# FlagEval-Arena: A Side-by-Side Comparative Evaluation Platform for Large Language Models and Text-Driven AIGC

**Jing-Shu Zheng***, **Richeng Xuan***, **Bowen Qin**, **Zheqi He**
**Tongshuai Ren**, **Xuejing Li**, **Jin-Ge Yao**, **Xi Yang**
BAAI FlagEval Team
{jszheng, rcxuan, jgyao, yangxi}@baai.ac.cn

## Abstract

We introduce FlagEval-Arena, an evaluation platform for side-by-side comparisons of large language models and text-driven AIGC systems. Compared with the well-known LM Arena (LMSYS Chatbot Arena), we reimplement our own framework with the flexibility to introduce new mechanisms or features. Our platform enables side-by-side evaluation not only for language models or vision-language models, but also text-to-image or text-to-video synthesis. We specifically target at Chinese audience with a more focus on the Chinese language, more models developed by Chinese institutes, and more general usage beyond the technical community. As a result, we currently observe very interesting differences from usual results presented by LM Arena. Our platform is available via this URL: https://flageval.baai.org/#/arena.

## 1 Introduction

Advances in large language models (LLMs) and the broader field of AI-generated content (AIGC) have been blazingly fast, causing a significant challenge in evaluation. Traditional benchmarks, often static and limited in scope, fail to capture the nuances of real-world interactions. The emergence of LM Arena, or formerly known as the LMSYS Chatbot Arena (Zheng et al., 2023; Chiang et al., 2024) [1], have addressed those limitations to a significant extent. LM Arena is designed to compare and evaluate the performance of various LLMs in a side-by-side fashion. By allowing real users to interact with two models anonymously and to vote afterwards, the platform offers a dynamic and realistic data to assessing model quality.

While being a valuable evaluation platform to the community, LM Arena has some limitations in coverage or usage: (1) LM Arena is most widely known in English context, with limited evaluation and inclusion for non-English languages or cultures (Zheng et al., 2024); (2) Due to its big impact in LLM evaluation, the user base of LM Arena heavily skews toward the technical community, henceforth almost dominantly reflecting the preferences or use cases there. (3) Non-experts, especially those who are still new to modern AI, may struggle to initiate their use of such a system. (4) The four-type coarse-grained voting system[2] offers a limited level of nuance and does not capture the degree of preference or specific strengths/weaknesses (Dhar and Simonson, 2003).

As an attempt to address these limitations, in this paper we describe FlagEval-Arena, our side-by-side platform with additional mechanisms or features. Specifically,

- Our platform uniformly integrates the evaluation for text-driven AIGC, namely large language models, vision-language models, text-to-image and text-to-video generation.
- We implement a new design of UI which is expected to be more lightweight, user-friendly, and also prompting for sligtly more fine-grained expression of preference.
- As beta features, we introduce two new modes: the deep thinking mode involving recent reasoning models, and the multi-models battle enabling more efficient comparisons among a customized number of systems.
- We target at Chinese audience with a more focus on the Chinese language and culture, via promoting our platform on Chinese social media. Based on the current data we collect, we have already observed some interesting trends that differ from LM Arena. Our

* Equally contributed to this project
[1] https://lmarena.ai/
[2] A user will select one of these four possibilities: A is better, B is better, tie, both are bad.



