VARCO Arena: A Tournament Approach to Reference-Free Benchmarking Large Language Models

Anonymous ACL submission

Abstract

Most existing benchmarking approaches for evaluating the output quality of large language models (LLMs) rely on comparing LLM responses to predefined references. Such methods, based on static datasets, quickly become outdated as LLM capabilities and use cases evolve. In this work, we introduce VARCO Arena—a novel, cost-effective, and robust benchmarking approach that leverages a single-elimination tournament structure to minimize the number of required comparisons while eliminating the need for static references or costly human annotations. We validate our approach through two experiments: (i) a simulation study that examines its robustness under various conditions, and (ii) an empirical evaluation using publicly available benchmark prompts. In both experiments, VARCO Arena consistently outperforms current LLM benchmarking practices, achieving stronger correlations with human-established Elo ratings. Our results demonstrate that VARCO Arena not only produces reliable LLM rankings but also provides a scalable, adaptable solution for qualitative evaluation across diverse, customized use cases. We release our demo and code at [URL placeholder].

1 Introduction

The versatility of Large Language Models (LLMs) stems from their generative capacity to address a wide array of tasks. The multi-faceted capability of LLMs enables flexible applications across numerous user scenarios (Ouyang et al., 2022; Köpf et al., 2024; Roziere et al., 2023). As versatility emerges as a core attribute of LLMs, the challenge of accurately gauging their skill becomes increasingly significant. In response to this challenge, numerous benchmarks evaluating LLM capabilities have emerged (Hendrycks et al., 2020; Srivastava et al., 2023; Zhong et al., 2024). Many LLM benchmarks employ formats amenable to automated scoring.

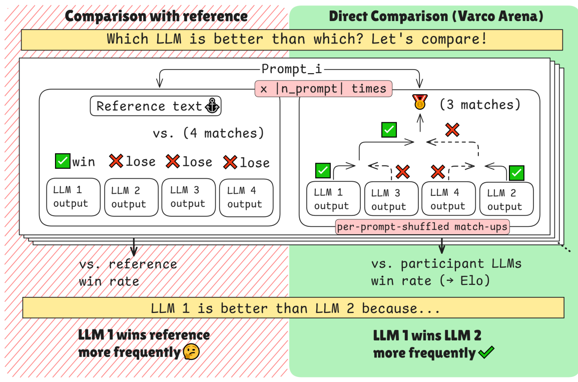

Figure 1: VARCO Arena directly compares LLM response pairs in single-elimination tournament rather than comparing references. In terms of deciding whether a certain LLM is better or worse compared to the other one, we suggest direct head-to-head comparison is more intuitive and results in better separability.

Examples include benchmarks for arithmetic problems (Gao et al., 2022; Patel et al., 2021), multiple-choice questions (Lin et al., 2022), and code execution (Austin et al., 2021; Chen et al., 2021). While these benchmarks are valuable, they primarily assess problem-solving abilities. The need for LLM benchmarks extends beyond this, as the primary value of LLMs lies in the versatility of their generative behavior. Researchers have primarily relied on pairwise comparisons to evaluate and rank the LLMs based on human preference annotations. A representative example of this approach is Chatbot Arena (Chiang et al., 2024), which computes Elo ratings based on a massive number of human votes. While Chatbot Arena benefited from reliable, open-ended prompts generated by a large user base, obtaining such a high volume of annotations from dedicated annotators remains challenging.

To address these challenges, several benchmark datasets for generation tasks have been developed and applied with an LLM judge (Zheng et al., 2024) to quantify LLM capabilities. Using these benchmarks and their reference responses, LLMs are

	No. Comp. (↓)	Judge	Eval. Type
Chatbot Arena	unknown	human	head-to-head
Current Practice	$n_{\text{model}} \cdot \lvert X \rvert$	LLM	reference-based
VARCO Arena	$(n_{\text{model}} - 1) \cdot \lvert X \rvert$	LLM	head-to-head

Table 1: Comparison between Current Practice and Varco Arena. $|X|$ and n_{model} represents size of benchmark dataset, and number of candidate LLMs to rank respectively. Human annotators are considered much more costly than LLM judge counterpart.

ranked via pairwise comparisons between their outputs and the reference responses provided by LLM judges. Representative examples include AlpacaEval (Li et al., 2023), Arena-Hard-Auto (Li et al., 2024), and MTBench (Zheng et al., 2023). There are two major advantages to relying on reference responses: (1) The number of comparisons required for ranking scales linearly with the number of LLMs. (2) The reference response establishes a quality standard for evaluating LLM outputs.

However, we argue that using reference responses to mediate comparisons between LLM outputs is suboptimal and that head-to-head comparisons yield more accurate assessments. In this context, we propose VARCO Arena as a novel evaluation framework that directly compares LLM responses while requiring fewer comparisons. Even with fewer comparisons and without relying on reference responses, VARCO Arena achieves a stronger correlation with the human-established Elo rankings from Chatbot Arena.

