Towards Personalized Evaluation of Large Language Models with An Anonymous Crowd-Sourcing Platform

Mingyue Cheng¹, Hao Zhang¹, Jiqian Yang¹, Qi Liu^{1*}, Li Li¹, Xin Huang¹, Liwei Song¹, Zhi Li², Zhenya Huang¹, Enhong Chen¹

¹Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence, Hefei, China,

² Shenzhen International Graduate School, Tsinghua University, Shenzhen, China {mycheng,qiliuql,huangzhy,cheneh}@ustc.edu.cn,{zh2001,yangjq,lili0516,wuli_error,songliv}@mail.ustc.edu.cn, zhilizl@sz.tsinghua.edu.cn

ABSTRACT

Large language model evaluation plays a pivotal role in the enhancement of its capacity. Previously, numerous methods for evaluating large language models have been proposed in this area. Despite their effectiveness, these existing works mainly focus on assessing objective questions, overlooking the capability to evaluate subjective questions which is extremely common for large language models. Additionally, these methods predominantly utilize centralized datasets for evaluation, with question banks concentrated within the evaluation platforms themselves. Moreover, the evaluation processes employed by these platforms often overlook personalized factors, neglecting to consider the individual characteristics of both the evaluators and the models being evaluated. To address these limitations, we propose a novel anonymous crowd-sourcing evaluation platform, BingJian, for large language models that employs a competitive scoring mechanism where users participate in ranking models based on their performance. This platform stands out not only for its support of centralized evaluations to assess the general capabilities of models but also for offering an open evaluation gateway. Through this gateway, users have the opportunity to submit their questions, testing the models on a personalized and potentially broader range of capabilities. Furthermore, our platform introduces personalized evaluation scenarios, leveraging various forms of human-computer interaction to assess large language models in a manner that accounts for individual user preferences and contexts. The demonstration of BingJian can be accessed at https://github.com/Mingyue-Cheng/Bingjian.

CCS CONCEPTS

• Computing methodologies → Natural language processing.

Qi Liu is corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW '24 Companion, May 13-17, 2024, Singapore, Singapore

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0172-6/24/05

https://doi.org/10.1145/3589335.3651243

KEYWORDS

Large Language Model, Personalized Evaluation, Crowdsourcing Platform

ACM Reference Format:

Mingyue Cheng¹, Hao Zhang¹, Jiqian Yang¹, Qi Liu^{1*}, Li Li¹, Xin Huang¹, Liwei Song¹, Zhi Li², Zhenya Huang¹, Enhong Chen¹. 2024. Towards Personalized Evaluation of Large Language Models with An Anonymous Crowd-Sourcing Platform. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651243

1 INTRODUCTION

The advent of Large Language Models (LLMs) has marked a significant milestone in the journey toward Artificial General Intelligence (AGI) [5, 7], opening new frontiers in our ability to process and understand complex human languages at an unprecedented scale. As these models become increasingly sophisticated, their evaluation transcends traditional paradigms [9], challenging the very notion of a singular, definitive ground truth. In the realm of large language model development, the absence of a clear-cut benchmark necessitates a reimagined approach to evaluation—one that accommodates the nuanced and multifaceted nature of tasks these models are designed to tackle. This shift underscores the critical role of evaluation methodologies in not only benchmarking current capabilities but also in driving the evolution of model sophistication. The quality and depth of these evaluation mechanisms, therefore, directly influence the trajectory of large language model advancements, making it imperative to explore and refine our evaluative frameworks to keep pace with the rapid advancements in this domain.

In the quest to evaluate LLMs, researchers are diligently working to gauge an expansive range of model capabilities, from coding proficiency to domain-specific expertise. These efforts [1, 6, 11] play a crucial role in refining evaluation methodologies and shedding light on the complex competencies of LLMs. Yet, in spite of these advances, current methods face significant shortcomings that demand attention. One major issue is the dependence on centralized datasets, which narrows the evaluation to a set of predetermined challenges and fails to encompass decentralized, real-world problems. Additionally, most evaluation frameworks do not adequately consider the integration of personalized user data [3, 10], an essential factor that could provide deeper insights into how models perform across varied user interactions. These challenges highlight the necessity for inventive evaluation approaches that not only broaden the scope of problem collection to include decentralized

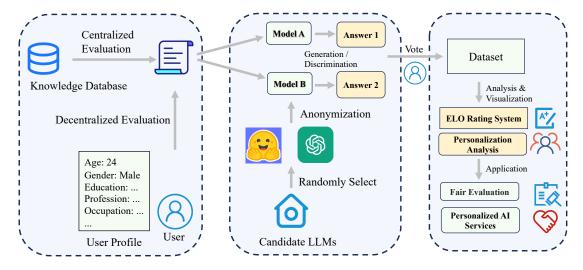


Figure 1: The illustration of evaluation pipeline of BingJian platform.

issues but also factor in individual user contexts, thereby making the evaluation results more relevant and applicable.