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 583–591

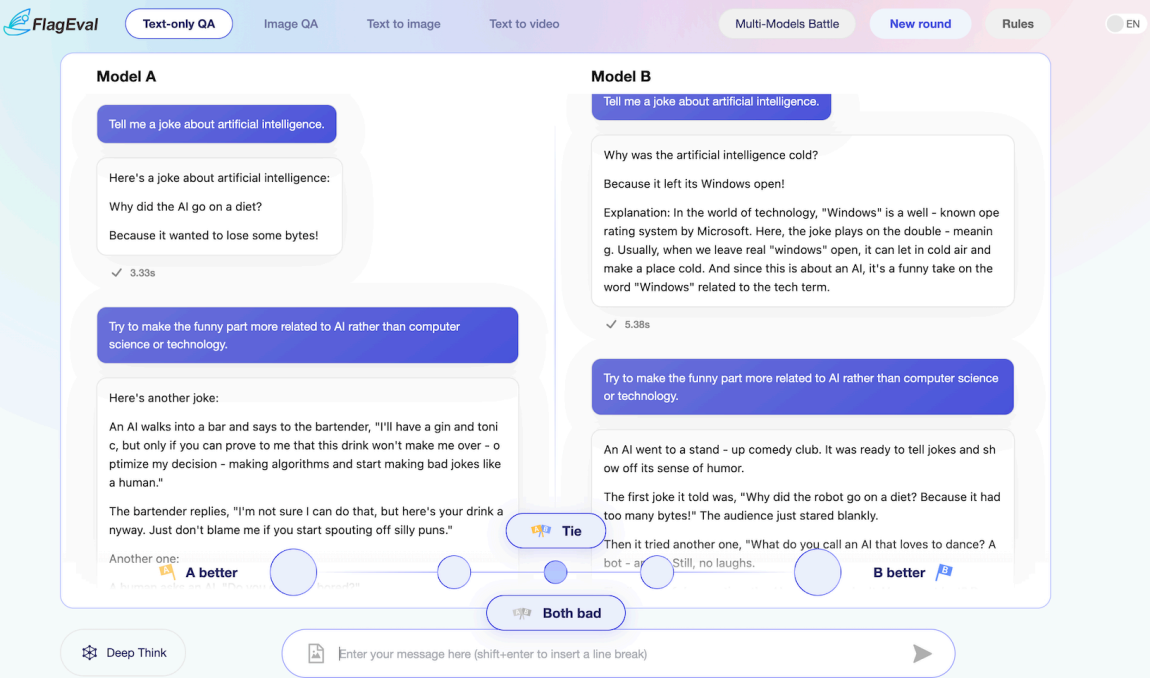


Figure 1: Main User Interface. Users can vote their preference with one simple click.

findings reveal new patterns of usage from a different group of people and cultural context, and also magnifying some key limitations of arena-style side-by-side evaluation.

## 2 User Interface and Functionality

With preference for more flexibility and the convenience of significant modifications, we do not use the FastChat framework[3] open-sourced by the LMSYS team behind LM Arena, although we borrowed some of the key ideas. We instead reimplement our framework from scratch, which enables easier modifications and adaptation. [4]

### 2.1 UI Design

The basic mechanism is the same form of side-by-side comparison as LM Arena: a user provides a prompt, receives two responses from two anonymous systems whose identity will be revealed after voting. We have made some changes based on preliminary user study and the target for a much broader range of audience.

#### 2.1.1 General Display

Rather than a direct adoption of the original Gradio interface[5] in LM Arena, we design a new user interface with a strong preference of visual simplicity, as shown in Figure 1. Apart from the simple UI structure, our platform initially will also provide a randomized set of human-crafted prompts for newbie users to begin with or to learn from, making FlagEval-Arena more friendly to users outside of the technical community. We have also adapted FlagEval-Arena on small-screen mobile devices such as smartphones. See Figure 2 for an instance. The mobile adaptation makes it easier to make a visual query immediately after receiving an image or taking a photo from camera. The default mechanism is that the identities of systems will only be revealed after voting. We have also implemented a mechanism to detect and block identity-revealing responses, and exclude them from data analysis or system ranking.

### 2.2 Multimodality

Apart from the most popular large language models, FlagEval-Arena is designed to integrate many other kinds of AIGC comparison, as one can find on the top of the interface in Figure 1:

- By default, the webpage will land in the *Text-only QA* arena which is intended for comparing standard LLMs.
- In the *Image QA* arena, two Vision-Language Models (VLMs) will be sampled as a user is

[3] https://github.com/lm-sys/FastChat

[4] See also our demo video for a walkthrough: https://www.youtube.com/watch?v=uI2Alx06-gI

[5] https://gradio.app/

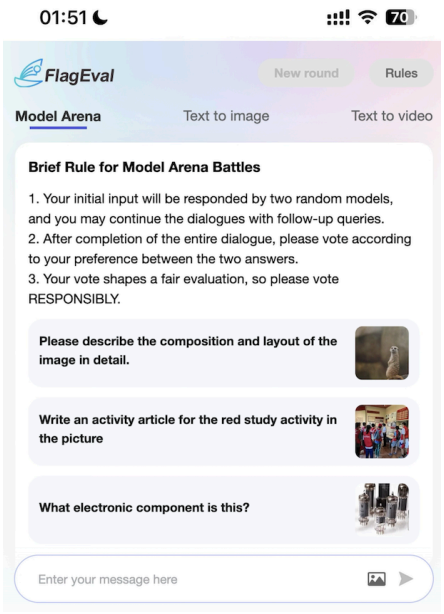


Figure 2: Mobile UI, showing rules and starter prompts

expected to upload an image as the required context for the textual prompt.