VARCO Arena is designed to conduct a single-elimination tournament for each prompt across all participating LLM responses and then compute Elo ratings (Elo and Sloan, 1978) based on the outcomes of these matches. By organizing tournament-style matches for multiple LLMs on each prompt, we obtain relative win rates between all possible model pairs with fewer comparisons than those required by reference-based methods. These win rates are subsequently interpreted as Elo ratings, enabling us to generate a comprehensive ranking of candidate LLMs.

We validate the effectiveness of VARCO Arena against current LLM benchmarking practices through two experiments. First, we conduct a simulation study (Section 4.2) to evaluate the reliability and robustness of our tournament approach under various controlled factors, including the number of participating LLMs, the number of test prompts, and the accuracy of the judge model. Second, we empirically demonstrate that running VARCO Arena with several judge LLMs on public benchmark (Section 4.3) prompts yields a stronger correlation with human-established Elo rankings from Chatbot Arena than does the reference-based approach.

These experiments confirm that our benchmarking approach outperforms current practices in terms of ranking reliability while requiring the lesser number of comparisons (Table 1). Notably, the simulation results—which emulate outcomes under various conditions—underscore that the superiority of our tournament approach is not coincidental but rather reflects the inherent advantages of direct pairwise comparisons over mediated comparisons using reference responses.

In this work, we make the following contributions:

1. We introduce VARCO Arena, a novel reference-free benchmarking method for LLMs that leverages a single-elimination tournament structure to perform head-to-head comparisons of LLM responses across test prompts. This approach eliminates the dependency on static references and enhances benchmarking flexibility.
2. Through extensive simulation studies and empirical evaluations, we demonstrate that VARCO Arena not only yields more reliable LLM rankings compared to traditional reference-based methods but also achieves this with significantly fewer comparisons, thereby reducing evaluation costs.
3. We release our demo and code to facilitate the further research and reproducibility; the code is available at [URL placeholder]

2 Preliminaries: Quantifying Generation Ability

Quantifying an LLM's generation ability is crucial, but it presents several challenges. The outcomes of comparisons between two LLMs are often probabilistic, influenced by various factors such as the provided test prompts and the inherently subjective nature of human preferences. A straightforward approach is to evaluate an LLM's performance across diverse test prompts, which approximates its real-world performance. Two popular measures for LLM generation ability are the win rate against reference responses and the Elo rating.

2.1 Win Rate in Comparison to the Reference

AlpacaEval (Li et al., 2023) and Arena-Hard-Auto (Li et al., 2024) are representative benchmark datasets that aim to quantify LLM response quality using reference responses. These benchmarks employ an LLM judge with a standardized prompt for automated evaluation of generation capability. Given the prompts, LLM responses, and reference responses, the judge LLM is tasked with determining whether the reference or the response is preferred. The LLM's win rate across the test prompts in the benchmark is used as a measure of its generation proficiency.

2.2 Elo Rating

The Elo rating system, introduced by Elo and Sloan, has since become a popular method for quantifying performance levels in competitive sports and, more recently, for evaluating the generative abilities of LLMs. The primary purpose of the Elo rating system is to represent a participant's skill level as a single scalar value, enabling the prediction of relative win rates between participants who have never directly competed. Assessing the superiority of one LLM over another in terms of generative capability shares several similarities with determining winners in competitive sports. Varying test prompts may yield different results, similar to how weather or other factors can influence the outcome of a sports match. While a superior team or LLM is not guaranteed to outperform its inferior counterpart in every instance, it tends to succeed more frequently. With Elo ratings, one can expect the relative chance of winning between a pair of players. The expected win rate of the player i against j is computed as follows:

$$\mathrm{P}(i > j) = \frac{1}{1 + 10^{(R_j - R_i)/400}} = \frac{1}{1 + 10^{\Delta_{\mathrm{ij}}/400}} \tag{1}$$

Computing accurate Elo ratings requires a sufficient number of matches to estimate relative win rates between pairs of participants. The resulting matrix of relative win rates is then used to estimate Elo ratings through logistic regression, similar to multivariate logistic regression with a sigmoid function. In Equation 1, $P(i > j)$ represents the expected win rate of participant i over j, R_i is the Elo rating of participant i, and Δ_{ij} is the Elo rating difference between participants j and i ($R_j - R_i$).

Chatbot Arena presents a leaderboard of LLMs, with Elo ratings computed from matches evaluated by a large user base. Users prompt a pair of LLMs and submit their judgments of which one responded better. Although Chatbot Arena relies on manual evaluation of matches, it offers an intuitive method for comparing LLMs using Elo ratings.

3 VARCO Arena

We propose VARCO Arena, a reference-free approach for benchmarking large language models (LLMs) via single-elimination tournaments. Instead of comparing model responses against fixed references, VARCO Arena directly compares outputs from different models, determining superiority through head-to-head matchups for each prompt in our benchmark dataset. Repeated tournaments across prompts yield reliable leaderboards that reflect the relative performance of each model.

We begin by motivating our approach over current reference-based evaluations (Section 3.1). Next, we detail how VARCO Arena performs tournaments and results in Elo ratings (Section 3.2 and Algorithm 1). Finally, we discuss the reason why aggregating the results from tournaments promises reliable ranking (Section 3.3).

