Towards Comprehensive Evaluation of Open-Source Language Models: A Multi-Dimensional, User-Driven Approach

Qingchen Yu[1]
[1] School of Management, Shanghai University, Shanghai, China
zhgqcyu@outlook.com

Abstract

With rapid advancements in large language models (LLMs) across artificial intelligence, machine learning, and data sci-ence, there is a growing need for evaluation frameworks that go beyond traditional performance metrics. Conventional methods focus mainly on accuracy and computational metrics, often neglecting user experience and community interaction—key elements in open-source environments. This paper intro-duces a multi-dimensional, user-centered evaluation framework, integrating metrics like User Engagement Index (UEI), Community Response Rate (CRR), and a Time Weight Factor (TWF) to assess LLMs' real-world impact. Additionally, we propose an adaptive weighting mechanism using Bayesian op-timization to dynamically adjust metric weights for more accurate model evaluation. Experimental results confirm that our framework effectively identifies models with strong user engagement and community support, offering a balanced, data-driven approach to open-source LLM evaluation. This frame-work serves as a valuable tool for developers and researchers in selecting and improving open-source models. All resources are available at https://github.com/Duguce/UserDriven-LLMEval.

1 Introduction

In recent years, large language models (LLMs) in the field of natural language processing (NLP) have achieved remarkable advancements, driving performance improvements across various applications such as machine translation, text generation, and automated question answering (Brown et al., 2020; Yang et al., 2024). Since the introduction of GPT-3, open-source LLMs have continued to expand in scale and performance, drawing substantial interest from developers and researchers alike (Zheng et al., 2025; Liang et al., 2024; Chen et al., 2024). As the number of models increases rapidly, selecting the most suitable LLM among numerous options has become a critical challenge in practical applications. Existing methods for evaluating LLMs primarily focus on performance testing, usually measuring accuracy or other technical metrics on standardized datasets (Devlin et al., 2019; Raffel et al., 2020). However, performance-based evaluations alone often fall short of comprehensively capturing a model's real-world application value. This is particularly true in open-source environments, where user experience and community engagement are increasingly recognized as key factors in evaluating a model's actual impact.

In open-source communities, the practical value of LLMs depends not only on their technical performance but also on user feedback and community support and interaction. For example, user interaction data on platforms like Hugging Face [1] and GitHub [2]—such as download counts, likes, issue reports, and pull requests—provides essential insights for evaluating models, reflecting the real-world demand for and user experience with these models. Therefore, traditional evaluation methods that focus solely on performance metrics have significant limitations, as they fail to capture the full impact of open-source LLMs. Based on this observation, this paper proposes a multi-dimensional, user-driven evaluation framework. By integrating metrics such as User Engagement Index (UEI), Community Response Rate (CRR), and a Time Weight Factor (TWF), we aim to establish a more practically valuable framework for comprehensive LLM evaluation.

To enhance the flexibility and adaptability of the evaluation framework, this paper further introduces an adaptive weight optimization mechanism. Since the impact of user interaction and community response may vary across different models, a fixed

[1] https://www.huggingface.co
[2] https://www.github.com

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 1–7
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

weight allocation is often inadequate for all models. Therefore, we employ a Bayesian optimization approach to automatically adjust the weights of each metric, ensuring that different models receive a fair and accurate evaluation across all evaluation dimensions. This adaptive weight optimization mechanism effectively improves the scientific rigor and representativeness of evaluation results, providing a more objective reference for model selection.

Additionally, this paper introduces a TWF to address the balance in scoring between newer and older models. Models released more recently may have limited accumulated user and community data, and traditional scoring methods often treat these models unfairly. The introduction of the TWF reduces time-related bias in scoring to a certain extent, ensuring that evaluation results maintain a high level of fairness across models with different release dates.

The main contributions of this paper include the following:

- We propose a multi-dimensional evaluation framework based on user engagement and community response rate, integrating real user and community feedback data to provide a panoramic perspective for evaluating models in open-source settings.

- We introduce a time weight factor to address fairness issues in scoring between newer and older models, enhancing temporal consistency in evaluations.

- We design an adaptive weight mechanism based on Bayesian optimization, allowing the weights of each metric to adjust automatically according to a model's specific performance, thereby enhancing the flexibility and scientific rigor of the evaluation framework.