- The *Text-to-Image* (T2I) arena accepts a user prompt for image creation and then renders the images from two T2I systems.
- Likewise, the *Text-to-Video* (T2V) arena supports a comparison between two video synthesis systems. Given that current T2V systems usually require too much time to generate the short video, we only enable user voting on offline-generated videos based on a diverse set of pre-defined prompts.

#### 2.2.1 Increased Granularity

The original LM Arena allows a user to vote for one of these four choices: System A is better, System B is better, tie, both are bad, henceforth no mechanism to express the degree of preference. To address limitation while avoiding an increased burden to the voting process, we add one-level of preference degree such that users can cast an easier vote when they are hesitating on a less significant difference between the two systems in comparison (Dhar and Simonson, 2003). As a result, each vote can be made among a choice of six (see also bottom of Figure 1).

### 2.3 New Features

As more and more people gradually identified their preferred LLM products, the incentives of simultaneously using of two systems and voting become decreased, which has been reflected on some declines in our traffic. Starting very recently, we introduce two new beta features. One for a dedicated comparison involving recent deep thinking models, and the multi-model battle which involves more than two systems to respond to increase efficiency in vote collection.

#### 2.3.1 Deep Thinking Mode

One of the most significant recent trends is inference-time scaling, popularized by the o1-series (OpenAI, 2024) from OpenAI. Many model providers start to add a "deep thinking" mode to indicate a different model that spends a significant amount of time in "thinking" before providing the real answer.

Our initial integration of o1-like models did not turn out to be much informative, as we found that our user group have a strong preference over non-thinking models that output an answer much more quickly. Therefore, we specifically design a Deep Thinking mode for more patient users who would like to test for more challenging prompts. That said, more recent reasoning models have become faster and faster, so the chance of two models becoming simultaneously slow would not be high, making it still usable for less patient users in that they can always start reading the response from one candidate while waiting for the other system who may take longer time to reason. In this mode, at least one recent large reasoning model that supports "deep thinking" will be sampled, along with another such system or one of the most advanced non-thinking model. For fair comparison, we do not allow the user to unfold the thinking process until a vote based on the responses has been made (Figure 3).

#### 2.3.2 Battles among Multiple Models

One round of comparison most typically involves a battle of two, making it very clumsy if a user would like to try the same prompt on more candidate models. With this pain point in mind, we introduce a new mode to support multi-model battles, enabling a direct comparison of a customized number of systems using one same prompt. To express preference for more than two candidates, we change the simple one-click preference voting to a pointwise 5-scale rating. As shown in Figure 4, once a rating has been received, the battle is considered to be complete, and the identity will be displayed after each rating. Since the ratings are

