SEAL Showdown
Technical Report

Scale AI

September 2025

Contents

Abstract

We introduce SEAL Showdown, a live leaderboard designed to reflect human preferences over large language models (LLMs) in a natural chat setting. Unlike static benchmarks, Showdown collects preference data *in situ* by periodically prompting users to compare responses from their current model with a randomly selected opponent. We rank models using the Bradley-Terry model, which we augment with style controls to account for confounding factors such as response length, Markdown formatting, and loading time. Our preliminary results place GPT-5 Chat at the top of the leaderboard, followed by Claude Opus 4.1. Our analysis reveals strong user preferences for certain style features, such as response length and formatting. We also find that models with extended thinking capabilities do not consistently outperform their non-thinking counterparts, suggesting that additional test-time compute offers limited benefit for everyday conversational tasks. By capturing user interactions in a natural conversational setting, Showdown complements existing LLM benchmarks by offering a transparent and dynamic view into human preferences in the wild. For real-time model rankings, please visit https://scale.com/showdown.

1 Introduction

The SEAL Showdown leaderboard is a ranking over large language models (LLMs) designed to reflect human preferences in a natural chat setting. We collect human preference data *in situ* using an internal, LLM-API-agnostic chat application where registered human users can freely converse with a variety of open- and closed-source models, and may freely switch between available language models at any time.

During conversations, users are periodically prompted to participate in a side-by-side model comparison. Users are given the option to skip a comparison before any model responses are revealed to reduce bias. If the user agrees to continue, they are served a pair of model responses. One response comes from the model the user was actively conversing with (the *in-flow* model), while the other comes from an opponent selected by a sampling strategy (the *out-of-flow* model). This approach differs from prior work, which typically samples both models in a pair, with no prior conversation context [9]. Both responses are streamed in tandem, and the identities of both models are hidden from the user during the side-by-side comparison.

The user reports their preference between responses by choosing one of four options: "left preferred," "right preferred," "both good," and "both bad" (Figure 1). After their selection, the model identities are revealed and the conversation continues, using the winning response—or the in-flow response, in the case of a tie—in context. The user is also given a nudge by the chat interface to switch to the winning model, but is not forced to change models. Thus, there may be multiple model endpoint changes throughout the conversation, and a single conversation may contain multiple model comparisons.

The side-by-side comparison results are then used to generate approximate rankings. Showdown ranks models using the Bradley-Terry model [5]. We detail the methodology used to select model pairings, obtain Elo ratings, and control for confounding factors such as response length, Markdown formatting, and loading time in Section 2.

As of September 20, 2025, GPT-5 Chat is the top ranked model on Showdown, followed by Claude Opus 4.1. In Section 3, we present our preliminary ranking, followed by analyses of the effect of style on user preferences; the distribution of users and prompts; and the relative performance of thinking versus non-thinking model variants.

To strengthen our confidence in the robustness of our rankings, we make commitments regarding model selection and deprecation, data confidentiality, and user privacy safeguards. We describe these policies in Section 4.

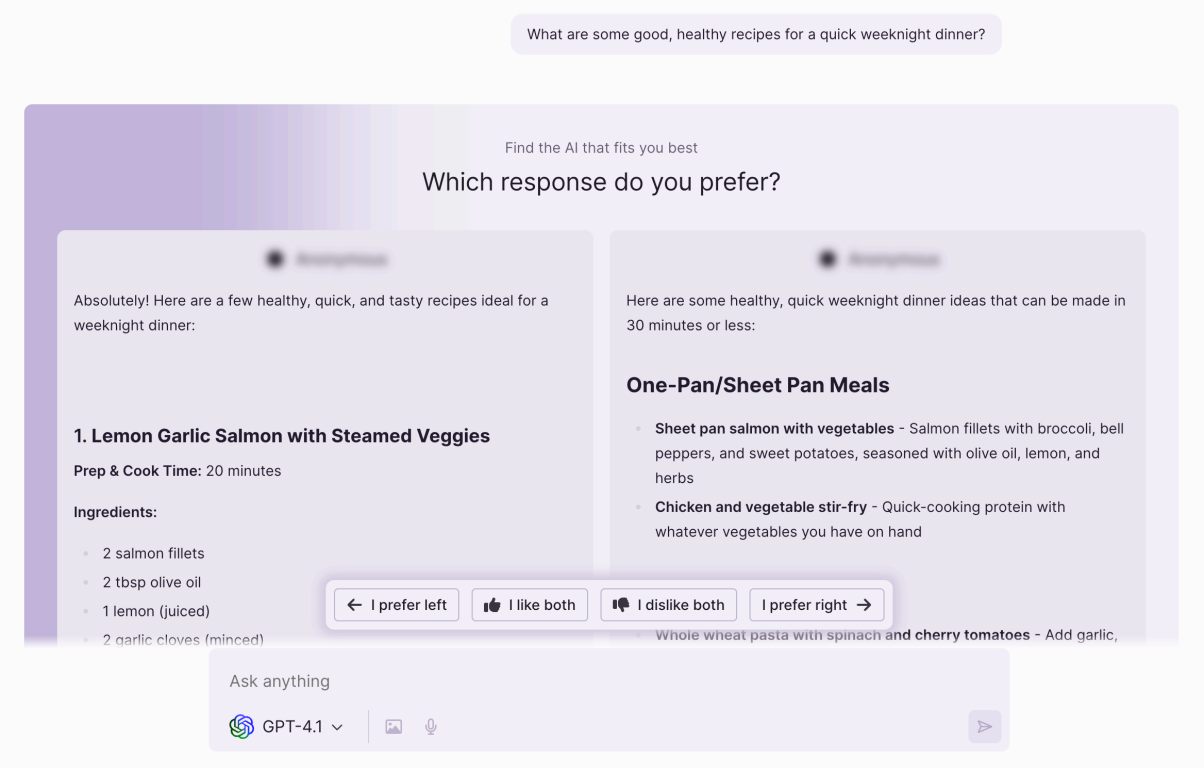

Figure 1: SEAL Showdown's model comparison interface.

2 Methodology

Our methodology is designed to produce a robust ranking of large language models by controlling for common biases. First, we collect model comparison data using a sampling strategy that prioritizes under-evaluated pairs to ensure the data is balanced. We then rank the models by estimating their strength using the Bradley-Terry model, which calculates a score for each model based on its head-to-head performance. Finally, to ensure our rankings reflect underlying capabilities rather than superficial aspects of presentation, we enhance our model to explicitly control for stylistic factors such as response length, formatting, and loading time.