To address the prevailing challenges in the evaluation of LLMs, we design a platform named BingJian aimed at facilitating comprehensive model assessment. On this platform, responses from different models are presented to users, who then are encouraged to select the most appropriate answer. To ensure a fair evaluation of model capabilities, BingJian employs an ELO rating system [8] that adjusts model scores based on user selections. Furthermore, BingJian is designed as an open, crowdsourced platform. Just as ImageNet [2] significantly advanced the field of computer vision by constructing a high-quality image dataset through crowdsourced annotations, we aim to revolutionize the evaluation of large language models. Traditional objective assessments [4] fall short of capturing the full capabilities of LLMs. By incorporating human crowdsourcing evaluations, we introduce the most authentic form of human feedback. Humans assess models based on various intangible aspects, such as the quality of generated text, knowledge conveyed, and presentation style, offering a comprehensive evaluation beyond quantifiable metrics. Naturally, a crowdsourced platform can gather a wide variety of evaluation results. These outcomes are closely related not only to objective facts but also to the personalized information of the evaluators. To this end, our platform compiles a comprehensive dataset of evaluator personalization information. We aim to delve into this personal data to uncover the cognitive relationships between humans and LLMs. This endeavor provides a more holistic depiction of the evaluation results for LLMs, revealing how the individual characteristics of evaluators influence their interactions with and assessments of these models. By exploring these nuanced relationships, we can enhance our understanding of model performance from a perspective that integrates human subjectivity, thereby enriching the evaluation process.

2 THE PROPOSED BINGJIAN

In this section, we introduce our BingJian platform, as depicted in Figure 1. Initially, we establish a comprehensive question database



Figure 2: Centralized and decentralized evaluation interface.

encompassing a wide array of domains, along with gathering a suite of large language models for assessment. Following this, we develop interfaces incorporating both centralized and decentralized approaches to evaluation, enabling the collection of varied data sets pivotal for evaluating the multifaceted capabilities of LLMs. Next, we provide a detailed introduction to the BingJian platform, focusing on the interface design, the evaluation process, and the data analysis and visualization.

2.1 Interface Overview

On the login page, we first encourage users to fill out their profile information, including age, gender, profession, and educational background, to facilitate subsequent analysis of personalized large language model evaluation results. Then, as shown in Figure 2, the BingJian evaluation interface primarily consists of two parts: centralized evaluation and decentralized evaluation.

2.1.1 Centralized Evaluation. The central evaluation is to evaluate the performance of LLM on a constructed question set, which covers general knowledge in various domains such as natural sciences, humanities, economics, etc. In each interaction, the users may choose a question displayed on the page. Then, the system will randomly invoke two different models to answer the question

and anonymously display their answers along with the explanations and analyses, on the evaluation interface. Such an anonymous mechanism could help eliminate bias against different LLMs to some extent, effectively ensuring the fairness of the evaluation.

Furthermore, we also integrate a question-recommendation system. Based on users' past question browsing history, we select questions that align with their interests to push to them while ensuring, as much as possible, that users do not encounter the same question twice. From the perspective of user experience, this approach significantly enhances the engaging nature of the crowd-sourced evaluation process. When considered from an evaluation standpoint, this strategy aims to broaden users' evaluation activities across various related disciplines, thereby minimizing the risk of inaccurate assessments. This personalized and dynamic question recommendation not only caters to the users' preferences but also enriches the evaluation dataset by capturing a wider spectrum of user interactions and responses, leading to more robust and comprehensive insights into model performance.

2.1.2 Decentralized Evaluation. Due to the continuous iteration and updating of LLMs, the dataset in the constructed question bank may be incorporated into the model's training corpus in the present or future, leading to biases in the evaluation results. To address this challenge, we further design the decentralized evaluation module shown in the right column of Figure 2. Users can input custom questions into the dialogue box and then click the button below to evaluate the generated answers. This feature greatly alleviates the potential issue of question leakage in the evaluation nowadays and further improves the fairness of the leader board while supporting the open-domain question.