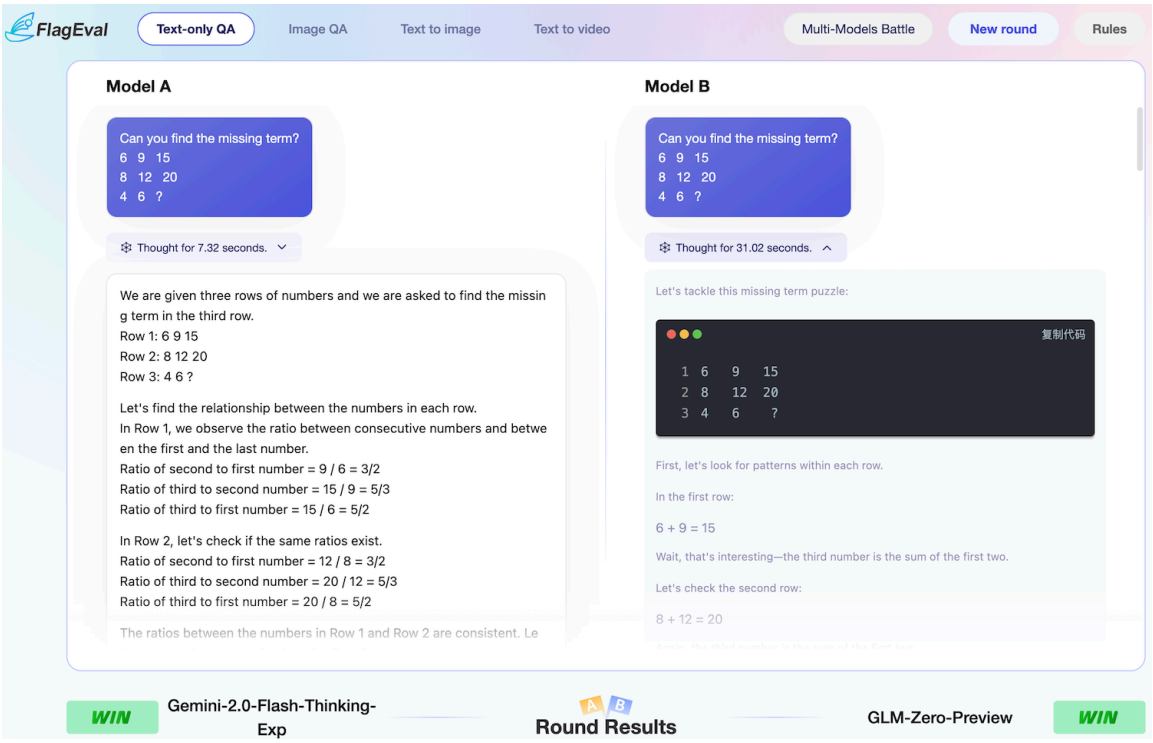


Figure 3: Deep thinking mode: once voting is complete, the thinking process could be unfolded, if provided.

made among comparisons with many other candidates, we do not interpret the 5-scale rating as absolute scoring. As one round involving $K$ models will induce $\binom{K}{2}$ pairwise preference votes if all of them get rated, this new mode will largely increase prompt efficiency in terms of gathering voting data.

## 3 Results and Preliminary Analysis

Given that LM Arena has been a solid platform to characterize the user group of the entire technical community with English being the major form of data (Zheng et al., 2024), the main motivation for us to build and deploy another arena platform is the will to target for a different user group, mostly for the Chinese language, social context, and beyond a narrow range of technical members.

### 3.1 The Different Group of Audience

Although we are showing the English UI screenshots in this paper for the convenience of readers worldwide, the most dominant use of our platform is in its Chinese UI, which is structurally the same but all phrases displayed in Chinese. To attract more votes from a more diverse range of real-world users outside the AI community, we have promoted our platform on many social media channels in China such as WeChat and Douyin (Chinese version of TikTok). Launched in late September 2024, we have collected tens of thousands of valid votes and the votes are still growing.

We conduct analysis to better understand the usage in the Chinese context. Take text-only usage for instance, we identified a group of classes via clustering and manually named those categories, as shown in Table 1. We also conduct a manual labeling process on a sample of around 2k prompts to understand the distribution. [6] Different from a much coarse-grained categorization of use cases as reported in LM Arena (Li et al., 2024b), we can observe that our targed user group was dominated by information seeking and writing queries. This conforms to similar findings on Chinese usage reported by Anthropic based on their analysis on Claude traffic (Tamkin et al., 2024).

### 3.2 More Preferences Expressed

Another notable difference from LM Arena is that we observe much fewer tie votes. In Table 2, we list the percentages from a sample of LM Arena (shared in Chiang et al. (2023)) and ours from FlagEval-Arena. The huge difference is most likely a direct cause from our new UI design that

[6] Our earlier attempts with cost-effective LLMs turn out to be rather error-prone for longer Chinese prompts, thus we opt for manual sample analysis at this stage.