2.1 Data collection

In Showdown, M distinct LLMs are evaluated through a sequence of N pairwise comparisons. Each comparison consists of a model pair (m_1, m_2) and human rating h. The models m_1 and m_2 are sampled from the set of valid model pairings, $\mathcal{A}$. The human rating is a value $h \in \{0, 0.5, 1\}$ assigned by a human evaluator. A value of 1 indicates that the response from m_1 is preferred over the response from m_2, denoted $m_1 \succ m_2$. Conversely, a value of 0 indicates that m_2 is preferred over m_1, denoted $m_1 \prec m_2$, and 0.5 indicates a tie, $m_1 \sim m_2$. The collection of N records is denoted as $\mathcal{D} = \{(m_1^{(k)}, m_2^{(k)}, h^{(k)})\}_{k=1}^{N}$, and forms the dataset used for ranking and analysis.

During a comparison, users are presented with the in-flow model and one out-of-flow model. Thus, one of (m_1, m_2) is fixed. From a ranking standpoint, however, we wish to avoid oversampling models that are more popular on Scale's LLM chat platform, as the underlying popularity distribution is non-uniform (Figure 2, left). One appropriate way to allocate comparisons is over model *pairs*, by prioritizing under-evaluated and high-variance pairs [9, 39], and is formally defined as

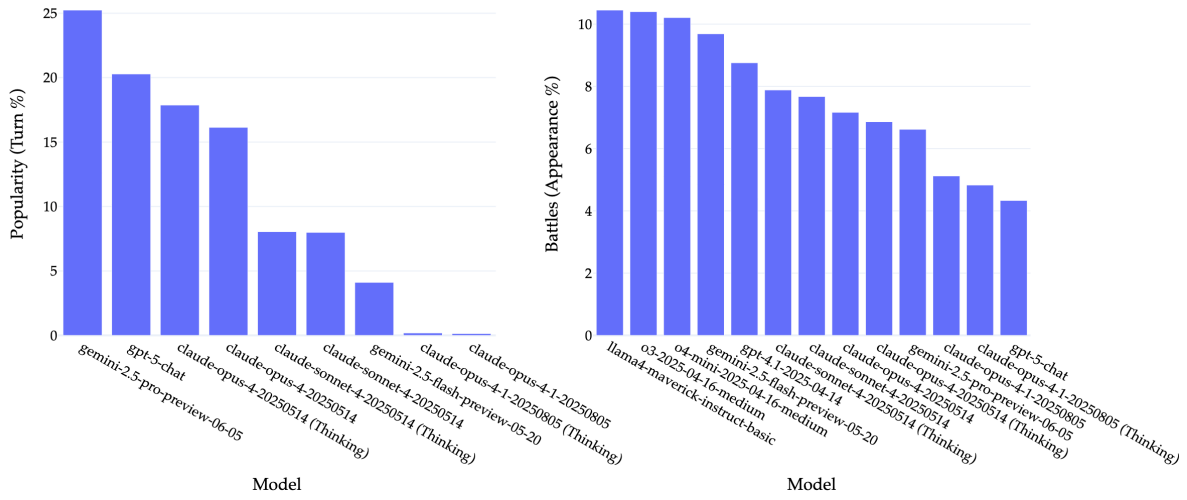

Figure 2: Popularity (*left*) and battle sampling distribution (*right*) for Showdown models. Popularity measures the number of turns where users converse with a given model. While the popularity distribution heavily favors GPT-4.1, the battle sampling distribution does not follow the popularity distribution and exhibits less skew.

$$P_t(a) \propto \sqrt{\frac{\Sigma_{t,a,a}}{N_t(a)}} - \sqrt{\frac{\Sigma_{t,a,a}}{N_t(a)+1}} \tag{1}$$

where $P_t(a)$ is the probability of sampling the pair $a \in \mathcal{A}$ at time t, $\Sigma_{t,a,a}$ is the estimated variance for the win-rate of a at time t, and $N_t(a)$ is the number of times the pair a has been selected up to time t.

Thus, we sample model pairs using a two-step strategy. First, we sample a pool of model pairs according to P_t. Next, we opportunistically serve model pairs from the pool to active users by finding matches, i.e., assigning the pair (m_1, m_2) to a user conversing with m_1 or m_2. While the resulting sampling probability does not match the active sampling probability P_t because of the second step, we observe that the battle distribution is more balanced than the popularity distribution in practice (Figure 2, right).

2.2 Ranking preliminaries

To generate rankings for M models from $\mathcal{D}$, we use the Bradley-Terry model, which models the outcome of a comparison by assigning a strength coefficient to each competitor. In our setting, each model m is assigned a strength coefficient $\beta_m \in \mathbb{R}$, and the probability of $m_1 \succ m_2$ is modeled as

$$P(m_1 \succ m_2) = \sigma(\beta_{m_1} - \beta_{m_2})$$

where $\sigma(\cdot)$ is the logistic function, $\sigma(x) = 1/(1+\exp(-x))$.

The coefficients β are estimated using maximum likelihood estimation, minimizing the expected cross-entropy loss

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{N} \sum_{k=1}^{N} \ell\big(h^{(k)}, \sigma(\beta_{m_1^{(k)}} - \beta_{m_2^{(k)}})\big) \tag{2}$$

where ℓ represents the binary cross-entropy loss between the predicted win probability, $P(m_1^{(k)} \succ m_2^{(k)})$, and the observed outcome, $h^{(k)}$, when it is not a tie ($h^{(k)} \in \{0, 1\}$). For tie outcomes ($h^{(k)} = 0.5$), we naturally generalize the loss to a soft target.

Because the Bradley-Terry model is under-determined with respect to constant shifts in the strength coefficients, we designate, without loss of generality, an *anchor model* m_0 for which $\beta_{m_0} = 1000$. For Showdown, we anchor on Llama 4 Maverick. The coefficients are then adjusted using a scaling factor of 400 to obtain Elo ratings [12].

To produce a ranking from the strength coefficient estimates, we compute a confidence interval for each coefficient via bootstrapping. We then assign an approximate rank to model m by counting the number of competitors it outperforms, ignoring competitors with overlapping confidence intervals:

$$\text{rank}(m) = 1 + \sum_{m' \in \{1,\ldots,M\}} \mathbb{I}[m' \succ m]$$

Thus, a model has rank 1 if it is not outperformed by any competitor, and has rank $n + 1$ when it is outperformed by n competitors.

