Performance Comparison Analysis of Large Language Models Based on Objective Ability Testing and Subjective User Experience

Yitong Wang[1*]
[1]Yuci Longhu School, 030600 Jinzhong city, China

Abstract. At present, large language models have profoundly transformed many fields, including human-computer interaction, content creation, and code generation. They demonstrate astonishing, even human-like abilities in various tasks. At the same time, it has also brought about huge risks and social concerns, such as the existence of prejudice, the leakage of private data, and the abuse of false information dissemination. Therefore, it is necessary to conduct a comprehensive and objective assessment of large language models. This article will combine relevant data and analyze the main language models from both objective ability testing and subjective user experience aspects through some evaluation methods. In the collected articles, the respective data and limitations have all been explained. All have undergone in-depth research and reached a conclusion. These articles provide data and conclusion support for this paper, which can be summarized to draw new conclusions. The performance evaluation of large language models is a complex study. This article systematically reviews the main viewpoints and methodological analyses of this study. Each method has its own advantages and limitations. In the future, more comprehensive, authoritative and all-round evaluation methods should be studied. This can provide a framework reference for future research.

1 Introduction

At present, there are already many well-known large language models, such as GPT-4, Gemini, and LLaMA. These models have profoundly transformed numerous fields including human-computer interaction, content creation, and code generation. Their astonishing and even human-like capabilities demonstrated in various tasks. At the same time, they also bring huge risks and social concerns, such as the existence of prejudice, the leakage of private data, and the abuse of false information dissemination, etc.

Large language models can apply deep neural network models to describe the vector representation and generation probability of massive texts, thereby effectively expressing the lexical, syntactic and semantic features of language. Their unique advantages in tasks such as language understanding, dialogue interaction, content creation and logical reasoning can facilitate the efficient creation of personalized digital resources, conversational human-computer collaborative learning and literacy orientation It provides strong support in areas

*Corresponding author's email: fdswey1246647351@gmail.com

such as educational evaluation [1]. At present, almost all advanced models (State-of-Art models, SOTA) in The field Of Natural Language Processing (NLP) have evolved based on the large model architecture of Transformer. With Generative pre-training (GPT), Bidirectional Encoder Representation from Transformers (BERT), GPT-3, DALL-E, Switch Large-scale pre-trained models such as Transformer, Huawei Pangu, Unisplendour Wudao, ERINE, and M6 have emerged rapidly. The field of artificial intelligence research is undergoing a paradigm shift from supervised learning to "big data + large models" under unsupervised learning conditions, that is, training based on massive wide-area data and automatically adapting to a wide range of downstream applications through fine-tuning and learning the model of the task [2].

Therefore, a comprehensive and objective assessment of large language models is necessary. This article will combine relevant data to analyze the main language models through some evaluation methods in terms of objective ability testing and subjective user experience.

2 Literature Review

This article will analyze many advanced models (such as GPT-4o, Claude 3.5 and Gemini 2.0) from different scenarios. This article studies the performance differences among various large language models from the perspectives of objective ability testing and subjective user experience.

MMLU was established in 2020 by a group of researchers from four renowned universities. In September 2024, OpenAI's 01 preview model achieved the highest record among all models on the MMLU. In contrast, GPT-4, which was launched in March 2023, scored 86.4%. It is worth noting that as one of the earliest models tested on MMLU, RoBERTa achieved an accuracy rate of only 27.9% in just five years in 2019. The economic growth rate has risen to the current level, reaching as high as 64.4 percentage points [3]. This evaluation method also has many limitations. It still faces significant criticism, with claims that the benchmark tests have errors or are overly simplistic. Developers will use non-standard prompt techniques to report MMLU scores, which can improve performance but may lead to misleading comparisons.

In terms of subjective experience, based on the assessment of Chatbot Arena and according to online introductions, Chatbot Arena is an open-source platform jointly developed by LMSYS and UC Berkeley SkyLab, aiming to promote the development and understanding of large language models through real-time, transparent, and community-driven evaluations. Users can anonymously compare and interact with multiple LLMS, and rank the models through voting using the Elo scoring system. The platform supports multiple publicly released models, including open-source weights and commercial apis, and continuously updates the ranking list based on real user feedback. As of November 2024, the top-scoring models on the Arena Hard-Auto leaderboard were o1-mini (92.0), o1-preview (90.4), and Claude-3.5-Sonnet (85.2) [3]. However, it also has limitations. Since the tests are random, it is impossible to conduct fair and repeatable benchmark tests on a group of models using the same set of prompt words. Moreover, it is impossible to eliminate malicious or repetitive test data, which contaminates the evaluation results.