2.2 Crowdsource Evaluation

Just as ImageNet revolutionized computer vision research by involving a large number of contributors in labeling images, our platform aims to harness the power of crowdsourcing to establish a comprehensive evaluation process and database for large language models. Next, we outline the specific evaluation process following the three primary data collection goals, respectively. Through the implementation of crowdsourced evaluation, our objective is twofold: firstly, to objectively delineate the capabilities of numerous LLMs, and secondly, to foster the development of a benchmarking system through human-computer interaction assessments that will propel further advancements in LLM technologies. This benchmark will serve as a valuable resource for researchers and developers alike, offering a standardized framework against which the progression of model capabilities can be rigorously tested and compared.

2.2.1 General Knowledge Mastery. To evaluate the general knowledge mastery of the models, we have created question banks in various domains such as nature, science, the humanities, and economics. The multiple-choice questions will be presented to LLM with the following prompt: For the following questions, please give the correct option and explanation. <Question>, (A) <Answer1>, (B) <Answer2>, (C) <Answer3>, (D) <Answer4>. Then, the models are required to provide the correct options for the given questions. By matching their responses with the correct answers, we can calculate the accuracy of the model's answers and also preliminary conclude its capabilities over various domains.



Figure 3: Evaluation process of generative ability.

2.2.2 Generative Ability. In addition to the problem answer, explainability is also an important evaluation indicator. As shown in Figure 3, users are encouraged to score the model's generative ability based on the answers and their analysis. The evaluation mainly involves a comparison (i.e., "JUST AS GOOD", "A IS BETTER", "B IS BETTER", "JUST AS BAD") of the outputs generated by different models. Besides, it also includes a quantitative evaluation of the model's generative capacity, as assessed using an exscoring system ranging from 1 to 5. Furthermore, given the substantial amount of user profile information collected, we highlight that this data can be utilized to analyze the correlation between the model's responses and personalized user profiles, providing potential research opportunities in the realm of personalized AI services. For instance, certain user groups may be more inclined to prefer professional explanations, while others may favor imaginative responses.

2.2.3 Discriminate Ability. Evaluating the discriminate ability of large language models is essential to ensuring their reliability, addressing biases, and benchmark performance, providing valuable insights into their generalization capabilities in real-world applications. In this vein, our evaluation framework goes beyond simply assessing the generative prowess of large language models. We also scrutinize their ability to evaluate and judge different answers by engaging them in a comparative analysis. Specifically, for a given question, our system employs a double-blind method where it randomly selects two models, referred to as A and B, to provide answers without revealing their identities. These responses are then displayed on the user interface. Finally, users are invited to participate by scoring the responses given by models C and D to evaluate the discriminating ability of the large language models. Through this multi-tiered evaluation process, we can identify strengths and weaknesses in the models' abilities to discriminate between highand low-quality responses. Such insights are instrumental for iterative improvements, leading to more sophisticated and reliable AI systems. Ultimately, by enhancing the discriminative capabilities of LLMs, we can better tailor them to a variety of applications, ensuring that they not only produce content that is engaging and informative but also critically sound and contextually appropriate.

2.3 Analysis & Visualization

2.3.1 ELO Rating Mechaminsm. Inspired by the renowned chess ranking system, the Elo Rating System (ELO) provides a dynamic and intuitive framework for gauging relative strengths. We initialize models with ELO ratings, in which winners gain ELO points

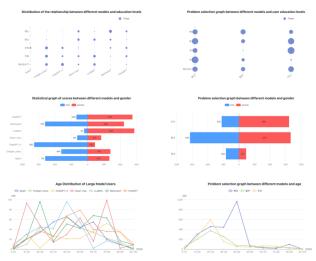


Figure 4: Visualization of correlation between the model's responses and personalized user group profiles.

while losers lose points, ensuring a continuous adjustment of model rankings based on relative model performance. To be specific, it updates a participant's rating (R') based on the outcome (S) of each evaluation record, which can be expressed as:

$$R' = R + K \cdot (S - E),$$

where R' is the updated ELO rating, R is the pre-match ELO rating, K is a constant determining the rating change magnitude, S is the match outcome (1 for a win, 0 for a loss, 0.5 for a draw) and E is the expected outcome calculated using a logistic function, i.e. $E = \frac{1}{1+10^{((R_B-R_A)/400)}}$, in which R_A and R_B are the initial ELO ratings of the two participants. The logistic function ensures that the expected outcome aligns with the participants' relative ratings, making ELO a dynamic and reliable measure of skill in various competitive scenarios. This innovative methodology allows for a fair and nuanced assessment of the model's capability, transcending traditional evaluation metrics.