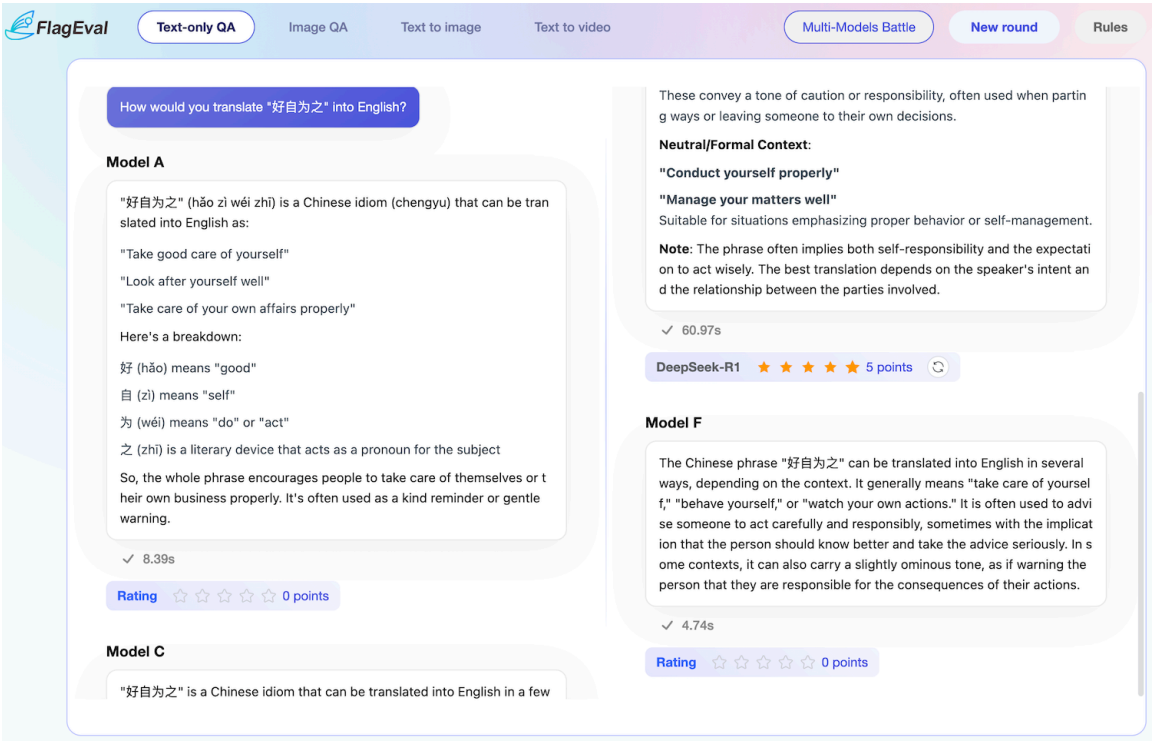


Figure 4: Multi-models battle: model identity revealed only after rating

| Category | Pr(%) | Description |
|---|---|---|
| Info. seeking | 68.01 | Often starting with "how", "please explain", etc. for searching |
| Writing | 13.46 | Constrained, contextual or creative writing |
| Program-related | 8.03 | Queries related to code generation, explanation, or debugging |
| Factoid Q&A | 2.92 | Common knowledge QA such as "When was World War I?" |
| Academic Q&A | 2.51 | Asking domain-specific knowledge in various academic subjects |
| Reasning puzzles | 1.94 | Mainly includes fun reasoning questions and brain teasers. |
| Math problems | 1.69 | Standard math problems |
| Multilingual | 0.97 | Translation, summariation, or Q&A from a different language |
| Situational consulting | 0.72 | Related to the user's emotions or assumptions |
| Process/format | 0.62 | Data analysis and formatting requests |

Table 1: Identified Dominant Categories and Descriptions

| Voted for | LM Arena | FlagEval-Arena |
|---|---|---|
| A preferred | 31.59% | 41.02% |
| Tie | 18.75% | 8.64% |
| Tie (both bad) | 17.16% | 6.27% |
| B preferred | 32.50% | 44.07% |

Table 2: %Votes from LM Arena and FlagEval-Arena

has added one-level of granularity, making users more leaning towards an indicated preference between two candidates.

### 3.3 System Ranking

We mainly include modern systems developed by companies in China for comparison based on APIs provided by their official services.[7] We applied the Bradley-Terry models (Bradley and Terry, 1952) as adopted by LM Arena (Chiang et al., 2024, 2023) with reweighing to utilize our more fine-grained votes that contain a different strength in preference: [8]

$$w = \begin{cases} 1, & \text{A is much better} \\ 0.75, & \text{A is better} \\ 0.5, & \text{Tie} \\ 0.25, & \text{B is better} \\ 0, & \text{B is much better} \end{cases} \quad (1)$$

We find that on Chinese-oriented data with use cases focusing more on information seeking and writing, the ranking generally differs from a tech-focused ranking, as the latest strongest models can produce generally correct or useful responses to such queries, making it difficult to distinguish. Interestingly, some of the strongest models in English (e.g., Claude 3.5 series (Anthropic, 2024)) failed to join the best performed systems in Chinese. [9] We provide current rankings (as of March 2025) of all four arena settings in Appendix (Sec. B, truncated due to space limit).