The proposed evaluation framework not only offers a new perspective for evaluating open-source LLMs but also provides developers and researchers with a scientific reference for optimizing model design and enhancing user experience. We hope this study will offer valuable support for selecting, improving, and advancing open-source LLMs in the future.

2 Related Works

Existing methods for evaluating LLMs primarily focus on standardized datasets, using metrics such as accuracy and F1 scores to gauge model performance on specific tasks (Liang et al., 2023; Yu et al., 2024, 2025a). While these methods provide a direct reference for evaluating a model's technical performance, in real-world applications, user feedback and community interaction are equally important components of a model's overall impact. Moreover, many models may be fine-tuned on particular datasets, potentially resulting in overfitting, which limits their ability to accurately reflect performance across diverse scenarios (Elangovan et al., 2024; Yu et al., 2025b).

In recent years, increasing research attention has been directed toward user experience and community support for models (Chang et al., 2024). In open-source projects, user interaction and community engagement are regarded as critical factors in measuring a project's value. Metrics such as download counts and likes on the Hugging Face platform, as well as stars and issue reports on GitHub, are increasingly used as indicators of a model's popularity and community activity level. However, most current evaluation frameworks are limited to single-dimensional metrics of user or community engagement, lacking a comprehensive, multi-dimensional analysis. This paper constructs a multi-dimensional evaluation system based on user engagement, community response rate, and a time-weighting factor, complemented by an adaptive weight optimization method, to provide a more holistic, user-centered perspective for evaluating LLMs.

3 Methodology

3.1 Data Collection and Preprocessing

Our evaluation framework is based on multi-dimensional open-source data collected from the Hugging Face and GitHub platforms, which authentically reflect the popularity and user engagement of open-source LLMs. By systematically collecting this data, we aim to establish a user experience-centered, comprehensive evaluation framework for LLMs.

Specifically, the Hugging Face platform is currently the leading open-source platform for LLMs and serves as the primary channel for users to download these models, while GitHub is the main hosting platform for open-source projects, gathering attention and feedback from developers worldwide. The integration of data from both platforms provides comprehensive insights into model usage and developer community engagement. Therefore, we selected the following data metrics:

- **Monthly Downloads:** This metric indicates the number of times the model was downloaded by users in the past month, directly reflecting the model's actual usage by users.
- **Total Likes:** This metric represents overall user satisfaction with the model. A higher number of Likes suggests greater user approval.
- **Total Stars:** This metric reflects the model's popularity; a higher number of Stars indicates a higher level of attention within the open-source community.
- **Open Issues and Closed Issues:** These represent unresolved and resolved user feedback, respectively. Open Issues indicate current pending user feedback, while Closed Issues reflect the responsiveness of the development team to user feedback.
- **Open PRs and Closed PRs:** These represent the number of unmerged and merged pull requests, respectively. PR data is used to assess community contributions and improvements to the model, with Closed PRs particularly reflecting the development team's receptivity to community suggestions.

The data for Monthly Downloads and Total Likes is sourced from the Hugging Face platform, while the other metrics are obtained from GitHub.

To ensure data consistency, the raw data collected was standardized through the following processes.

Outlier Treatment. Extreme values were handled using a truncation method to reasonably limit their influence on the scoring.

Normalization. Since the scales of different metrics vary, Min-Max normalization was applied to scale each metric to the [0,1] range, ensuring consistency in scoring dimensions:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

3.2 Evaluation Framework Design

The user feedback-based comprehensive evaluation framework for LLMs proposed in this paper conducts a holistic evaluation by utilizing multi-dimensional metrics, including user engagement, community participation, and response efficiency. This framework combines metric selection, adaptive weight optimization, and time-weighted processing to ensure the scientific rigor and objectivity of the scoring system.

Specifically, we constructed the following key metrics based on the collected raw data to reflect the model's performance across different dimensions:

UEI. This metric combines user download counts and cumulative feedback, incorporating time normalization to mitigate the impact of model release duration. It is defined as follows:

$$\begin{aligned} \text{UEI}_i = {} & \frac{\text{Total Likes}_i}{T_{\text{model},i}} \\ & + \frac{\text{Total Stars}_i}{T_{\text{model},i}} \\ & + \text{Monthly Downloads}_i \end{aligned} \tag{2}$$

CRR. The Community Response Rate measures the efficiency of the model team in responding to user feedback and is defined as follows:

$$\text{CRR}_i = \frac{\text{Closed Issues}_i}{\text{Open Issues}_i + \text{Closed Issues}_i} \tag{3}$$

Here, Closed Issues_i and Open Issues_i represent the numbers of resolved and unresolved user feedback for model i, respectively.