3.1 Comparing to a Reference is not Always Helpful

Although reference texts are a standard way to evaluate and rank large language models (LLMs), they introduce potential failure modes. Beyond the fact that a single reference might not capture every dimension of correctness, relying solely on a reference can lead to unreliable rankings of LLMs.

Consider an ideal scenario with a judge capable of perfectly distinguishing the quality of any two outputs. If we choose to compare LLM responses directly to rank them using Elo ratings (Equation 1), all head-to-head comparisons are utilized. In contrast, reference-based evaluation for differentiating LLMs can exhibit failure modes, as shown in Equation 2.

$$\begin{array}{c}\mathrm{M}_1(X_i)\\ \text{vs.} \rightarrow\\ \mathrm{M}_2(X_i)\end{array} \begin{cases} \mathrm{M}_1(X_i) > Y_i > \mathrm{M}_2(X_i) & \text{(helpful)} \\ \mathrm{M}_1(X_i) < Y_i < \mathrm{M}_2(X_i) & \text{(helpful)} \\ \mathrm{M}_1(X_i),\ \mathrm{M}_2(X_i) > Y_i & \text{(unhelpful)} \\ \mathrm{M}_1(X_i),\ \mathrm{M}_2(X_i) < Y_i & \text{(unhelpful)} \end{cases} \tag{2}$$

When the reference output (Y_i) for a prompt (X_i) successfully disambiguates the pair of LLM responses $M_1(X_i)$ and $M_2(X_i)$ (as in the first and second cases), comparison to the reference is effective for benchmarking. Otherwise, these comparisons do not help differentiate LLM performance. Consequently, the reference-based approach provides less information for ranking when multiple responses are either both correct or both incorrect relative to the reference.

3.2 Tournaments of LLMs over multiple prompts to Elo Ratings

Algorithm 1 Tournaments of LLMs over prompts

Require: prompts $X = \{x_1,\ x_2,\ ...,\ x_i\}$, LLMs $M = \{m_1,\ m_2, ...,\ m_j\}$, outputs $O_{i,j} = m_j(x_i)$

Ensure: Ranked LLMs with Elo ratings

function `Match`(m_1, m_2, x)
  return m_1 if `IsBetter`($O_{x,1}, O_{x,2}$)
  else m_2
end function
function `SingleElim`(M, x, res)
  if $|M| = 2$ **then**
    res.append(`Match`(M[0], M[1], x))
    return res[-1]
  end if
  mid $\leftarrow \lfloor |M|/2 \rfloor$
  left $\leftarrow$ `SingleElim`(M[:mid], x, res)
  right $\leftarrow$ `SingleElim`(M[mid:], x, res)
  return `SingleElim`(left + right, x, res)
end function
function `Tournaments2Ranks`(X, M)
  res $\leftarrow$ []
  for $x_i \in X$ **do**
    `SingleElim`(`Shuffled`(M), x_i, res)
  end for
  return `ComputeElo`(res)
end function

Figure 1 and Algorithm 1 illustrate how VARCO Arena benchmarks LLMs via a tournament approach. Here, $|X|$ denotes the number of prompts in the benchmark dataset. Each execution of VARCO Arena runs a tournament among participant LLMs for every prompt in the dataset.

The use of tournament structures for LLM benchmarking offers both benefits and challenges. A major advantage of a single-elimination tournament is efficiency. As shown in Table 1, the number of matches scales linearly with the number of participants and even lower compared to using references. However, single elimination tournament only identifies a champion, leaving the relative ordering of other participants unclear.

To retain tournament's efficiency while obtaining a fine-grained ranking, we propose aggregating tournament results over multiple prompts with randomized initial match-ups for each prompt. Performing multiple tournaments with random initialization offers several benefits:

1. It resolves ties among non-champion participants from previous tournaments.
2. It mitigates the impact of unfavorable match-ups in any single tournament.
3. Aggregating match results allows for precise win rate estimation via Elo ratings, resulting in a well-aligned overall ranking.
4. More matches are allocated to high-performing participants while ensuring every participant is evaluated at least once per prompt.

In Section 3.3, we further explain how aggregating multiple tournaments could yield an reliable ranking of LLMs. We also provide an analysis of the number of matches each LLM faces, offering a comprehensive view of the method's efficiency and effectiveness.

3.3 Why Aggregating Multiple Tournaments Yields Reliable Ranks?

Our approach aggregates match outcomes from multiple tournaments to approximate the complete set of pairwise comparisons—akin to the comparisons made in merge sort. In a single-elimination tournament, every participant advances based solely on match outcomes, a process that mirrors the merging steps in merge sort. Notably, a single-elimination tournament executes only the comparisons strictly necessary for determining the winner, omitting many comparisons that merge sort would perform further.

We posit that the missing pairwise match-ups in any one tournament can be recovered by aggregating tournaments conducted over different prompts. This hypothesis relies on the assumption—central to our use of the Elo model—that match outcomes are independent of the prompt. Consequently, matches across different prompts are considered equivalent.