Based on the assessment of LLM-as-a-Judge, it is known from online information that LLM-as-a-Judge is a method that utilizes the advanced text understanding and generation capabilities of large language models to perform evaluations, judgments, or decisions on specific tasks or outputs. For example, the winning rate of GPT-4 is 10% higher. The winning rate of Claude-V1 is 25% higher. However, GPT-3.5 expressed dissatisfaction with itself. For the existing data and almost non-existent differences, our research does not guarantee that the model can have its own enhanced bias. Wanting to conduct a study and comparison

is quite difficult, because we have to vary the quality of the case to easily fit one response in order to understand the characteristics of another model [4]. Limited mathematical and reasoning abilities regarding Master of Laws may lead them to get incorrect answers and cause problems with their grading. However, it also indicates the disadvantage of grading the basic mathematical problems it can solve. For instance, this article presents an example of a problem where GPT-4 made low-level wrong judgments regarding mathematics. Although GPT - 4 can solve this problem, but it still be misled about the wrong answer, finally make surreal judgment [4].

In these articles, the respective data and limitations are all explained. All of them conduct in-depth research and draw conclusions in one way. These articles provide data and conclusion support for this article, which can be summarized and new conclusions can be drawn.

3 Discovery

3.1 Advantages of closed Source Models

This paper conducts comprehensive experiments on 16 open-source models and 4 closed-source models. For instance, in qwen2.5v1-72b, the best-performing open-source model in pure images has an accuracy rate of only 19.43%. On the contrary, among the closed-source models, Doubao Seed-1.6 has the highest accuracy rate, which is only 41.32%. Compared with advanced closed-source models, the most powerful open-source models have an average accuracy rate reduction of 13.94% in a given scenario. In the Dimg zh with higher visual ability requirements for MLLM, the closed-source model increased by 125.84% gain compared to the large group in the same situation. These results suggest that the open and closed source model has a huge gap in scientific reasoning tasks [5]. Judging from its results, this reflects the leading edge of closed-source models in more subjective and complex tasks such as dialogue interaction and instruction compliance.

3.2 Disadvantages of LLM-AS-A-Judge

Researchers selected four LLM-AS-A-Judge benchmark datasets and expanded them through data synthesis to construct our evaluation dataset. Our evaluation indicators can quantify the impact of scoring bias on the scoring accuracy and scoring tendency of the evaluation model. The experimental results clearly show that scoring bias can affect the robustness and reliability of the existing scoring evaluation models. Evaluation models with different capabilities, scales and architectures have varying sensitivities to various scoring biases. This paper shows that there is a scoring bias. When the scoring cue is disturbed, the scoring given by LLM will be shifted.

3.3 Limitations of systematic bias

In terms of value judgment, compared with continuous scoring, when LLMs are asked to express their opinions in a binary way, they are significantly more inclined to give a "dissenting" judgment. For instance, the average "support" ratio of LLMs for value statements dropped from 74.5% in the continuous format to 60.7% in the binary "support/oppose" format. Text sentiment analysis: Similarly, in sentiment analysis tasks, LLMs is more inclined to judge the text as "negative" in the binary format compared to continuous scoring. For instance, the proportion of LLMs judging as "positive" dropped from 39.9% under the continuous format to 24.6% under the baseline binary condition. "No"

preference: More interestingly, in a controlled experiment of sentiment analysis, when "No" was set to represent "positive", the probability of LLMs choosing "No" significantly increased, suggesting that the model may have a deeper preference for the word "No" itself [6]. This article illustrates the limitations of systematic bias, so the current assessment results must be used carefully and with an understanding of their context.

4 Discussion

This article's review indicates that the evaluation of large language models is developing rapidly, and various evaluation methods can more comprehensively and authoritatively demonstrate their true value.

There are still some challenges that need to be addressed at present. For instance, how to construct a more comprehensive and reasonable evaluation system is one of the relatively urgent challenges that need to be solved. Moreover, after collecting data and analyzing it online, it was found that the LLM-as-a-Judge method could be extended but also had deviations. Chatbot Arena can understand people's feedback and also has uncontrollable variables. Therefore, in order to make the evaluation of large language models more comprehensive and authoritative, it is possible to attempt to construct a hybrid framework. It needs to avoid the problems existing in the above assessment methods and integrate their advantages.

