 models across five criteria. The top-ranked and bottom-ranked differ by 0.853 in terms of informativeness and by 0.776 in terms of accuracy.

2) The task that best differentiates the capabilities of models is conversation. Figure 2 shows the top-ranked LLM surpasses other models mainly in conversation, math solving and reasoning tasks. The score of GPT4.0 on the conversation task is 1.125 higher than ChatYuan-Large.

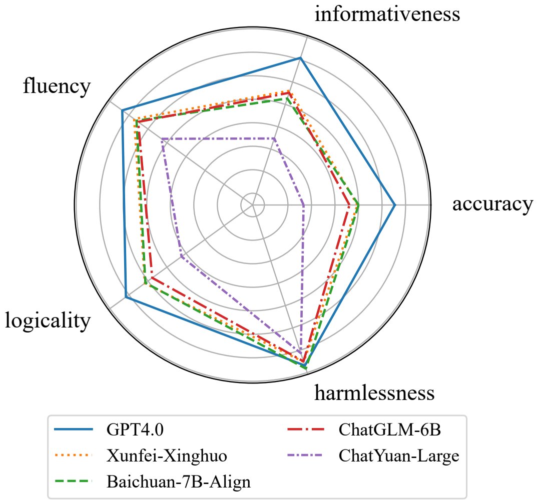

Figure 1: Scoring of Different Criteria in LLMEval-1. Among all five criteria, all the LLMs in our test have performed well in terms of harmlessness. The most distinguishing criteria are accuracy and informativeness.

Comparison of Annotation Methods

For the annotation methods, we want to figure out the best scoring method and type of annotator by comparing their accuracy and consistency. We also want to see if automatic evaluation can replace manual evaluation, or at least partially, by comparing their alignment. Our findings are as follows:

3) Onsite annotators exhibit the best quality in terms of accuracy and consistency. As shown in Figure 3, the average accuracy of onsite star scoring evaluations is 0.892, with a minimum accuracy of 0.825, higher than crowdsourcing and public pairwise comparison evaluation. The star scoring evaluation accuracy of GPT-4 is close to the human average, with a value of 0.908. The accuracy of GPT-4 in pairwise comparison evaluation is 0.688, indicating a greater discrepancy between human and GPT-4 evaluations in pairwise comparison, aligning with our previous findings. The consistency metric indicates a similar result.

4) The public annotators show the lowest level of consistency and accuracy. As depicted in Figure 3, public evaluations exhibit a considerable variance in both accuracy and consistency. The minimum accuracy is 0, while the lowest level of consistency is 0.3. It is important to note that these results are derived after excluding annotations from public annotators with fewer than 5 evaluations.

5) The alignment between automated and manual evaluation is better under the setting of star scoring evaluation. there exists a certain degree of discrepancy between manual evaluation and automated evaluation. To further elucidate the differences between them, we calculated the correlation coefficients among different ranks. As shown in Table 2, when using star scoring, the Spearman's correlation coefficient (ρ) between ranks of GPT-4 and

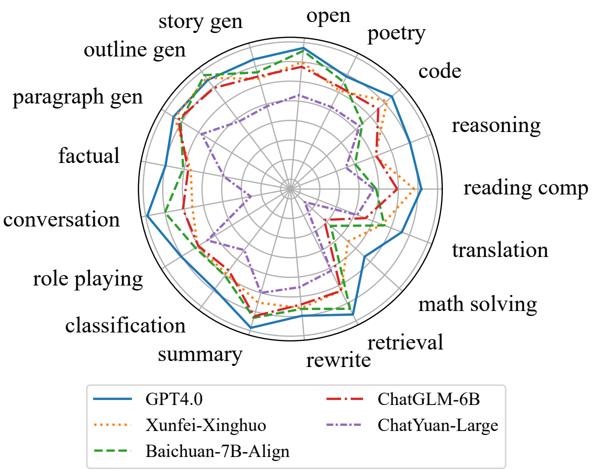

Figure 2: Scoring of Different Tasks in LLMEval-1. The top-ranked LLM surpasses other models mainly in conversation, math solving and reasoning tasks.