Considering only the initial comparisons, which are randomly sampled, the aggregate number of comparisons is $|X| \cdot n_{\text{model}}/2$. Since n_{model} is typically on the order of tens and $|X|$ comprises at least hundreds, this number exceeds the total possible match-ups, $\binom{n_{\text{model}}}{2}$[1]. Furthermore, as shown in Table 1, the remaining matches—totaling $|X| \cdot (n_{\text{model}} - 1)$—either contribute additional merge sort structures or enhance the accuracy of estimating relative win rates among LLM participants.

Moreover, considering that each unique pair meets in at least $|X|/(2(n_{\text{model}} - 1))$ matches[2] across the benchmark, this frequency is sufficient for a fair estimation of their relative win rates.

In summary, aggregating tournaments not only reconstructs the full set of pairwise comparisons for a merge sort but also ensures that each pair of models faces one another often enough to yield accurate win rate estimations, leading to reliable Elo ratings. Based on these considerations, we propose that conducting tournaments over a benchmark prompts will yield a reliable ranking of LLMs.

4 Experiments

We run two experiments with different settings for comparing VARCO Arena and current practice of reference-based benchmarking. In the first experiment (Section 4.2) we conduct a simulation study to control various factors that affect LLM benchmarking. This simulation tests our foundational propositions for VARCO Arena design (mentioned in Section 3.1 and 3.3) under a more controlled, simplified environment immune to noisy factors such as potential biases of LLM judges (Park et al., 2024).

The other is empirical experiments (Section 4.3). We use `gpt-4o` and `-mini` as well as other several popular models such as `Claude3.5`, `Qwen2.5`, `Llama3.1`, and `Gemma2` to validate the effectiveness of our tournament approach against current LLM benchmarking practices. By presenting both simulation and empirical results, we aim to demonstrate the effectiveness of our tournament approach. In Section 4.1, we first describe the common experimental settings before getting into specific details of each experiments in the following subsections.

[1]Note that merge sort produces a set of match-ups with no duplicates.

[2]A model participates in between $|X|$ and $\lceil \log_2 n_{\text{model}} \rceil$ matches per tournament. Dividing this by the number of possible match-ups per model, $2(n_{\text{model}} - 1)$, yields the expression above.

4.1 Chatbot Arena Leaderboard Ratings as Ground-Truth LLM Rankings

We compare the results of each benchmarking approach against the rankings from the Chatbot Arena leaderboard. Chatbot Arena is widely regarded as one of the most reliable leaderboards due to its extensive collection of human preference annotations. Given the large number of votes and the diverse set of prompts used for model comparisons, the resulting rankings are considered sufficiently accurate to serve as ground truth.

4.2 Experiment 1: Simulation Study

We designed a simple simulation to emulate a probabilistic model of LLM matches, adhering to the Elo rating system in a controlled environment. In line with the Elo model's assumptions, our judge is configured to follow Equation 3 exactly. The judge stochastically determines the winner of each LLM match solely based on the Elo rating difference (Δ_{ij}) and the judge's accuracy (P_{judge}). As described in Equation 3, the outcome of a single match is sampled according to $P_{\text{predict}}(i > j)$, which is computed as the product of the judge's accuracy and the likelihood of model i beating model j based on the Elo gap (P_{gt}).

$$\begin{aligned} P_{\text{predict}}(i > j) &= P_{\text{judge}} \times P_{\text{gt}}(i > j) \\ &= P_{\text{judge}} \times \frac{1}{1 + 10^{\Delta_{\text{ij}}/400}} \end{aligned} \tag{3}$$

The details of our simulation settings are as follows:

Ground-truth Elo ratings (initial parameter of the simulation, and at the same time, the gold ranking to reproduce): We extracted Elo ratings from the English category of Chatbot Arena as of June 23. This Elo ratings are the estimates from massive user-submitted judgments (approximately 60% of the total submissions to the platform). For the simulation, any set of Elo ratings could be used, but we opted for real values computed from human preferences.

Judge Accuracy (P_{judge}): In practice, a judge's accuracy is an adaptive value that depends on factors such as the specific prompt-response pair and the manner in which LLM judges are prompted. In this simulation study, we control this parameter, varying it from 0.6 to 0.9.

Number of Participant LLMs (n_{model}) and Benchmark Dataset Size ($|X|$): We varied these factors to assess the robustness of both the tournament and reference-based approaches under diverse conditions. This allowed us to explore how reliability changes with the number of participants in both data-poor and data-rich environments.

Simulation Procedure:

1. Select the participant LLMs and obtain their Elo ratings for the simulation.
2. Compute the expected relative win rates (P_{gt}, see Equation 3) from the participants' Elo ratings.
3. Sample match outcomes for each LLM pair according to the benchmarking approach. The winner of each match is sampled from P_{pred} (Equation 3), which depends solely on the Elo gap (Δ_{ij}) and the judge's accuracy (P_{judge}).
4. Repeat step 3 for the designated number of test prompts ($|X|$).
5. Calculate scores for ranking:
   (a) For the reference-based approach (current practice), use the win rate against the reference model (`gpt-4-1106-preview`, which has an Elo rating of 1233).
   (b) For the tournament approach (VARCO Arena), compute Elo ratings from all match outcomes.
6. Rank the models based on these scores.
7. Compute the Spearman correlation between the simulated rankings (from step 6) and the ground-truth rankings (from step 1).