As shown in table 1. Through the articles written by Bai Tai Lao Ren, large-scale multi-task language understanding methods were used to test large language models. It can be seen that the accuracy rate of Claude-3.5 has reached 88.3%, which is currently the strongest comprehensive knowledge ability. The performance of GPT-4o and DeepSeek V3 is closely following the first place [7].

Table 1. Data on accuracy rates for each model

Ranking	Model name	Accuracy rate	Release time
1	Claude-3.5	88.30%	In June 2024
2	GPT-4o	87.20%	In May 2024
3	DeepSeek-V3	87.10%	In 2024
4	Qwen2.5-72B	85.00%	In 2024

As shown in table 2. Through the article written by Anru Shao Nian Chu Ru Meng 662, it can be known that GPT-4o performs best in terms of user experience and dialogue quality by testing large language models using the Chatbot Arena method. Claude-3.5 is only 3 points behind the first place, and the strength is very close. DeepSeek V3 has a certain gap compared with the previous two models [8].

Table 2. Data on the scores of each model through the Chatbot Aren method

Ranking	Model name	Score	Test time
1	GPT-4o	1411	On May 13, 2024
2	Claude-3.5	1408	On May 13, 2024
3	DeepSeek-V3	1367	On May 13, 2024
4	Qwen2.5-72B	1324	On May 13, 2024
5	Llama-3.1-405B	1318	On May 13, 2024

In terms of evaluating the performance of the model, from AI Index Report, it can be found that each model has different levels of data contamination when using MMLU evaluation methods [3]. Therefore, new problems can be added to the reasoning ability to reduce data pollution. Of course, the advantages of the MMLU method should also be retained, such as an extremely wide test range, which can more comprehensively evaluate

the performance of the model.

In terms of bias to different models, from CodeShare points out that experiments show that it is feasible to predict human preferences using a small amount of data when using the Chatbot Arena evaluation method. Therefore, a relatively accurate assessment can be made using a small amount of data. It is also necessary to retain its advantages, such as the genuine user experience, which can reduce users' prejudice against different models.

In terms of self-bias in the assessment, from discovery of Zhang et al, self-bias may occur when using the LLM-as-a-Judge evaluation method [4]. So multiple referees need to be added to improve accuracy. However, it is also necessary to retain its advantages. It has a very high assessment efficiency and can reduce assessment time, etc.

5 Conclusion

The performance evaluation of large language models is a complex study. This paper systematically reviews the main viewpoints and methodological analyses of this study. Through in-depth analysis of mainstream evaluation methods such as MMLU, Chatbot Arena, and LLM-as-a-Judge, it concludes that the performance differences among large language models are significant. Research has found that different models have significant differences in objective ability and subjective experience, and the choice of evaluation methods itself will profoundly affect the evaluation results. Each method has its own advantages and limitations. However, this paper also has some limitations. It only studies objective ability tests and subjective user experience, and does not conduct a comprehensive evaluation of large language models. This article only evaluates the currently mainstream large language models and does not assess all of them. In the future, more comprehensive, authoritative and all-round evaluation methods should be studied. This can provide a framework reference for future research.

References

1. M. Liu, Z. M. Wu, J. Liao et al. Educational Applications of Large Language Models: Principles, Current Situations and Challenges - From Lightweight BERT to Conversational ChatGPT. Modern Educational Technology, 33(08):19-28. (2023)
2. Z. Y. Zhao, G. B. Zhu, J. Q. Wang. The Implications of ChatGPT for Language Large Models and New Development Ideas for Multimodal Large Models. Data Analysis and Knowledge Discovery, 7(03):26-35 (2023)
3. N. Maslej, L. Fattorini, R. Perrault, Artificial Intelligence Index Report 2025. arXiv:2504.07139. (2025)
4. L. Zheng, W. L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, ... & I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36, 46595-46623. (2023)
5. J. C. Ruan, J. Dan, X. Gao, T. Liu, Y. Z. Fu, Y. Y. Kang. Mme-SCI: A comprehensive and challenging science... Available at: https://www.arxiv.org/pdf/2508.13938 (2025)
6. Y. L. Lu, C. H. Zhang, W. Wang. Systematic Bias in Large Language Models: Discrepant Response Patterns in Binary vs. Continuous Judgment Taskss. Available at: https://www.arxiv.org/pdf/2508.13938 (2025)
7. Y. B. Wang, X. G. Ma, G. Zhang, et al., MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. arXiv:2406.01574. (2024)
8. A. X. Liu, B. Fang, B. Xue, et al., DeepSeek-V3 Technical Report. arXiv:2412.19437v2. (2024)