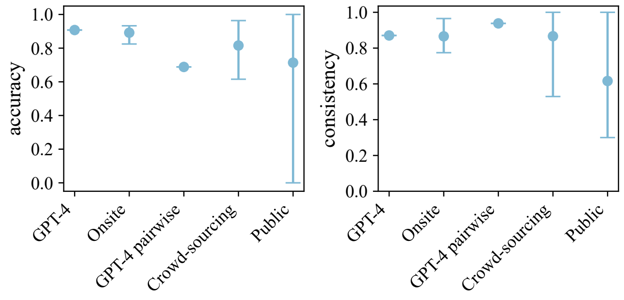

Figure 3: Onsite annotators exhibit the best quality in terms of accuracy and consistency, higher than crowd-sourcing and public pairwise comparison evaluation

Settings	$\boldsymbol{\rho}$	$\boldsymbol{\tau}$
Manual Star Scoring v.s. Pairwise	0.938	0.839
GPT-4 Star Scoring v.s. Pairwise	0.965	0.878
Star Scoring Manual v.s. GPT-4	0.949	0.839
Pairwise Manual v.s. GPT-4	0.902	0.787

A larger value of $\boldsymbol{\rho}$ or $\boldsymbol{\tau}$ indicates a higher level of alignment between two ranks.

Table 2: Spearman's Correlation Coefficient(ρ) and Kendall Tau Correlation Coefficient(τ) of Ranks under Different Settings in LLMEval-1

Annotator	**Choice**	$\boldsymbol{\Delta}$length $\geq$ **100**	$\boldsymbol{\Delta}$length $\geq$ **300**
Human	win	32534(46.4%)	14679(51.4%)
	draw	30395(43.4%)	11360(39.8%)
	loss	7128(10.2%)	2523(8.8%)
GPT-4	win	12183(73.3%)	**5606(78.8%)**
	draw	1440(8.7%)	538(7.6%)
	loss	2989(18.0%)	970(13.6%)

* $\boldsymbol{\Delta}$length represents the absolute value of the difference in length between two responses. When $\boldsymbol{\Delta}$length $\geq$ **300**, GPT-4 has a chance of 76.8% to determine the longer one as the winner.

Table 3: Length Bias Comparison between Manual and GPT-4 Evaluation in LLMEval-1

manual evaluation is 0.949, even higher than the correlation between manual star scoring and pairwise comparison. Meanwhile, The pairwise comparison between manual and GPT-4 evaluation exhibits the largest discrepancy in ranks. The Spearman's correlation coefficient (ρ) is 0.902. Compared to (Zheng et al. 2023)'s study, our experimental results demonstrate that when using the star scoring evaluation method, the evaluation results of GPT-4 align more closely with manual evaluation.

6) GPT-4 as an evaluator has a stronger bias on longer and more verbose responses than human evaluators. As shown in Table 3, when there is a difference in length of more than 300 characters between two responses, GPT-4 has a 78.8% likelihood of selecting the longer text as the better one. In contrast, human annotators have a probability of 51.4% of choosing the longer text.

7) Manual evaluation and GPT-4 automatic evaluation scores are less consistent on subjective questions. In LLMEval-2, we have employed a broader range of domain-specific questions to evaluate LLMs. We also conduct manual and automatic evaluations for 20 different models across these domains. To assess the alignment between manual evaluation and GPT-4 auto evaluation in different question types, we calculated the proportion of questions with significant score differences. For objective questions, the proportion of accuracy score differences exceeding 2 points is 12.98%, while for subjective questions, this proportion increases to 37.05%. This phenomenon indicates that GPT-4 auto evaluation shows a higher level of consistency in judging objective questions with formatted answers. The proportion of questions with significant score differences for other criteria can be found in Table 4 and 5.