TWF. To account for the impact of release time on cumulative metrics (such as Total Likes and Total Stars), a Time Weight Factor W_time is introduced, defined as follows:

$$W_{\text{time},i} = \frac{T_{\text{ref}}}{T_{\text{model},i} + \epsilon} \tag{4}$$

Here, T_{ref} represents the reference time window, $T_{\text{model},i}$ denotes the number of months since model i was released, and ϵ is a bias term.

To achieve a comprehensive score across multiple metrics, this paper employs an adaptive weight optimization mechanism based on Bayesian optimization, allowing for automatic adjustment of each metric's weight and enhancing the flexibility of the scoring system. The scoring formulas for each metric are defined as follows:

$$\text{FinalScore}_i = w_1 \cdot \text{UEI}_i \cdot W_{\text{time},i} + w_2 \cdot \text{CRR}_i \tag{5}$$

Here, UEI_i represents the User Engagement Index, CRR_i represents the Community Response

Rate, and w_1 and w_2 are weight parameters that satisfy $w_1 + w_2 = 1$.

The optimization objective is to maximize the average variance in model scores, with the calculation formula defined as follows:

$$\max_{w_1,w_2} \frac{1}{N(N-1)} \sum_{i \neq j} |\text{FinalScore}_i - \text{FinalScore}_j| \quad (6)$$

Bayesian optimization automatically searches for weight combinations (w_1, w_2) to maximize the average distance between model scores, thereby enhancing the effectiveness of the evaluation framework.

4 Experiments

4.1 Experimental Setup

Datasets This study collected multi-dimensional data on 24 well-known open-source LLMs from the Hugging Face and GitHub platforms. These models were released by notable institutions such as Meta, Google, and Alibaba. The dataset includes information on user engagement and community feedback, providing a rich foundation for comprehensive model evaluation. Data collection was primarily conducted through each platform's API to ensure data timeliness and accuracy. To maintain consistency and comparability, all data used in this experiment was collected up to November 9, 2024. During data preprocessing, we performed outlier treatment and normalization to enhance data reliability and the robustness of the analysis.

Metrics Based on the constructed comprehensive evaluation framework, this study designed three core metrics: UEI, CRR, and TWF to thoroughly evaluate the performance of open-source models in real-world applications. These metrics, formally defined in Section 3, encompass dimensions such as user interaction, community support, and temporal adaptability of the models. In the experiments, we determined the optimal weight combination for each metric through Bayesian optimization to generate the final comprehensive score.

4.2 Main Results

We first used Bayesian optimization to determine the optimal weight combination for the metrics, resulting in final optimal weights of w_1=3.0 and w_2=1.0. This outcome indicates that UEI holds a higher weight in the comprehensive evaluation of the models, while the influence of CRR is relatively smaller.

This weight allocation aligns with real-world conditions, as information such as user download counts and likes more directly reflects a model's use in actual scenarios. Thus, these factors hold a higher weight in our scoring system, making the evaluation results more closely aligned with actual user experience. In comparison, although community response rate is also significant for the model's sustainable development and iterative improvement, its lower weight emphasizes the priority of widespread user adoption in model evaluation. Through this weight distribution, our evaluation framework achieves a reasonable balance between user experience and community feedback, ensuring the scientific rigor and representativeness of the scoring system.

Figure 1 presents the scores of various models and the contribution of each metric to those scores. In the figure, different colored blocks represent the weighted contributions of UEI * TWF and CRR to each model's score, while the green line indicates the final score of each model.

Table 1 provides a more detailed breakdown of the evaluation results, listing key metrics for each model, including the UEI, CRR, TWF, and the final computed score. These results offer a more granular view of how user interaction and community support influence model rankings.