We perform 50 trials for each simulation configuration to mitigate randomness from tournament brackets and sampling.

4.3 Experiment 2: Assessing VARCO Arena Using Various LLM Judges

To empirically validate our proposal, we evaluated the reliability of both VARCO Arena and reference-based approach over the top 20 models from the Chatbot Arena leaderboards. This experiment employs actual prompt inputs and LLM outputs, distinguishing it from the earlier simulation study.

4.3.1 Dataset: Test Prompts and LLM Responses Used

Testing the benchmarking approaches requires: (1) test prompts and (2) the corresponding responses from LLMs. For the benchmark dataset, we selected Arena-Hard-Auto (Li et al., 2024). The prompts in Arena-Hard-Auto were carefully curated from Chatbot Arena user queries. This dataset consists of 500 prompts—two instances for each of 250 subtopics. Although AlpacaEval (Li et al., 2023), which comprises 800 prompt-reference pairs, could serve as a viable testbed, we opted for Arena-Hard-Auto because its design aligns more closely with Chatbot Arena. Arena-Hard-Auto uses responses from `gpt-4-0314` as the reference outputs. For ranking, we utilized the reserved outputs of the top 21 models from the Arena-Hard-Auto Browser.[3]

4.3.2 Participant LLMs

For ranking, we selected 20 LLMs from the top of the ChatBot Arena leaderboard in the *hard prompts* category, as these models most closely align with Arena-Hard-Auto.

4.3.3 LLM Judges

We used several aligned LLMs as judges for testing both benchmarking approaches. LLMs of our choice are `gpt-4o` family of models (OpenAI et al., 2024), `Claude3.5`, and a selection of open-weight models: `Qwen2.5` (Qwen et al., 2025), `Llama3.1` (Grattafiori et al., 2024), and `Gemma2` (Team et al., 2024). For pairwise comparisons of responses, we employed the judging prompt suggested in LLMBar (Zeng et al., 2024) (See Appendix A.6.2). The same judge prompt was applied consistently across both the tournament and reference-based approaches. To mitigate position bias (Wu and Aji, 2023), the order of model responses was alternated during evaluation. Further details on the LLM-as-a-judge configuration are provided in Appendix A.6.

The two experimental settings are summarized as follows:

Experiment 1 (Simulation Study): This experiment uses the ground truth Elo ratings of the models to initialize the simulation. We vary control parameters for the benchmarking approaches—including the judge's accuracy (P_{judge}), the number of test prompts used ($|X|$), and the number of participant LLMs (n_{model})—to determine which benchmarking approach more accurately reproduces the participants' ranking. For each configuration, we conduct 50 trials of experiments.

[3] Extracted from the 2024 Jul 6 commit (`fd42026`).

Experiment 2 (Empirical Runs): This experiment assesses the two benchmarking approaches using empirical runs with various LLM judges. We select the top 20 LLMs from Chatbot Arena and used their reserved outputs on Arena-Hard-Auto test prompts. For both the tournament and reference-based approaches, we employ the Spearman correlation coefficient to measure how well the results align with the ground truth leaderboard rankings. In our empirical study, we conduct 500 trials for each experimental setting.

5 Results and Discussion

We assess the reliability and robustness of VARCO Arena as a means for LLM benchmarking, comparing it against the current reference-based approach. Our results from both simulation study and empirical runs indicate that the tournament approach of VARCO Arena yields rankings that align more closely with the ground-truth Elo leaderboards. We present our findings using whisker plots and tables in the following sections.

5.1 Experiment 1: Simulation Study Results

Figure 2 illustrates noticeable differences in Spearman correlation, indicating that the tournament approach is more reliable than the reference-based method. The consistent performance gap across various conditions—namely, the number of participants, the number of test prompts, and judge accuracy (n_{model}, $|X|$, and P_{judge})—demonstrates the robustness of the tournament approach. Although the simulation simplifies real-world complexity, a similar performance gap was observed in the empirical findings (Experiment 2, Figure 3). This consistency suggests that the robust performance of VARCO Arena is not coincidental or limited to a specific empirical setting of ours.

5.2 Experiment 2: Empirical Results

As hinted in the previous section, the empirical results in Figure 3 show that VARCO Arena consistently outperforms the reference-based approach. Although the performance gaps are less pronounced than in the simulation, the same trend persists. In Table 2, we report the median values for VARCO Arena and the reference-based approach using the `gpt-4o` family of judges while varying the number of test prompts ($|X|$). These results consistently demonstrate that VARCO Arena outperforms the reference-based method. Note that VARCO Arena shows similar or superior reliability even in extreme data-poor benchmark condition ($|X| = 50$).