#### 3.3.1 The Impact of Style

We have also controlled for the effect of style by adding extra length and "style features" into the Bradley-Terry regression process. This is the standard technique in statistics, and has been used in LLM evaluations (Dubois et al., 2024). The general idea is to include confounding variables in BT regression, in order to attribute any increase in strength to the confounder, as opposed to the model. We use the normalized length difference, the number of markdown headers, the number of lists, and the number of bolded texts, following LM Arena (Li et al., 2024a). We find that the controlled scores from different LLMs are even closer, with many systems staying in the same band, while the ranking of some style-heavy or lenthy systems drops from the top. Interestingly, the controlled scores for VLM become slightly more diverged, indicating that image QA testing more on visual capabilities might be more differentiable among current systems.

Do the changes indicate that the votes from our targeted user group are heavily affected by output length and style? We suggest to take a grain of salt on this interpretation, as style control analysis only suggests a strong correlation between style and user voting, rather than causation. On the one hand, the style of language is usually more subtle or complex than lengths or fonts (e.g., more recent discussion on sentiment (Chen et al., 2025)) while model developers can optimize for "aesthetics" (Jiang et al., 2024) in various ways. On the other hand, a qualitative sample analysis on the platform suggests a potential trend that models producing well-formatted responses are usually also more comprehensive or caring in terms of content, which might be a signal that better LLMs are partially driven by better product mindsets and stronger model development. Limited by current scale of usage, we prefer to leave more convincing conclusions upon further analysis in the future after we get more traffic.

### 3.4 Limitations

While addressing some limitations of LM Arena, our FlagEval-Arena inherited a few notable weaknesses as any current arena-style system, including but not limited to: relative shortage of multi-turn usage, sensitive to sample size and domain shift, noise in user voting, human voting bias, etc. While we are working on further improvements, we would prefer to promote our platform to a broader audience inside more specialized communities to gather more difficult prompts that can help distinguish between top-tier models.

[7] We also include some of the most well-known open-weight models and API-based systems via third-party providers with verified validity.

[8] Preliminary analysis on current data does not show notable difference in final ranking had we ignored the strength of preferences. It might be fair to say that our new UI contributes more in decreasing tie votes that are not informative for ranking.

[9] This conforms to the LM Arena leaderboard `https://lmarena.ai/?leaderboard` in "Chinese" category.

## 4 Related Work

Our FlagEval-Arena is directly motivated by the well-known LM Arena, also known as the LMSYS Chatbot Arena (Zheng et al., 2023; Chiang et al., 2024). We adopt basically the same statistical methods to induce ranking (Angelopoulos et al., 2024) and style control mechanism as used by LM Arena (Li et al., 2024a). For video generation, we later realized that the LMSYS team have also released VideoArena (LMSYS, 2024) in a separate website. The design resembles popular short video platforms, making annotation fast and addictive. Our FlagEval-Arena support comparisons for text-to-video systems on Day 1 since released. We are studying the strengths and weaknesses of the different scheme before a decision on whether to migrate towards that direction. There are also related efforts originated from Chinese institutes (Team, 2023; OpenCompass, 2024). Built on FastChat, the focus there is more on LLMs/VLMs among technical community, while we are keen on a better initial understanding of domestic AIGC usage comprehensively. We are also happy to see that our UI design has inspired a recent change in the CompassArena UI (OpenCompass, 2024). Additionally, there are studies suggesting potential bias for pretty and more detailed responses in humans (Chen et al., 2024; Park et al., 2024). Moreover, more recent studies have revealed potential vulnerability to ranking manipulation (Huang et al., 2025; Singh et al., 2025). We are working on more understanding to what extent human bias might affect our new features, along with close monitoring on potentially unusual traffic or votings while strictly limiting the number of systems involved from the same organization.

## 5 Conclusion

We present FlagEval-Arena, our side-by-side evaluation platform with a different targeted audience from the well-known LM Arena. Our simpler design makes it more natural to use and to express a preference, while current findings also reveal interesting behavioral differences from a Chinese-centric user group. We will continue our analysis once some of our new features have gathered sufficient traffic, especially on some potentially new trends on our deep thinking mode. We are working on a more detailed report to describe more detailed or principled analysis, and also plan to release part of our collected data and accompanied evaluation scripts to the public under a permissive license, after more accumulation plus necessary post-processing to filter out sensitive or personally identifiable information.