2.3 Style control

How information is presented can strongly influence user preferences; a user's choice may not be based purely on the factual accuracy of a response, but also on stylistic elements such as its length, tone, or formatting. In Showdown, we empirically observe that response length has a measurable effect on win rates. When the m_1 response is 2,000 tokens shorter than the m_2 response, it wins only 20% of the time; when it is 2,000 tokens longer, its win rate rises to 67% (Figure 3).

To measure preferences over core model capabilities, it is crucial to disentangle these stylistic effects from the substance of the response. To this end, we introduce controls for stylistic features into our model. We begin with a baseline model and augment it with style controls.

Baseline model Starting with Equation 2, we add a fixed intercept and an **in-flow indicator** to capture any effects from a model being part of the ongoing conversation.

Style-control model Next, we add the following style features to the logistic regression:

1. **Token count difference**. The difference in the number of tokens between the two responses. This feature accounts for variation in verbosity across models.
2. **Markdown formatting difference**. The difference in the count of Markdown elements in each response's parse tree. This feature measures the degree to which each model's response is well-formatted, as our chat interface presents Markdown-formatted text as rich text.
3. **Loading time difference**. The difference in the time taken to load each response. Because both responses are streamed in real-time, this feature accounts for the delay between the faster and slower model in each comparison.

Therefore, we augment Equation 2 by adding s additional style parameters, $\gamma \in \mathbb{R}^s$, and style features, $\phi \in \mathbb{R}^s$.

$$\hat{\beta}, \hat{\gamma} = \arg\min_{\beta,\gamma} \frac{1}{N} \sum_{k=1}^{N} \ell\left(h^{(k)}, \sigma(\beta_{m_1^{(k)}} - \beta_{m_2^{(k)}} + \gamma^\top \phi^{(k)})\right) \tag{3}$$

By directly controlling for the style elements, we learn strength coefficients that are more robust to variation in style.

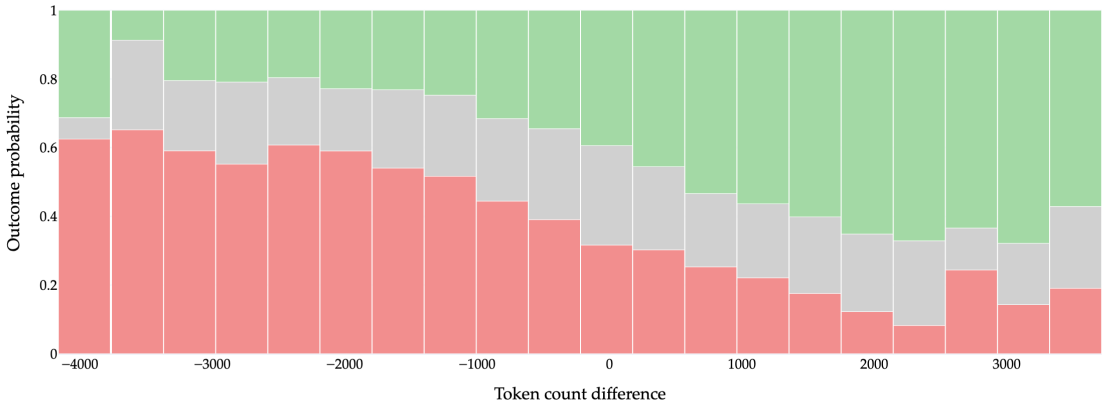

Figure 3: Comparison outcome by difference in token count (left - right). The win rate increases as the difference in token counts grows. Green denotes wins for the left model ($m_1 \succ m_2$); grey, ties ($m_1 \sim m_2$); and red, losses ($m_1 \prec m_2$).

3 Results

Our initial release tests a selection of language models from a variety of providers. These models are evaluated across different capability tracks, including those with and without "thinking" capabilities. Over time, we will gradually phase out older models as their successors accumulate sufficient samples to be included in the leaderboard (Section 4).

Our preliminary ranking results are shown in Table 1. As of launch, GPT-5 Chat tops the leaderboard, followed by Claude Opus 4.1. To better understand the underlying data, we first examine the effects of styles on user preferences (Section 2.3), then analyze the prompt and user distributions (Sections 3.2 and 3.3). Following this analysis, we observe an interesting result concerning thinking versus non-thinking models: although thinking variants appear to be favored on some prompts, most prompts do not show a consistent advantage from thinking. (Section 3.4). We further discuss formatting issues with some model APIs that appear to affect their rankings (Section 3.5).

3.1 Effect of style on user preferences

Our preliminary analysis of feature effects in the style-controlled model shows that user preference increases with response length, richer formatting, and longer loading times. While the preference for length and formatting is expected [35, 40], the positive correlation with loading time is noteworthy. We hypothesize that users may not prefer latency itself, but rather subconsciously associate the longer wait with a more thorough or higher-quality response, perceiving it as the model "thinking" more deeply. However, it's also possible our style-control model still leaves out unobserved confounders that could explain the observed correlations between these styles features and user preferences.

Ablations with individual style features (Table 2, Table 8) further show how style control helps alleviate the effects of style features.

1. Controlling for loading time penalizes slower models. For instance, after applying this control, the rank gap between Gemini 2.5 Pro and the faster Gemini 2.5 Flash narrows. Similarly, non-thinking variants rank higher than thinking variants after controlling for loading time. We discuss this issue further in Section 3.4.

2. Adjusting for response length penalizes models that tend to be more verbose. Gemini 2.5 Pro and Gemini 2.5 Flash, for example, often produce longer outputs (Appendix A.2) and consequently see their rankings decrease after this control is applied, indicating that their baseline preference is partly driven by a user bias for longer responses.

3. Accounting for Markdown usage helps separate a response's substance from its presentation. This control tends to boost the rankings of models that provide raw or plainly formatted text: o4-mini sees a significant improvement in Elo score, while o3 sees improvements in both Elo score and ranking. See Figure 8) for Elo score changes. We discuss further in Section 3.5.

Rank	Model	Score	CI
1	gpt-5-chat	1112.6	+9.1/-7.3
2	claude-opus-4-1-20250805	1093.6	+9.0/-6.9
3	claude-sonnet-4-20250514	1075.0	+6.4/-4.7
3	gpt-4.1-2025-04-14	1068.5	+7.1/-4.5
3	claude-opus-4-1-20250805 (Thinking)	1067.2	+7.9/-9.7
3	claude-opus-4-20250514	1065.8	+5.9/-7.5
4	gemini-2.5-pro-preview-06-05	1059.8	+8.4/-6.9
8	claude-opus-4-20250514 (Thinking)	1039.3	+6.3/-7.0
8	claude-sonnet-4-20250514 (Thinking)	1032.9	+5.4/-5.6
8	gemini-2.5-flash-preview-05-20	1028.0	+7.1/-5.7
9	o3-2025-04-16-medium	1025.0	+5.3/-5.6
12	llama4-maverick-instruct-basic	1000.0	+5.5/-5.8
12	o4-mini-2025-04-16-medium	993.1	+5.4/-5.2

Table 1: The Showdown leaderboard as of September 20, 2025. For each Claude model, we host two versions: one with and one without extended thinking (denoted by the "Thinking" tag). For reasoning models that have multiple levels of thinking effort, we use the default setting and mark it as a suffix ("-medium"). All other models are queried with the default settings.