Table 3 presents the outcomes when using other LLMs as judges, with a fixed number of prompts ($|X| = 500$). The results for `Claude3.5-sonnet`, `Llama3.1-8b`, and `Qwen2.5-7b` follow a similar trend. However, smaller models (`Gemma2-2b` and `Qwen2.5-0.5b`) appears to be less reliable for benchmarking. Hence, we recommend using evaluation-specialized judge LLMs or, at least, generative judge models with around 7B parameters regardless of using VARCO Arena or considering reference-based approach.

Spearman corr. (↑)	$\|X\|$ = 50	100	250	475	500
comp. to ref. (4o)	0.895	0.935	**0.963**	0.966	0.964
tournament (4o)	**0.905**	**0.940**	0.960	**0.970**	**0.970**
comp. to ref. (4o-mini)	0.895	0.908	0.917	0.916	0.912
tournament (4o-mini)	**0.901**	**0.919**	**0.931**	**0.933**	**0.933**

Table 2: Spearman correlation (↑) varying over size of the benchmark set ($|X|$) for each benchmarking approach. Comp. to ref. refers to reference-based approach.

$\|X\|$ = 500	claude3.5 sonnet	llama3.1 8b-it	qwen2.5 7b-it	qwen2.5 0.5b-it	gemma2 2b-it
comp. to ref.	0.924	0.820	0.756	0.089	**0.592**
tournament	**0.930**	**0.850**	**0.811**	-0.124	0.552

Table 3: Spearman correlation (↑) result using other LLMs as a judge. Comp. to ref. refers to reference-based approach.

5.3 Incorporating a New LLM into an Existing Leaderboard

While our main focus has been on ranking multiple LLMs at once, it is also useful to consider the common scenario of adding a single new model to an existing leaderboard, which is also frequent use-case for leaderboards. We explored two approaches: (1) a *binary search*-like placement method, and (2) using the top-performing model response as a reference. Our findings indicate that the latter approach is more reliable (Table 4). Further details and discussions are provided in Appendix A.4.

6 Related Works

6.1 Elo-based LLM Benchmarking

Recent studies have leveraged Elo ratings derived from human preferences to benchmark LLMs (Boubdir et al., 2023). For instance, Chatbot Arena (Chiang et al., 2024) employs Elo as an evaluation metric, while RAGElo (Rackauckas et al.,

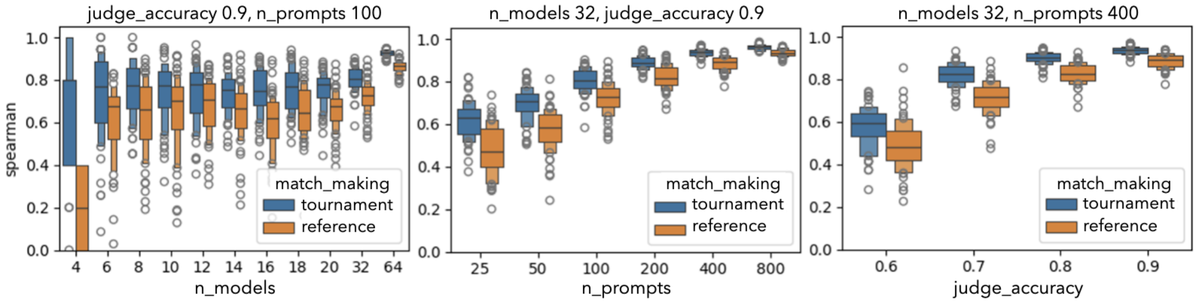

Figure 2: Simulation results comparing the tournament and reference-based approaches. The tournament method consistently outperforms the reference-based approach in Spearman correlation across various control variables: the number of participant LLMs (n_{models}), the number of benchmark prompts ($|X|$), and judge precision (P_{judge}).

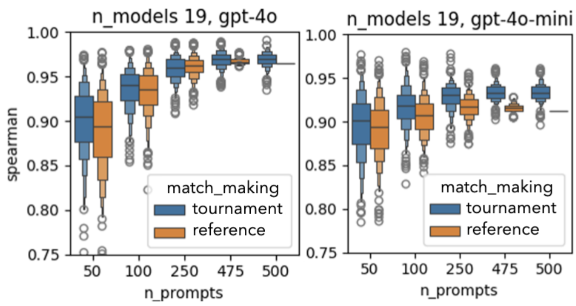

Figure 3: Results of VARCO Arena (tournament) and reference-based approach with `gpt-4o` (left) and `gpt-4o-mini` (right) judge. VARCO Arena constantly records higher Spearman correlation coherent with the Experiment 1 result (Figure 2). Results summary is on Table 2.