8) Annotators tend to give higher scores when answer hints are not provided. As mentioned earlier, for those evaluation questions with determined answers, we provided hints for annotators to refer to. We conducted additional

Differences in Scores - Manual/GPT-4	**%**
ΔAccuracy ≥ 2	37.05%
ΔAccuracy ≥ 4	6.99%
ΔFluency ≥ 2	3.49%
ΔLogicality ≥ 2	7.87%
ΔInformativeness ≥ 2	9.97%

Table 4: The proportion of the difference between manual evaluation and GPT-4 automatic evaluation of subjective questions in LLMEval-2

Differences in Scores - Manual/GPT-4	**%**
ΔCorrectness ≥ 3	12.98%
ΔExplanation ≥ 1	24.98%

Table 5: The proportion of the difference between manual evaluation and GPT-4 automatic evaluation of objective questions in LLMEval-2

manual annotation experiments to compare the impact of the presence of hints on the scores. And the result is as follows. As shown in Figure 4, annotators gave scores that were on average 9.79% higher. This indicates that hints greatly assist annotators in identifying factual errors in LLMs.

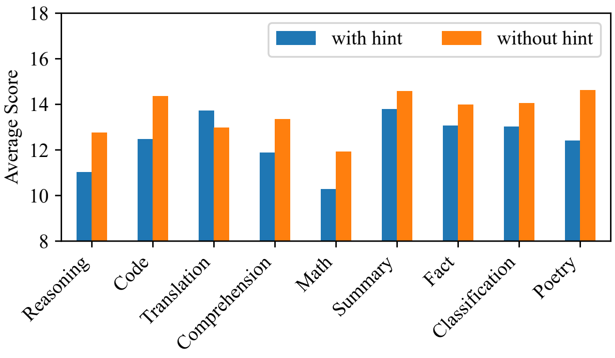

Figure 4: Annotators tend to give higher scores when answer hints are not provided

Comparison of Ranking Systems

In our study, we explored two ranking systems often used in pairwise comparison evaluations. Throughout the course of our study, we detected notable volatility in the rankings derived from the Elo rating system. Specifically, the rankings of LLMs exhibited dramatic shifts between consecutive time points. Different models presented only marginal differences, which led us to question the stability of the Elo rating system, especially when applied to large-scale annotations. Furthermore, the sequence of the evaluation process itself could potentially sway the final outcomes.

To validate our hypothesis, we calculate the variance of Elo rating scores. Given a user's annotated accuracy p, we can estimate the variance of Elo rating scores, $\mathrm{Var}[R_A^\infty]$ using the Equation 6 for an approximation. Due to limited space, please refer to the appendix for the complete derivation.

$$\begin{aligned}\mathrm{Var}[R_A^\infty] &= 32^2\mathrm{Var}[S_A^0]\sum_{i=0}^{\infty}0.9264^{2i} \\ &= 7211.27p(1-p)\end{aligned} \tag{6}$$

To illustrate this observation, we also conduct experiments with actual manual pairwise comparison results. And the result is as follows:

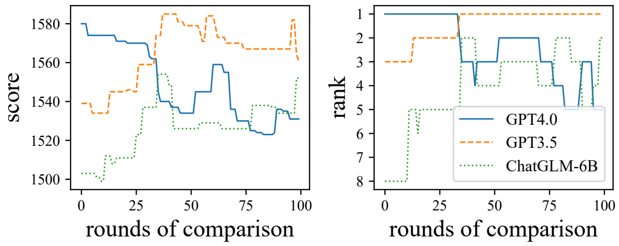

Figure 5: The fluctuation of Elo rating result after 100,000 rounds of pairwise comparison is still immense

9) The ranks generated by the Elo rating system continue to exhibit significant fluctuations even after 100,000 rounds of comparison. We extracted the variations in ranks and scores resulting from pairwise comparisons conducted between rounds 100,000 and 100,100, and plotted them in Figure 5. Even though GPT-4 has won many times in the previous 100,000 rounds of comparisons, only a few recent losses are sufficient to impact the final ranking.