Model	**Load Time**	**Length**	**Markdown**	**Combined**
gemini-2.5-pro-preview-06-05	1→2	1→4	1→1	1→4
gpt-5-chat	2→1	2→1	2→2	2→1
claude-opus-4-1-20250805	2→4	2→1	2→3	2→2
claude-opus-4-1-20250805 (Thinking)	2→7	2→3	2→3	2→3
gemini-2.5-flash-preview-05-20	5→3	5→10	5→3	5→8
claude-sonnet-4-20250514	5→4	5→3	5→5	5→3
gpt-4.1-2025-04-14	5→3	5→3	5→5	5→3
claude-opus-4-20250514	6→7	6→3	6→5	6→3
claude-opus-4-20250514 (Thinking)	6→9	6→8	6→5	6→8
claude-sonnet-4-20250514 (Thinking)	7→8	7→9	7→9	7→8
o3-2025-04-16-medium	9→11	9→10	9→5	9→9
o4-mini-2025-04-16-medium	12→13	12→13	12→12	12→12
llama4-maverick-instruct-basic	13→11	13→12	13→13	13→12

Table 2: The effect of different style control features on the ranking. Models are ordered by their ranking under the vanilla ranking model, i.e., without style control. For more detailed Elo score changes, see Figure 8.

3.2 Prompt distribution

Showdown prompts are primarily composed of conversational tasks, including Open QA, Closed QA, Chitchat, and Generation (Figure 4). In contrast, technical tasks, such as coding and reasoning, constitute a smaller portion of the dataset, at 12.4% and 4.0%, respectively. This task distribution indicates that model comparisons arise in a natural conversational setting. On a 5-point difficulty scale, we found the majority of prompts to have a rating of 3 or lower, indicating a complexity solvable with knowledge generally acquired through a high school or un-

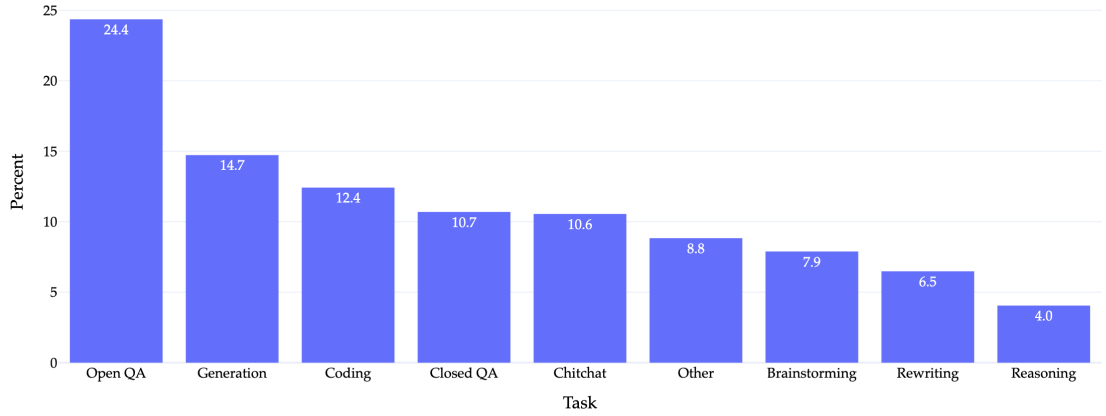

Figure 4: The distribution of task types for all historic Showdown prompts. Showdown prompts have a high proportion of conversational tasks like Open QA, Chitchat, and Generation. Coding (12.4%) is the most common technical task. Our task taxonomy is defined in Appendix B.1.

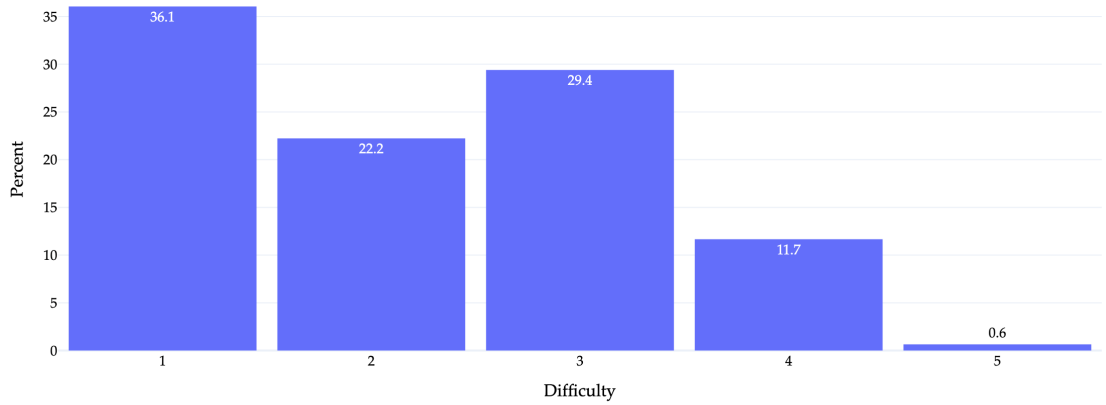

Figure 5: The distribution of prompt difficulty levels, classified with a 5-point scale (Appendix B.3).

dergraduate education (Figure 4). These task and difficulty distributions were characterized using LLM-based classifiers (Appendix B).

3.3 User distribution

To ensure data integrity, participation in Showdown is limited to verified users, and each individual is restricted to a single account to prevent the use of multiple identities. The resulting user base is globally distributed, highly educated, and predominantly young. Geographically, the user base is primarily concentrated in Asia (32.3%), North America (25.0%), and Europe (20.4%). Academically, the vast majority of users have a college degree (88.6%), and more than half hold an advanced or professional degree (56.6%); those with a high school education or less represent the smallest segment (5.1%). The user population also skews young, with a median age of 31. Users aged 18-34 constitute the majority (58.2%), with the 25-34 bracket forming the largest single cohort at 35.0%. English remains the dominant language for prompts at 64.8%.