$\lvert\Delta_{\text{rank}}\rvert$ (↓)	gt=1-6	7-13	14-19 (20)	total avg.
binary search (4o)	**0.92**	1.84	2.13	1.72
comp. to 1st (4o)	1.98	**1.55**	**1.57**	**1.39**
binary search (4o-mini)	1.27	1.82	**1.21**	1.5
comp. to 1st (4o-mini)	**1.00**	**1.43**	1.43	**1.37**

Table 4: Comparison of the binary search method versus using the top-performing model's response as a reference (*comp. to 1st*) for inserting a new LLM into the leaderboard. We report the mean rank deviation ($\lvert\overline{\Delta}_{\text{rank}}\rvert$) from the ground-truth leaderboard as an additional error metric. For further details, see Algorithm 2 in Appendix.

2024) uses it as a ranking metric. These implementations highlight the benefits of applying Elo ratings for open-ended text quality evaluation.

6.2 Reference-free Evaluation

Reference-free evaluation has emerged as a promising alternative to static reference-based methods from the era of Natural Language Generation. Advances in LLM capabilities have enabled models to assess open-ended responses effectively (Jauhiainen and Guerra, 2024). When reference quality is poor, reference-free metrics like XComet (Guerreiro et al., 2023) works as a more reliable alternatives.

7 Conclusion

We introduced VARCO Arena, a reference-free LLM benchmarking approach that employs a tournament-style framework with direct pairwise comparisons to evaluate the generative capabilities of LLMs. VARCO Arena offers a cost-effective, scalable, and adaptable solution for benchmarking response quality. Our simulation study and empirical experiments demonstrate that VARCO Arena consistently achieves higher rank reliability compared to current reference-based methods, as evidenced by stronger correlations with human-established Elo ratings. Given its robust performance and flexibility, we believe VARCO Arena can serve as a reliable automated tool for model selection and ranking across diverse and evolving use cases. Future work will explore broader applications, such as benchmarking LLMs in multi-modal settings (e.g., those incorporating visual or audio inputs and outputs).

Limitations

Although we tested robustness of the tournaments performed by VARCO Arena, it still has added factor for randomness such as match bracket initialization which does not applies to reference-based method. Rooms for improvement exist for more informative match-making algorithm that would achieve better ranking than single-elimination tournaments within same or less number of matches.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 339–352, Singapore. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

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.

Arpad E Elo and Sam Sloan. 1978. The rating of chessplayers: Past and present. *(No Title)*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal

Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jussi S. Jauhiainen and Agustín Garagorry Guerra. 2024. Evaluating students' open-ended written responses with llms: Using the rag framework for gpt-3.5, gpt-4, claude-3, and mistral-large. *ArXiv*, abs/2405.05444.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From live data to high-quality benchmarks: The arena-hard pipeline.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan

Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. *Preprint*, arXiv:2407.06551.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zackary Rackauckas, Arthur Câmara, and Jakub Zavrel. 2024. Evaluating rag-fusion with ragelo: an automated elo-based framework. *Preprint*, arXiv:2406.14783.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand

Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *Preprint*, arXiv:2307.03025.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations (ICLR)*.

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. Curran Associates, Inc.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

A Appendix

A.1 Machine Requirements for Experiments

Except the part we inferenced open-weight models such as `Llama`, `Qwen` and `Gemma`, our experiments are mostly do not require GPU usage. Inference are done on one A100 GPU, but T4 would be enough for reproducing our experiments. Otherwise, our experiments require querying API and post-processing those with CPU. Experiments could be run on personal desktops. The lowest specification of the machine we deployed had `i5-8400 CPU`, 16 GiB RAM.

A.2 Assuring Statistical Significance of the Results within Budget for proprietary models

To ensure a statistically significant number of trials for each experiment while staying within budget, we utilize OpenAI's Batch API to prepare full-grid match outcomes (i.e., all-play-all matches for every prompt) in a cache file, allowing us to reuse these outcomes. Each empirical experiment consists of 500 trials per setting, with results represented using whisker plots or summary statistics such as median values. When experimenting with a subset of the Arena-Hard-Auto benchmark ($|X| < 500$), we sample a stratified subset of the benchmark dataset for each new trial.

A.3 Elo ratings from VARCO Arena compared to Human Annotations

Figure 4 shows the Elo ratings computed out of VARCO Arena. For judge, we used `gpt-4o`. As mentioned in the caption, the Elo ratings are bootstrapped median value from 500 trials. 95% confidence intervals also plotted as an error bar, which look negligible in scale compared to observed values. Matches are performed over Arena-Hard-Auto benchmark dataset (500 prompts).

A.4 Binary search vs. Win rate over reference

A.4.1 Binary Search

We tried binary search placement of a newly added LLM to the leaderboard without reference text in Table 5. Details of how we implemented binary search are attached in Appendix 2. It turns out that binary search based on leaderboard ranks is not as reliable as the current approach of scoring the newcomer to the reference outputs. The number of judge operations performed is equivalent to the matches allocated to the least-performant model in a tournament, which is $|X|$ (i.e. maximum possible matches that an LLM could have is $|X| * \log_2 n_{\text{model}}$). Within the size of the benchmark prompts ($|X|$), binary search is incompatible with the current approach of using reference instead.