Case Study From the results, we observe that models with high user engagement metrics and developed by organizations with active community support tend to achieve higher final scores. For example, Qwen2.5-72B-Instruct and Llama3.2-3B-Instruct demonstrate outstanding performance in both user downloads and community response. These models have gained substantial user approval, and the development teams actively address feedback and update the codebase, fostering a positive interaction between users and developers. This finding highlights the critical role of user-oriented engagement and prompt community response in promoting widespread model adoption in practical applications.

Conversely, models such as ChatGLM-3-6B and Yi-34B-Chat rank relatively lower in the final evaluation. As seen in Table 1, these models exhibit lower UEI and CRR scores, indicating lower levels of user adoption and community responsiveness. While technical performance remains a key fac-

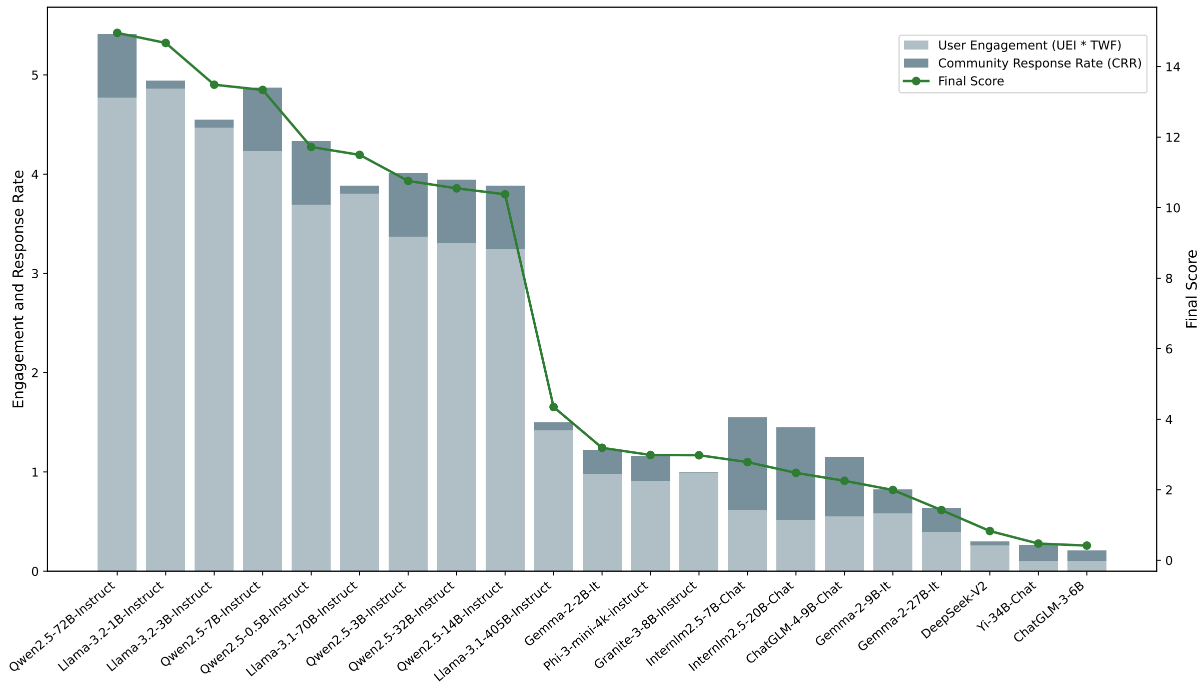

Figure 1: Breakdown of Final Scores with User Engagement and Community Response Contributions Across Open-source LLMs.