An analysis of the number of battles per user reveals that Showdown serves a broad audience of casual users rather than a community of highly engaged enthusiasts, with approximately

Continent	%
Asia	32.3
North America	25.0
Europe	20.4
South America	11.6
Africa	8.6
Oceania	2.1

(a) Continent

Education	%
Graduate	21.3
Professional	35.3
College	32.0
High school and below	5.1
Other	6.4

(b) Education

Age group	%
18–24	23.2
25–34	35.0
35–44	22.0
45–54	12.4
55+	7.4

(c) Age group

Language	%
English	64.8
Spanish	9.2
Japanese	3.6
Portuguese	3.4
Turkish	2.3
French	1.9
Italian	1.8
Arabic	1.5
Korean	1.4
Indonesian	1.2
Other	9.0

(d) Language

Table 3: User percentages across demographics.

60% of participants engaging in three or fewer battles (Figure 6). Furthermore, influence is not concentrated among a few "power users," as users with fewer than 20 battles account for 95.3% of all battle activity.

3.4 Thinking versus non-thinking model variants

One surprising finding is that increased test-time compute does not necessarily lead to consistent improvements in Showdown ranking. We analyze head-to-head battles between the thinking and non-thinking variants of Claude models as a case study, as the two are the same language model with varying amounts of test-time compute [1].

We find that while extended thinking helps on higher-difficulty prompts, there is limited effect on less difficult ones (Figure 7). The use of additional test-time compute does not appear to give thinking variants a decisive advantage over non-thinking counterparts. Moreover, because thinking variants generally respond more slowly, their relative ranking declines once style control is applied.

3.5 Formatting issues

We observed that some model APIs (e.g. o3, o4-mini) do not consistently format their responses in Markdown. Instead of prompting models to adopt Markdown formatting where appropriate, we used their raw API output in Showdown battles. These models tended to rank lower, consistent with a strong user preference for well-formatted responses. Our use of style control mitigates the effect of this preference, leading to measurable increases in scores for models with low Markdown usage (Table 2, Figure 8). However, we recognize this as a potential confounder and plan further analysis to quantify and mitigate its impact.

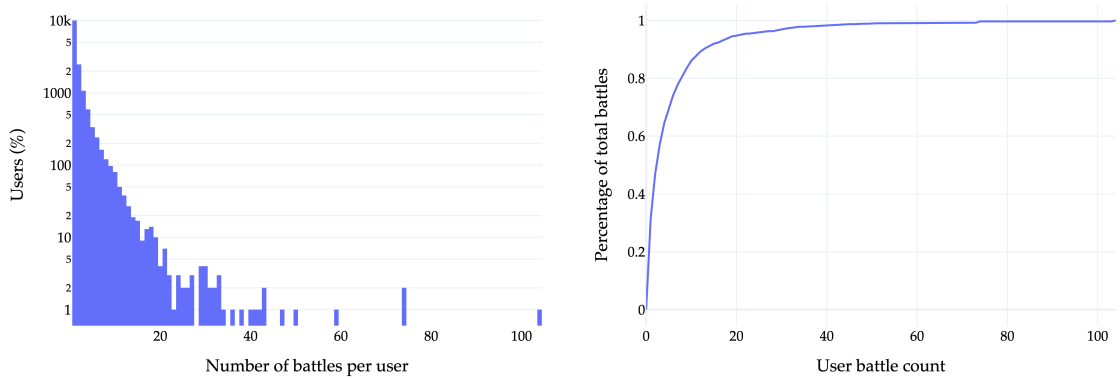

Figure 6: The user distribution, grouped by the number of total battles per user. Most users participate in a handful of battles, and 95.3% of overall battles are from users with fewer than 20 battles in their conversation history.

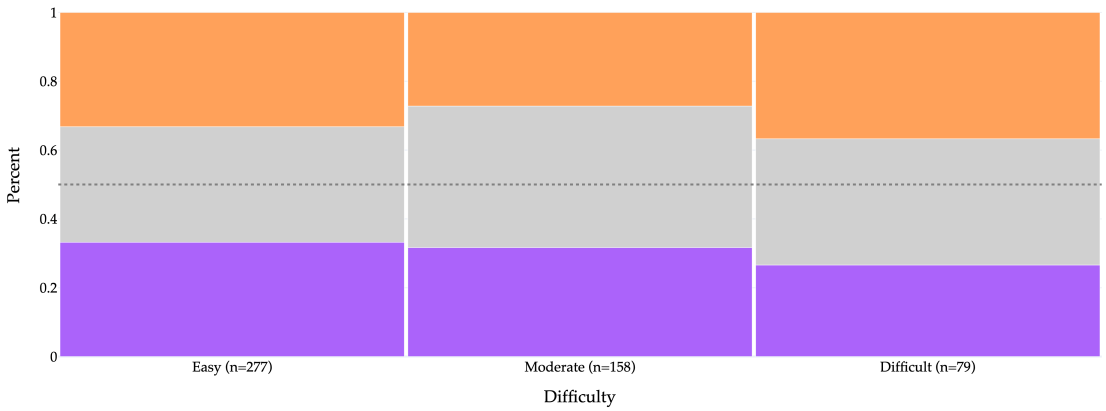

Figure 7: Battles between thinking and non-thinking variants of Claude models, grouped by prompt difficulty. Thinking variants outperform their non-thinking counterparts on higher-difficulty prompts, but these prompts are not sufficiently common for this to make a difference downstream. Difficulty ratings (5-point scale) are collapsed into three categories for visualization: easy (1–2), moderate (3), difficult (4–5).

4 Leaderboard Policies

To ensure the continued integrity of rankings derived from Showdown preference data, we make the following policies and commitments.

Model selection, inclusion, and deprecation To maintain a relevant and transparent leaderboard [39], we adhere to the following principles for model lifecycle management.

1. Models may be removed from the leaderboard when they are (a) no longer publicly accessible or (b) superseded by newer versions within the same capability track for the same provider. For transparency, we will maintain a public list of deprecated models post-launch. This policy is to ensure the validity of the ranking algorithm.
2. Our public leaderboards exclusively feature publicly available models, and all models added to the leaderboard will have their scores reported for at least 30 days' time. Prior to removing any model from the leaderboard, we will post a notice of its forthcoming removal at least 30 days prior, and display a visible countdown of the number of days remaining prior to removal on the leaderboard. This ensures that there are never "silent removals" of models from the leaderboard.

Ranking integrity To protect the integrity of our rankings from manipulation and overfitting, we enforce the following data confidentiality and access restrictions.