2.3.2 Crowd Analysis. To delve into the relationship between the responses generated by the model and the diverse backgrounds of the user groups, we embarked on a comprehensive visual analysis, as delineated in Figure 4. This analysis scrutinized the evaluation data through the lens of various demographic dimensions, such as age, gender, profession, occupation, and educational attainment. Our information collection requires user approval. Our objective in this exploratory endeavor is to detect and understand patterns that could inform and enhance the customization of services provided by large language models. For example, we might discover that specific demographic segments have a predilection for responses that are steeped in professional jargon or technical detail, while others might demonstrate a preference for responses that are more creative or narrative in nature. Moreover, by examining the assessment results from these diverse demographic vantage points, we can identify unique opportunities for research into tailored AI services. This granular analysis not only aids in refining the user experience but also serves as a foundational step towards the development of AI systems that are sensitive to the nuanced needs and preferences of different user groups. By integrating these insights into the iterative

design of large language models, we can move closer to achieving a level of personalized interaction that mirrors the adaptive and discerning nature of human communication.

3 CONCLUSION

This paper introduced a personalized, anonymized crowd-sourcing platform for evaluating the capacity of large language models, providing users with both centralized and decentralized evaluation entry points. Users are enabled to assess models' generative and discriminative capabilities within this framework. Moreover, the platform conducts a comprehensive analysis of users' personalized information in conjunction with model evaluation results, utilizing visual statistical charts to display relevant profile information. This innovative approach not only enriches the evaluation landscape by incorporating a human-centric perspective but also paves the way for a more nuanced understanding of model capabilities across diverse user backgrounds and preferences. The ongoing expansion of model integrations underscores our commitment to offer a robust and dynamic evaluation environment, poised to adapt and evolve with the advancing frontiers of large language model technologies.

Acknowledgements. This research was supported by grants from the National Key Research and Development Program of China (Grant No. 2021YFF0901003), the National Natural Science Foundation of China (Grants No. 62337001, U20A20229), and the Fundamental Research Funds for the Central Universities. This work also thanks the support of funding of SC5290005194. We thank the Hefei Artificial Intelligence Computing Center of Hefei Big Data Asset Operation Co., Ltd. for providing computational resources for this project.

REFERENCES

- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology (2023).
- [2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. Ieee, 248–255.
- [3] Jingtao Ding, Fuli Feng, Xiangnan He, Guanghui Yu, Yong Li, and Depeng Jin. 2018. An improved sampler for bayesian personalized ranking by leveraging view data. In Companion Proceedings of the The Web Conference 2018. 13–14.
- [4] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. arXiv preprint arXiv:2305.08322 (2023).
- [5] Junzhe Jiang, Shang Qu, Mingyue Cheng, and Qi Liu. 2023. Reformulating Sequential Recommendation: Learning Dynamic User Interest with Contentenriched Language Modeling. arXiv preprint arXiv:2309.10435 (2023).
- [6] Jiatong Li, Rui Li, and Qi Liu. 2023. Beyond Static Datasets: A Deep Interaction Approach to LLM Evaluation. arXiv preprint arXiv:2309.04369 (2023).
- [7] Yucong Luo, Mingyue Cheng, Hao Zhang, Junyu Lu, and Enhong Chen. 2023. Unlocking the potential of large language models for explainable recommendations. arXiv preprint arXiv:2312.15661 (2023).
- [8] Radek Pelánek. 2016. Applications of the Elo rating system in adaptive educational systems. Computers & Education 98 (2016), 169–179.
- [9] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018).
- [10] Tong Zhao, Julian McAuley, and Irwin King. 2014. Leveraging social connections to improve personalized ranking for collaborative filtering. In Proceedings of the 23rd ACM international conference on conference on information and knowledge management. 261–270.
- [11] Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, et al. 2023. Efficiently Measuring the Cognitive Ability of LLMs: An Adaptive Testing Perspective. arXiv preprint arXiv:2306.10512 (2023).