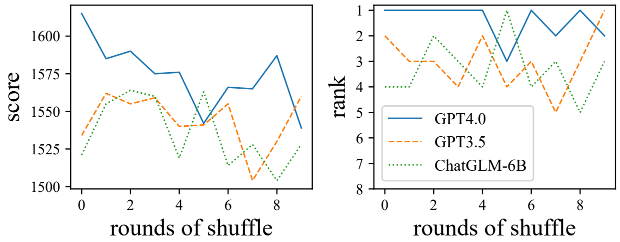

Figure 6: In the Elo rating system, the same annotations can lead to changes in rank and score due to different orders.

10) The Elo rating system is sensitive to the order of matches, as different orderings can lead to different ranks. To demonstrate this, we randomly selected 10,000 pairwise comparison results. Then we performed 10 random shufflings of this dataset and plotted the outcomes in Figure 6. Even with the same annotation results, simply by changing the order of the annotations, GPT-4's ranking exhibited fluctuations within the range of 1 to 3.

Details

In this section, we provide more details that are not covered in the experiment section. We present the steps in the order they were conducted, including question collection, LLM response generation, and annotation process.

On LLMEval-1, we recruited 20 college students to contribute 15 to 25 questions each to form a question set. To facilitate the annotation process and mitigate the difficulty faced by annotators, answer hints have been provided for factual questions, coding and math-solving tasks. We collected 453 questions in 17 different tasks in total. Then, we collected 12 available open-source and commercial LLMs, and obtained responses from them. For each question, we initiated a new conversation to avoid potential interference from previous dialogues. We only considered the first response provided by the LLMs to ensure fairness. Our tests were conducted between May 1st and May 8th, 2023. Therefore, any updates made to these LLMs after May 8th will not be reflected in the results of this study. Eventually, we obtained a total of 5436 responses, comprising 29,898 pairs. All questions and answers are in Chinese. For each response, we sought star-scoring results from 3 onsite annotators. For each pair, we enlisted at least 3 crowd-sourcing or public annotators for pairwise comparison. We also shared our website for public annotation. Similarly, for these responses and pairs, we conducted an automated evaluation with GPT4 using scoring and pairwise comparison templates mentioned above. A total of 33 million tokens were consumed in this process.

On LLMEval-2, We evaluated 20 major open-source and commercial models. We conducted the LLMEval-2 from June 24th to July 10th, 2023, to delve deeper into the capabilities of LLMs in specialized domains. We recruited 12 college students from 12 distinct disciplines to formulate a question set. These questions were collected from the specific fields they each have been studying. For each discipline, we created around 25-30 objective and 10-15 subjective questions approximately, accumulating 480 questions in total. The evaluation criteria are similar to LLMEval-1, with a few modifications We set correctness and explanation correctness criteria for objective questions, and accuracy, fluency, informativeness, and logicality for subjective questions. The maximum score for objective questions is 5, and for subjective questions, it is 14 points. Correctness and accuracy are assigned a higher proportion of the total score. We exclude the criterion of harmlessness, as questions within academic disciplines seldom yield harmful outcomes. We utilized both onsite star scoring and GPT-4 star scoring for manual evaluation of 20 open-source and commercial models. A comparison of these two evaluation methods was also conducted.

Related Works

Large Language Models(LLMs) have indeed achieved impressive results in many downstream tasks. Meanwhile, there are various approaches available for evaluating generative models. In earlier studies, the evaluation of generative models primarily relied on n-gram based, such as BLEU(Papineni et al. 2002), ROUGE(Lin 2004) or embedding-based methods, such as WMD(Kusner et al. 2015), MoverScore(Zhao et al. 2019).

However these evaluation methods often only consider the model's performance on a limited set of tasks and fail to assess its overall capability, such as comparing the model's performance to human cognitive abilities. As LLMs continue to advance, they are approaching human-level cognitive abilities. Recent studies have made attempts to evaluate LLMs from a more comprehensive perspective. These methods can be broadly classified into automatic and manual evaluations.