## Ethic Statement

Like any modern AIGC system or service, since our platform directly provides an interface for comparing AIGC systems, it could theoretically be used by malicious users for malicious purposes, along with potential concerns on copyright. While relying on AIGC service providers for governance and safety control, we have also adopted a safety-aware module on our side to block unsafe model output. That said, there would be no guarantee for a safeguard given various kinds of strategies of malicious prompting or jailbreaking known or unknown in the community.

## Acknowledgments

We sincerely thank Hui Wang for his massive and solid contributions to this project. His name should have appeared in the author list had OpenReview adopted a more efficient scheme for profile creation and verification. We thank all team members at our FlagEval team and the anonymous reviewers for feedback of the platform or the earlier draft of this paper. We also thank Bingrong Lyu and Xiaojing Xu for assistance in preliminary analysis. This work was supported by the National Science and Technology Major Project of China (Grant No. 2022ZD0116306). The observations, perspectives, and opinions expressed in this paper are those of the authors and do not necessarily reflect those of their institutions or funders.

## References

Anastasios Nikolas Angelopoulos, Wei-Lin Chiang, and Shishir Patil. 2024. Statistical extensions of the Bradley-Terry and Elo models.

Anthropic. 2024. Claude 3.5 Sonnet.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Connor Chen, Wei-Lin Chiang, Tianle Li, and Anastasios Angelopoulos. 2025. Does sentiment matter too?

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs

as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Wei-Lin Chiang, Tianle Li, Joseph E. Gonzalez, and Ion Stoica. 2023. Chatbot arena - new models & Elo system update.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

R. Dhar and I. Simonson. 2003. The effect of forced choice on choice. *Journal of Marketing Research*, 40:146–160.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*.

Yangsibo Huang, Milad Nasr, Anastasios Angelopoulos, Nicholas Carlini, Wei-Lin Chiang, Christopher A. Choquette-Choo, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Ken Ziyu Liu, Ion Stoica, Florian Tramer, and Chiyuan Zhang. 2025. Exploring and mitigating adversarial manipulation of voting-based leaderboards. *Preprint*, arXiv:2501.07493.

Lingjie Jiang, Shaohan Huang, Xun Wu, and Furu Wei. 2024. Textual aesthetics in large language models. *Preprint*, arXiv:2411.02930.

Tianle Li, Anastasios Angelopoulos, and Wei-Lin Chiang. 2024a. Does style matter? disentangling style and substance in chatbot arena.

Tianle Li, Wei-Lin Chiang, Yifan Song, Naman Jain, Lisa Dunlap, Dacheng Li, Evan Frick, and Anastasios N. Angelopoulos. 2024b. Chatbot arena categories: Definitions, methods, and insights.

LMSYS. 2024. VideoArena.

OpenAI. 2024. Introducing OpenAI o1-preview.

OpenCompass. 2024. CompassArena.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017, Bangkok, Thailand.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2025. The leaderboard illusion. *Preprint*, arXiv:2504.20879.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, and 2 others. 2024. Clio: Privacy-preserving insights into real-world ai use. *Preprint*, arXiv:2412.13678.

SuperCLUE Team. 2023. SuperCLUElyb.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

## A More on Style Control

In fitting the Bradley-Terry model, we found the linear coefficients for style features to be informative, henceforth showing them here along with the corresponding coefficients from LM Arena data samples released by Li et al. (2024a). In Table 3 we can see notably larger coefficients for length features and all markdown style features, indicating stronger correlation between better formatting and user preference in general Chinese usage.

| Style feature | LM Arena | FlagEval-Arena |
|---|---|---|
| Length | 0.191 | 0.231 |
| Header | 0.043 | 0.156 |
| Lists | 0.010 | 0.077 |
| Bold fonts | -0.001 | 0.170 |

Table 3: Linear coefficients of style regression

## B Current Leaderboards

In Table 4-7, we display the partial leaderboards (as of March 2025) for four types of AIGC models: text-only LLMs, vision-language models, text-to-image and text-to-video generation. Note that results from some very recent models or a few provided by relatively unstable service have been excluded due to high variance. We are working on solutions to gather more valid votes form them, and release more comprehensive results and analysis in an extended report.