Model Name	#Params	Publisher	Release	UEI	CRR	TWF	Score
ChatGLM-3-6B	6B	Tsinghua	2023/10/25	0.11	0.10	0.92	0.42
ChatGLM-4-9B-Chat	9B	Tsinghua	2024/6/4	0.23	0.60	2.40	2.25
Llama-3.2-3B-Instruct	3B	Meta	2024/9/25	0.75	0.08	6.00	13.49
Llama-3.2-1B-Instruct	1B	Meta	2024/9/25	0.81	0.08	6.00	<u>14.67</u>
Llama-3.1-70B-Instruct	70B	Meta	2024/7/23	1.27	0.08	3.00	11.50
Llama-3.1-405B-Instruct	405B	Meta	2024/7/23	0.47	0.08	3.00	4.35
Qwen2.5-72B-Instruct	72B	Alibaba	2024/9/19	0.80	0.64	6.00	**14.96**
Qwen2.5-32B-Instruct	32B	Alibaba	2024/9/19	0.55	0.64	6.00	10.55
Qwen2.5-14B-Instruct	14B	Alibaba	2024/9/19	0.54	0.64	6.00	10.38
Qwen2.5-7B-Instruct	7B	Alibaba	2024/9/19	0.71	0.64	6.00	13.34
Qwen2.5-3B-Instruct	3B	Alibaba	2024/9/19	0.56	0.64	6.00	10.76
Qwen2.5-0.5B-Instruct	0.5B	Alibaba	2024/9/19	0.62	0.64	6.00	11.72
Granite-3-8B-Instruct	8B	IBM	2024/10/3	0.08	0.00	12.00	2.98
DeepSeek-V2	236B	DeepSeek	2024/4/22	0.15	0.04	1.71	0.83
Gemma-2-27B-It	27B	Google	2024/6/24	0.16	0.24	2.40	1.42
Gemma-2-9B-It	9B	Google	2024/6/24	0.24	0.24	2.40	1.99
Gemma-2-2B-It	2B	Google	2024/6/24	0.41	0.24	2.40	3.19
Phi-3-mini-4k-instruct	3B	Microsoft	2024/4/23	0.53	0.25	1.71	2.99
Yi-34B-Chat	34B	01 AI	2024/5/13	0.05	0.16	2.00	0.47
Internlm2.5-20B-Chat	20B	Shanghai AI Lab	2024/7/3	0.17	0.93	3.00	2.48
Internlm2.5-7B-Chat	7B	Shanghai AI Lab	2024/7/3	0.21	0.93	3.00	2.78

Table 1: Comparative Evaluation of Open-Source LLMs Based on User Engagement and Community Response. The table presents the evaluation scores of various open-source large language models (LLMs) across multiple dimensions, including User Engagement Index (UEI), Community Response Rate (CRR), and Time-Weighted Factor (TWF). The highest Final Score is **boldfaced**, and the second-highest is <u>underlined</u>.

tor in LLM development, our findings suggest that user engagement and developer interaction play an equally crucial role in determining a model's long-term impact and usability.

Additionally, we observe that some models, such as Granite-3-8B-Instruct and DeepSeek-V2, receive relatively low scores despite their large parameter sizes. This result implies that model size

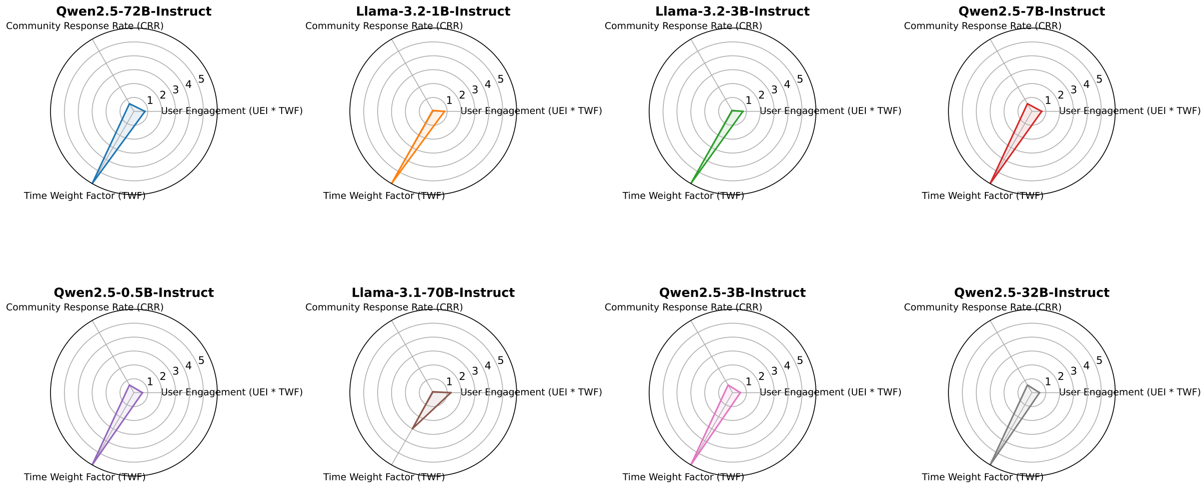

Figure 2: Comparative Analysis of Key Interaction Metrics Across Top 8 Open-source LLMs.

alone does not necessarily translate to higher user engagement or stronger community feedback. Instead, factors such as accessibility, documentation quality, and active issue resolution may significantly impact a model's real-world adoption.