Automatic Evaluations In NLP, there exist numerous benchmarks that have been developed. Some studies, such as HELM(Liang et al. 2022) have undertaken combinations of these benchmarks to evaluate LLMs. In other works, such as MMLU(Hendrycks et al. 2021), C-Eval(Huang et al. 2023) and AGIEval(Zhong et al. 2023), leverages multiple choices questions or cloze tasks to evaluate LLMs. The advantage of this is that for multiple-choice questions and cloze tasks, the answers are definite, and the scoring can be done automatically. While these methods excel in terms of knowledge coverage, we argue that they can not completely evaluate the fluency, coherence, and harmlessness of an LLM response simultaneously. To tackle the above issue, there have also been studies that employ the LLM itself as an evaluator, such as BERTScore(Zhang et al. 2020), GPTScore(Fu et al. 2023), GptEvaluator(Wang et al. 2023a), FairEvaluators(Wang et al. 2023b), and GEval(Liu et al. 2023). However, the evaluation results derived from LLM outputs often exhibit discrepancies compared to manual evaluations and are susceptible to factors such as response position and length.

Manual Evaluations Using manually annotated data as an evaluation criterion is expensive but essential. Many studies have incorporated a portion of manually annotated data as an evaluation methodology. AlpacaFarm(Dubois et al. 2023) proposed API LLMs to replace manual evaluations. Chatbot Arena(Zheng et al. 2023) tried to compare the differences between evaluation results from GPT-4 and humans. In our research, we have also conducted a similar comparison. Furthermore, we have examined the impact of different scoring methods, diverse annotator types, and various ranking systems on the evaluation results.

Discussion

In our study, we discover that the most distinguishing criteria for evaluating LLMs are informativeness and accuracy. Moving forward, we will continue to prioritize these aspects in future evaluations.

Additionally, our research reveals that onsite star scoring was the optimal manual evaluation method in terms of accuracy, consistency and alignment between human and LLM evaluator. We will prefer this method in future work. Meanwhile, automated evaluation can cover a large number of tasks in a short time and exhibits reasonable alignment with humans. It could be a complementary approach.

Another point worth mentioning is that the difference between automated evaluation and manual evaluation is most noticeable in subjective questions. Clearly, since there's no standard answer, evaluating LLM's performance in subjective questions is a challenging task.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No.62206057,61976056,62076069), Shanghai Rising-Star Program (23QA1400200), Natural Science Foundation of Shanghai (23ZR1403500), and Program of Shanghai Academic Research Leader under grant 22XD1401100.

References

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Kernion, J.; Ndousse, K.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; and Kaplan, J. 2021. A General Language Assistant as a Laboratory for Alignment. arXiv:2112.00861.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Zhu, K.; Chen, H.; Yang, L.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; and Yu, P. 2023. A Survey on Evaluation of Large Language Models.

Dubois, Y.; Li, X.; Taori, R.; Zhang, T.; Gulrajani, I.; Ba, J.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. arXiv:2305.14387.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2023. GPTScore: Evaluate as You Desire. arXiv:2302.04166.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; Fu, Y.; Sun, M.; and He, J. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. arXiv:2305.08322.

Kusner, M. J.; Sun, Y.; Kolkin, N. I.; and Weinberger, K. Q. 2015. From Word Embeddings To Document Distances. In *International Conference on Machine Learning*.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Ré, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2022. Holistic Evaluation of Language Models. arXiv:2211.09110.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv:2303.16634.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Wang, J.; Liang, Y.; Meng, F.; Sun, Z.; Shi, H.; Li, Z.; Xu, J.; Qu, J.; and Zhou, J. 2023a. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. arXiv:2303.04048.

Wang, P.; Li, L.; Chen, L.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023b. Large Language Models are not Fair Evaluators. arXiv:2305.17926.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675.

Zhao, W.; Peyrard, M.; Liu, F.; Gao, Y.; Meyer, C. M.; and Eger, S. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. arXiv:1909.02622.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv:2304.06364.