Table 4: FlagEval-Arena Top-10 LLMs with sufficient votes

| Rank(UB) | Rank(SC) | Model | Score | 95% CI | Votes |
|---|---|---|---|---|---|
| 1 | 3 | o1-mini-2024-09-12 | 1149.16 | +11.62 / -13.04 | 3207 |
| 1 | 2 | Doubao-pro-32k-240828 | 1135.29 | +10.09 / -10.70 | 3092 |
| 1 | 3 | Nanbeige2-Turbo-0611 | 1132.28 | +11.26 / -11.53 | 3494 |
| 1 | 1 | GLM-4-Plus | 1124.81 | +14.28 / -11.28 | 2103 |
| 2 | 3 | Yi-Lightning | 1112.56 | +10.84 / -13.21 | 2306 |
| 2 | 2 | DeepSeek-V3 | 1094.90 | +11.63 / -11.72 | 4344 |
| 3 | 1 | Hunyuan-Turbo | 1090.50 | +12.61 / -12.73 | 2091 |
| 3 | 3 | o1-preview-2024-09-12 | 1074.56 | +9.81 / -9.69 | 3115 |
| 4 | 3 | GPT-4o-2024-08-06 | 1069.80 | +12.30 / -11.92 | 3263 |
| 4 | 4 | Gemini-1.5-pro | 1045.03 | +12.39 / -7.87 | 3645 |

Table 5: FlagEval-Arena Top-10 VLMs with sufficient votes

| Rank(UB) | Rank(SC) | Model | Score | 95% CI | Votes |
|---|---|---|---|---|---|
| 1 | 1 | GPT-4o-2024-11-20 | 1063.78 | +15.38 / -15.81 | 241 |
| 1 | 2 | GPT-4o-2024-08-06 | 1054.80 | +15.42 / -18.95 | 325 |
| 1 | 3 | Hunyuan-Vision | 1047.37 | +14.22 / -13.36 | 512 |
| 1 | 3 | Step-1V-32k | 1037.28 | +12.80 / -10.86 | 245 |
| 2 | 3 | Step-1.5V-Mini | 1029.23 | +16.45 / -12.36 | 244 |
| 2 | 6 | Claude-3.5-Sonnet-20240620 | 1017.85 | +20.90 / -15.28 | 194 |
| 3 | 3 | Qwen-VL-Max-0925 | 997.25 | +12.20 / -10.53 | 535 |
| 3 | 4 | GLM-4V-Plus | 996.36 | +13.89 / -11.31 | 310 |
| 3 | 3 | Qwen-VL-Plus-1105 | 988.35 | +12.76 / -14.39 | 506 |
| 3 | 6 | Gemini-1.5-Pro | 986.99 | +12.02 / -16.44 | 438 |

Table 6: FlagEval-Arena Top Text2Image models with sufficient votes

| Rank(Abs) | Rank(UB) | Model | Score | 95% CI | Votes |
|---|---|---|---|---|---|
| 1 | 1 | Kolors | 1076.29 | +24.79 / -20.11 | 3035 |
| 1 | 2 | Doubao Image v2.0 | 1047.79 | +19.79 / -23.67 | 3047 |
| 2 | 3 | DALL-E3 | 1009.46 | +25.89 / -18.23 | 2826 |
| 2 | 4 | CogView3 | 1001.03 | +23.37 / -21.79 | 2822 |
| 2 | 5 | SenseMirage | 983.67 | +14.92 / -21.27 | 2983 |
| 3 | 6 | Hunyuan-Image | 969.11 | +18.03 / -16.03 | 3049 |

Table 7: FlagEval-Arena Top Text2Video models with sufficient votes

| Rank(Abs) | Rank(UB) | Model | Score | 95% CI | Votes |
|---|---|---|---|---|---|
| 1 | 1 | Kling 1.5 | 1173.60 | +22.47 / -15.57 | 328 |
| 2 | 2 | MiniMax 01 | 1108.18 | +19.65 / -13.35 | 847 |
| 3 | 2 | Runway Gen-3 | 1078.43 | +17.02 / -14.80 | 1201 |
| 4 | 3 | GLM-video | 1073.37 | +15.22 / -15.81 | 1183 |
| 5 | 3 | Jimeng P 2.0pro | 1056.64 | +14.14 / -14.77 | 1256 |
| 6 | 4 | Pixeling | 1017.97 | +17.55 / -18.04 | 1198 |
| 7 | 4 | Sparks-Video | 1017.97 | +16.76 / -21.19 | 1241 |
| 8 | 4 | Dream Machine 1.6 | 1004.68 | +17.90 / -18.33 | 1239 |
| 9 | 4 | WAN | 1000.36 | +16.04 / -18.03 | 640 |
| 10 | 5 | Kling 1.0 | 968.95 | +13.61 / -17.58 | 1274 |