1. We will not provide access to human preference data used in the calculation of these rankings to outside parties.
2. In addition, we will not provide access to data which comes from a substantially similar distribution as the data used in the calculation of these rankings, if that data was gathered within the past 60 days, to outside parties. This ensures there is always a lag between the data distribution used to generate downstream rankings and the data we provide to outside parties.

Safeguarding user privacy We implement strict measures to protect user privacy at every stage of our analysis. Our approach follows established best practices in privacy-preserving computational social science and aligns with precedents from recent research on large language model usage [7, 37, 17, 19]. Specifically, we adopt two key safeguards:

1. **Automated PII-Scrubbing and De-Identification.** All user messages are first processed through automated classifiers designed to remove or mask personally identifiable information (PII), including names, contact information, and other sensitive attributes. No researcher manually inspects raw user messages. All downstream analyses are conducted only on the de-identified and PII-scrubbed text outputs from these classifiers.
2. **Aggregation Thresholds for Privacy Preservation.** To further protect individual privacy, we enforce strict aggregation thresholds. All analyses conducted or reported are exclusively on segments with at least 100 distinct users. This prevents the identification or re-identification of individual users and ensures that no user-level data is exposed or analyzed.

5 Related Work

LLM benchmarks As LLMs have become more powerful, researchers have proposed a wide spectrum of static benchmarks spanning commonsense QA [45, 4], reasoning and mathematics [20, 11, 15, 21], coding [8, 2, 24], safety and alignment [26, 18, 36], multimodal evaluation [44,

32, 31, 25, 29], and tool-augmented or agentic settings [38, 28]. SEAL Showdown complements these efforts by focusing on natural, up-to-date, and dynamic real-world use cases, capturing human preferences in situ rather than relying on fixed test sets.

Crowdsourced leaderboards Crowdsourced leaderboards have become increasingly popular for evaluating LLMs, with platforms such as Chatbot Arena [10] embedding blind comparisons into conversational settings and ranking models with Bradley–Terry or Elo-style methods. While these platforms provide scalability and ecological validity, they are vulnerable to gaming and manipulation, as highlighted by critiques on leaderboard illusions and vote-rigging risks [39, 33]. The SEAL Showdown is also a crowdsourced leaderboard and could be subject to manipulation. We are committed to protecting the integrity of rankings through our strict leaderboard policies and robust ranking methods.

Human preference datasets A parallel line of work has released large-scale human preference datasets to train reward models. Examples include early efforts such as WebGPT [34] and Anthropic's HH-RLHF [3], and more recent collections , UltraFeedback [13], PKU-SafeRLHF [14], and HelpSteer [43, 42]. These datasets demonstrate the centrality of human feedback for shaping model behavior. The SEAL Showdown also uses human preference for ranking model performance, but does so in situ, reflecting real-world use cases and preferences.

Stylistic preferences LLM-Judges and reward models consistently show preferences for stylistic features such as verbosity [22, 16, 27, 6] and formatting [30], independent of actual response quality. These biases are amplified by LLMs trained with RLHF, leading some models to overproduce longer or richly formatted outputs. To address this, several methods attempt to disentangle substance from style, including regression-based controls [16, 22, 41], preference-conditioned models [6], and debiasing reward models [23]. Our work follows this line by explicitly modeling stylistic features, ensuring that rankings reflect underlying capabilities rather than superficial presentation.

6 Conclusion

We present SEAL Showdown, a live leaderboard measuring human preferences for LLMs in a natural chat setting. Unlike traditional benchmarks, our platform collects preference data from *in situ* model comparisons. To produce robust rankings, we apply the Bradley–Terry model augmented with style controls, which mitigate biases from factors such as response length, Markdown formatting, and loading time. Our analysis indicates that the resulting dataset closely reflects real-world usage: prompts are conversational, diverse in topic and difficulty, and the user base is primarily composed of casual users, with no evidence of strong skew toward enthusiasts.

This report highlights two key findings. First, Showdown users exhibit a clear preference for responses that are longer and more heavily formatted. Second, models with extended thinking capabilities do not consistently outperform their non-thinking counterparts, indicating that additional compute offers limited benefit for most everyday tasks. We believe Scale Showdown complements the existing landscape of LLM evaluations by providing an accurate, fine-grained, and transparent view of human preferences in real-world settings.

7 Authorship and acknowledgements

Please cite this work as ”Scale AI (2025)”.

Contributors

Research
David Lee
Lifeng Jin
Zihao Wang
Jaehwan Jeong
Zifan Wang
Summer Yue
Bing Liu

Engineering
Charles Liu
Victor Shen
Jonny Liu
David Zhang
Immanuel Huang
Shivam Verma
Aakash Sabharwal

Product
Daniel Berrios
Janie Gu
Kai Yang
Kevin Tien
Farzad Eskafi

Design
Tim Bauer
Chelsea Tang
Luis Esquivel

Communications
Joe Osborne
George Quraishi
Caton Lu

Legal
Michael Le

References

[1] Anthropic. "Claude's extended thinking". In: (Feb. 2025). Accessed: 2025-09-04. URL: https://www.anthropic.com/news/visible-extended-thinking.

[2] Jacob Austin et al. "Program synthesis with large language models". In: *arXiv preprint arXiv:2108.07732* (2021).

[3] Yuntao Bai et al. "Training a helpful and harmless assistant with reinforcement learning from human feedback". In: *arXiv preprint arXiv:2204.05862* (2022).

[4] Yonatan Bisk et al. "Piqa: Reasoning about physical commonsense in natural language". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 05. 2020, pp. 7432–7439.

[5] Ralph Allan Bradley and Milton E. Terry. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". In: *Biometrika* 39.3/4 (1952), pp. 324–345. ISSN: 00063444, 14643510. URL: http://www.jstor.org/stable/2334029 (visited on 08/22/2025).

[6] Jianfeng Cai et al. "Disentangling length bias in preference learning via response-conditioned modeling". In: *arXiv preprint arXiv:2502.00814* (2025).

[7] Aaron Chatterji et al. *How People Use ChatGPT*. Working Paper 34255. National Bureau of Economic Research, Sept. 2025. DOI: 10.3386/w34255. URL: http://www.nber.org/papers/w34255.

[8] Mark Chen et al. "Evaluating large language models trained on code". In: *arXiv preprint arXiv:2107.03374* (2021).

[9] Wei-Lin Chiang et al. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. 2024. arXiv: 2403.04132 [cs.AI]. URL: https://arxiv.org/abs/2403.04132.