These insights reinforce the necessity of multi-dimensional evaluation metrics when assessing open-source LLMs, as traditional accuracy-based benchmarks alone may not fully capture a model's practical influence. By incorporating user-driven engagement factors into LLM evaluation, our framework provides a more holistic perspective that can better guide model selection and improvement efforts.

We analyzed the metrics of the top 8 LLMs in the overall score rankings—UEI, CRR, and TWF—as shown in Figure 2. The radar chart clearly illustrates the differences in each model's performance across these metrics, revealing their strengths and areas for improvement in user engagement and community support.

Qwen2.5-72B-Instruct demonstrates a balanced performance across all metrics, with particularly high CRR, reflecting a strong balance between user engagement and community support. In contrast, Llama-3.2-1B-Instruct shows high user engagement but a lower CRR, indicating insufficient community interaction.

Additionally, Llama-3.1-70B-Instruct and Qwen2.5-0.5B-Instruct have relatively high Time Weight Factors, indicating they have maintained a long-term user interest. However, their CRR and UEI are relatively low, suggesting there is still room for improvement in community support and user engagement. Overall, high user engagement and active community response are key indicators of a model's performance and influence.

5 Conclusion

This paper proposes a multi-dimensional evaluation framework for open-source LLMs, which uses a comprehensive assessment of metrics such as user engagement, community response rate, and time-weighted factors to reveal differences in model performance in real-world applications. Based on data from the Hugging Face and GitHub platforms, we validated the effectiveness of this evaluation system. Experimental results show that user-oriented engagement and active community support have a significant impact on the final model scores.

In this paper, we observed that models with high user engagement and active community support tend to receive higher final scores, which underscores the importance of user experience and community response in the open-source model ecosystem. However, some models performed poorly in user engagement and community interaction, indicating room for improvement in user-oriented optimization strategies. This evaluation framework not only provides a powerful tool for comprehensive model evaluation but also offers insights for developers and researchers to optimize their model design and user support strategies.

Future work will focus on expanding the evaluation metrics to cover different application scenarios of the models. Additionally, to address the dynamic nature of platform data, future research can explore real-time updates and adaptive optimization methods for evaluation, thereby enhancing the timeliness and adaptability of the evaluation results.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Ding Chen, Shichao Song, Qingchen Yu, Zhiyu Li, Wenjin Wang, Feiyu Xiong, and Bo Tang. 2024. Grimoire is all you need for enhancing large language models. *arXiv preprint arXiv:2401.03385*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. 2024. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. *arXiv preprint arXiv:2405.18638*.

Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Yezhaohui Wang, Dawei He, Peng Cheng, Zhonghao Wang, et al. 2023. Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. *arXiv preprint arXiv:2311.15296*.

Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, et al. 2024. Internal consistency and self-feedback in large language models: A survey. *arXiv preprint arXiv:2407.14507*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Hongkang Yang, Zehao Lin, Wenjin Wang, Hao Wu, Zhiyu Li, Bo Tang, Wenqiang Wei, Jinbo Wang, Zeyun Tang, Shichao Song, et al. 2024. Language modeling with explicit memory. *Journal of Machine Learning*, 3(3):300–346.

Qingchen Yu, Shichao Song, Ke Fang, Yunfeng Shi, Zifan Zheng, Hanyu Wang, Simin Niu, and Zhiyu Li. 2024. Turtlebench: Evaluating top language models via real-world yes/no puzzles. *arXiv preprint arXiv:2410.05262*.

Qingchen Yu, Zifan Zheng, Ding Chen, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025a. Guessarena: Guess who i am? a self-adaptive framework for evaluating llms in domain-specific knowledge and reasoning. *arXiv preprint arXiv:2505.22661*.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu li, Feiyu Xiong, Bo Tang, and Ding Chen. 2025b. xfinder: Large language models as automated evaluators for reliable evaluation. In *The Thirteenth International Conference on Learning Representations*.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025. Attention heads of large language models. *Patterns*.