[10] Wei-Lin Chiang et al. "Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings". In: *ICLR*. 2024.

[11] Karl Cobbe et al. "Training verifiers to solve math word problems". In: *arXiv preprint arXiv:2110.14168* (2021).

[12] Rémi Coulom. "Computing "elo ratings" of move patterns in the game of go". In: *ICGA journal* 30.4 (2007), pp. 198–208.

[13] Ganqu Cui et al. "Ultrafeedback: Boosting language models with high-quality feedback". In: (2023).

[14] Josef Dai et al. "Safe rlhf: Safe reinforcement learning from human feedback". In: *arXiv preprint arXiv:2310.12773* (2023).

[15] Dheeru Dua et al. "DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs". In: *arXiv preprint arXiv:1903.00161* (2019).

[16] Yann Dubois et al. "Length-controlled alpacaeval: A simple way to debias automatic evaluators". In: *arXiv preprint arXiv:2404.04475* (2024).

[17] Tyna Eloundou et al. "First-person fairness in chatbots". In: *arXiv preprint arXiv:2410.19803* (2024).

[18] Samuel Gehman et al. "Realtoxicityprompts: Evaluating neural toxic degeneration in language models". In: *arXiv preprint arXiv:2009.11462* (2020).

[19] Kunal Handa et al. "Which economic tasks are performed with ai? evidence from millions of claude conversations". In: *arXiv preprint arXiv:2503.04761* (2025).

[20] Dan Hendrycks et al. "Measuring Massive Multitask Language Understanding". In: *ICLR*. 2021.

[21] Dan Hendrycks et al. "Measuring mathematical problem solving with the math dataset". In: *arXiv preprint arXiv:2103.03874* (2021).

[22] Zhengyu Hu et al. "Explaining length bias in llm-based preference evaluations". In: *arXiv preprint arXiv:2407.01085* (2024).

[23] Zeyu Huang et al. "Post-hoc reward calibration: A case study on length bias". In: *arXiv preprint arXiv:2409.17407* (2024).

[24] Carlos E Jimenez et al. "Swe-bench: Can language models resolve real-world github issues?" In: *arXiv preprint arXiv:2310.06770* (2023).
[25] Bohao Li et al. "Seed-bench: Benchmarking multimodal llms with generative comprehension". In: *arXiv preprint arXiv:2307.16125* (2023).
[26] Stephanie Lin, Jacob Hilton, and Owain Evans. "Truthfulqa: Measuring how models mimic human falsehoods". In: *arXiv preprint arXiv:2109.07958* (2021).
[27] Wei Liu et al. "Length desensitization in direct preference optimization". In: *arXiv preprint arXiv:2409.06411* (2024).
[28] Xiao Liu et al. "Agentbench: Evaluating llms as agents". In: *arXiv preprint arXiv:2308.03688* (2023).
[29] Yuan Liu et al. "Mmbench: Is your multi-modal model an all-around player?" In: *European conference on computer vision*. Springer. 2024, pp. 216–233.
[30] Do Xuan Long et al. "Llms are biased towards output formats! systematically evaluating and mitigating output format bias of llms". In: *arXiv preprint arXiv:2408.08656* (2024).
[31] Pan Lu et al. "Learn to explain: Multimodal reasoning via thought chains for science question answering". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 2507–2521.
[32] Pan Lu et al. "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts". In: *arXiv preprint arXiv:2310.02255* (2023).
[33] Rui Min et al. "Improving your model ranking on chatbot arena by vote rigging". In: *arXiv preprint arXiv:2501.17858* (2025).
[34] Reiichiro Nakano et al. "Webgpt: Browser-assisted question-answering with human feedback". In: *arXiv preprint arXiv:2112.09332* (2021).
[35] Ryan Park et al. *Disentangling Length from Quality in Direct Preference Optimization*. 2024. arXiv: 2403.19159 [cs.CL]. URL: https://arxiv.org/abs/2403.19159.
[36] Alicia Parrish et al. "BBQ: A hand-built bias benchmark for question answering". In: *arXiv preprint arXiv:2110.08193* (2021).
[37] Jason Phang et al. "Investigating affective use and emotional well-being on ChatGPT". In: *arXiv preprint arXiv:2504.03888* (2025).
[38] Yongliang Shen et al. "Taskbench: Benchmarking large language models for task automation". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 4540–4574.
[39] Shivalika Singh et al. "The leaderboard illusion". In: *arXiv preprint arXiv:2504.20879* (2025).
[40] Prasann Singhal et al. *A Long Way to Go: Investigating Length Correlations in RLHF*. 2024. arXiv: 2310.03716 [cs.CL]. URL: https://arxiv.org/abs/2310.03716.
[41] Wei-Lin Chiang* Tianle Li* Anastasios Angelopoulos*. *Does Style Matter? Disentangling style and substance in Chatbot Arena*. Aug. 2024. URL: https://blog.lmarena.ai/blog/2024/style-control/.
[42] Zhilin Wang et al. "Helpsteer 2: Open-source dataset for training top-performing reward models". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 1474–1501.
[43] Zhilin Wang et al. "Helpsteer: Multi-attribute helpfulness dataset for steerlm". In: *arXiv preprint arXiv:2311.09528* (2023).
[44] Xiang Yue et al. "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9556–9567.
[45] Rowan Zellers et al. "Hellaswag: Can a machine really finish your sentence?" In: *arXiv preprint arXiv:1905.07830* (2019).

Appendix A Effect of style control

A.1 Score difference by feature

In Figure 8, we show the effect of applying style control with individual style features on the score assigned to each model. This reveals more finer-grained effects of style control, showing for instance that o3 and o4-mini both benefit from the Markdown feature–a result that is masked when observing the difference in rankings alone (Table 2). Recall that Llama 4 Maverick is used as an anchor model; as such, its rank is fixed.

Figure 8: The effect of different style control features on each model's leaderboard score. Each column represents the difference between the vanilla score (i.e., without style control) and the score after applying style control with a single feature, with the rightmost column showing the score difference after applying all style features. Models are ordered by their vanilla score.

A.2 Response lengths

In Figure 9, we show the mean token count by model. We observe that Gemini models have the most verbose responses, which leads to their being the most penalized by the length style control (Figure 8).

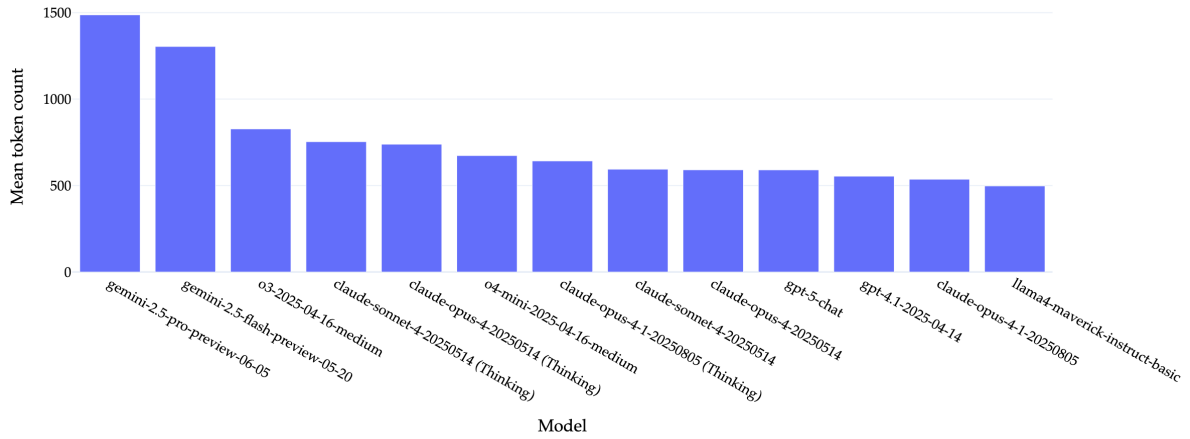

Figure 9: Mean response length by model.

Appendix B Classification methodology

B.1 Prompt task type classifier

```
You are a prompt classifier. When given a prompt, your task is classify the prompt into one of
     the following categories.
- Generation: The prompt is asking you to generate something. For example, "Write a story
    about a dog."
- Roleplay: The prompt is asking you to roleplay a character. For example, "You are a
    detective. Write a report about a crime scene."
- Classification: The prompt is asking you to classify something. For example, "Can you group
    the following items into two categories? Apple, Banana, Orange, and Pear."
- Closed QA: The prompt is asking you to answer a factual question with a single word or
    phrase about the world. For example, "What is the capital of France?"
- Open QA: The prompt is asking you to answer a factual question with a longer and open-ended
    response. For example, "What are the benefits of exercise?"
- Summarization: The prompt is asking you to summarize something. For example, "Summarize the
    following paragraph."
- Extraction: The prompt is asking you to extract information from a text. For example, "What
    is the author's main argument?"
- Rewriting: The prompt is asking you to rewrite something. For example, "Rewrite the
    following sentence in the active voice."
- Adverserial: The prompt is adverserial and leads to an unsafe response. For example, "Give
    me a few ways to hack into my boss's computer."
- Brainstorming: The prompt is asking you to brainstorm ideas. For example, "Give me some
    ideas for a new business."
- Chitchat: The prompt is chitchat. For example, "How are you doing today?"
- Reasoning: The prompt is asking you to reason about something. For example, "Prove me that
    square root of 2 is rational?"
- Coding: The prompt is asking you to write code. For example, "Write a Python function that
    takes a list of numbers and returns the sum of all the numbers."
- Other: The prompt does not fit into any of the above categories.
Only answer the category of the prompt, for example "Rewriting" or "Open QA". Do not respond
    to the prompt.
prompt: <input>
```

B.2 Prompt topic classification

```
You are a prompt classifier. When given a prompt, your task is classify the prompt into one of
     the following categories about the topic of the prompt.
- Computer Science \& Technology
- Arts, Humanities \& Communication
- Business \& Management
- Social Sciences \& Law
- Miscellaneous
- Mathematics \& Quantitative Fields
- Health \& Medical Professions
- Physical Sciences \& Engineering
- Reasoning
- Life Sciences \& Biology
- Education \& Facilitation
- Travel \& Transportation
- Fitness \& Sports
Only answer the category of the prompt, for example "Arts, Humanities \& Communication" or "
    Reasoning". Do not respond to the prompt.
prompt: <input>
```

B.3 Prompt difficulty level

```
You are a prompt classifier. When given a prompt, your task is classify the prompt into one of
     the following categories about the level of difficulty of the prompt.
Read the definition of each category carefully and choose the one that best fits the prompt.
- Ultra Low Domain Difficulty: The prompt is extremely easy to understand and answer. Most
    people with Primary School education can answer the prompt correctly. They are usually
    about general and commonsense topics, such as "What is the capital of France?"
- Low Domain Difficulty: The prompt is easy to understand and answer. Most people with High
    School education can answer the prompt correctly. They include world knowledge, basic
```

```
    domain knowledge, and basic reasoning. For example, "how can we interpret the reaction
    between a metal and an acid?"
- Medium Domain Difficulty: The prompt is moderately difficult to understand and answer.
    People with college education can answer the prompt correctly. They include more college
    level world knowledge, specialized domain knowledge, and college level math and reasoning.
     They also include professional requirements commonly seen in work. For example, "give me
    a list of the top 10 companies in the world. run a SWOT analysis on the top 5 companies."
- High Domain Difficulty: The prompt is difficult to understand and answer. People with
    advanced degrees can answer the prompt correctly. The prompt requires deep understanding
    of the domain, advanced reasoning, and deep experience with domain specific knowledge. For
     example, "compare and contrast the sparse attention mechanism and the attention mechanism
     in the Transformer model."
- Ultra High Domain Difficulty: The prompt is extremely difficult to understand and answer.
    They include Olympiad level questions, advanced math problems, and frontier domain
    specific knowledge that only a few people in the world can reliably answer. For example, "
    Determine all real numbers  such that, for every positive integer n, the  integer +  2 +
     · · · +  n is a multiple of n. (Note that  z denotes the greatest integer less than or
    equal to z. For example,  −  = −4 and  2 =  2.9 = 2.)"
Only answer the category of the prompt, for example "Low Difficulty" or "High Difficulty". Do
    not respond to the prompt.
prompt: <input>
```

B.4 User education level

We classified users using their resume information and the following criteria.

- High school and below: The user has completed high school or lower education. This may include middle school, elementary school, or no formal education.
- College: The user is currently enrolled in a college or university program. This may include an associate's degree, bachelor's degree, or any other undergraduate program.
- Professional: The user has completed a college or university program and is currently working in a professional field. This may include a bachelor's degree, master's degree, or any other professional certification.
- Graduate: The user has completed or is currently enrolled in a doctoral program or research program. This may include a PhD, MD, or any other advanced degree.
- Other: The user's education level does not fit into any of the above categories. This may include vocational training, online courses, or any other form of education that does not fall into the above categories.