A.4.2 Comparing to the most performant Model so far: Converting Elo Table back to Win Rate

Assuming we preserved a set of match results and model outputs from the last benchmarking, we could benefit from those to perform insertion. One could pick an appropriate *anchor* LLM as a reference in a leaderboard to estimate the skill of a

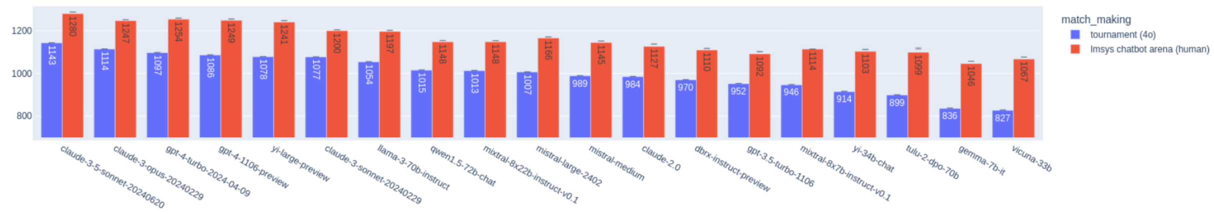

Figure 4: Elo ratings of the model with gpt-4o judge on the full set of Arena-Hard-Auto (Li et al., 2024) prompts. VARCO Arena result (bootstrapped median over 1000 samples of 500 trials) is in blue, plotted alongside the ratings from the ground truth leaderboard in red (Chatbot Arena, *Hard prompts category*). Error bars are 95% confidence intervals.

newcomer. Using previous matches from the tournaments that built the leaderboard could be used for estimating win rates over the reference. This is the same as converting the Elo table into a win rate leaderboard. Since the leaderboard is not built with full-grid matches but with tournaments, there would be some missing matches against the reference regardless we have picked. There are two ways to estimate the win rate over the reference model. We could just count the matches given are enough in amount, or we could also convert Elo ratings back to $P(i > a)$ to use it directly for scoring for the model ranks in the leaderboard. Reminding that Elo rating is purposed for expecting a likely outcome of the match, this should work. After this win rate of the newcomer model $P^*(n > a) = \frac{\text{count(n wins)}}{|X|}$ could be directly compared for enlisting.

A.5 Separability In terms of Confidence Interval

To see how well the two benchmarking approach (*anchored comparison* and tournament approach) separates LLMs in adjacent ranks, we provide scatter plot of Elo rating and win rate paired with error bar (95% confidence interval). We present the both results of using `gpt-4o` (Figure 5) and `gpt-4o-mini` (Figure 5) as a judge. Inside the each plot, inseparables indicates the cases where any pair of datapoint co-includes each other within their range of error bars, and overlap means a certain datapoint is within some other's range of error, when it is one-sided.

A.6 Judge configuration

A.6.1 Evaluation Prompt

We use the prompt from LLMBar. The prompt depicted in Figure A.6.2. We added 4 questions for criteria of our own to `Metrics.txt` prompt of (Zeng et al., 2024). You can refer to the original

$\|\Delta_{rank}\|$ (↓)	gt=1	2	3	4	5	6	avg.
binary search (4o)	0.09 (.04/-.03)	1.24 (.14/-.14)	**1.75** (.09/-.09)	**1.55** (.07/-.06)	1.26 (.08/-.08)	1.10 (.10/-.09)	**0.92**
anchored (4o)	**0.00** (0.00/0.00)	**1.01** (0.01/-0.01)	1.95 (0.02/-0.02)	2.00 (0.00/0.00)	**0.96** (0.02/-0.02)	**0.30** (0.04/-0.04)	1.98
binary search (4o-mini)	0.52 (.09/-.07)	0.85 (.12/-.11)	0.59 (.10/-.09)	2.03 (.02/-.02)	1.20 (.05/-.05)	2.45 (.07/-.06)	1.27
anchored (4o-mini)	0.00 (0.00/0.00)	0.00 (0.00/0.00)	1.00 (0.00/0.00)	2.00 (0.00/0.00)	2.00 (0.00/0.00)	1.00 (0.00/0.00)	**1.00**

7	8	9	10	11	12	13	avg.
1.31 (.10/-.10)	1.27 (.11/-.11)	2.22 (.14/-.12)	1.74 (.09/-.09)	2.27 (.12/-.11)	2.23 (.12/-.12)	1.86 (.07/-.07)	1.84
0.30 (0.04/-0.04)	3.68 (0.04/-0.04)	1.09 (0.03/-0.03)	1.03 (0.02/-0.01)	2.97 (0.02/-0.02)	0.78 (0.05/-0.05)	1.00 (0.00/0.00)	**1.55**
0.69 (.07/-.06)	0.85 (.09/-.09)	3.89 (.12/-.11)	1.95 (.06/-.05)	2.10 (.03/-.03)	2.37 (.10/-.11)	0.88 (.12/-.11)	1.82
0.51 (0.49/-0.51)	0.52 (0.48/-0.52)	3.50 (0.49/-0.51)	1.00 (0.00/0.00)	1.00 (0.00/0.00)	3.00 (0.00/0.00)	0.50 (0.50/-0.50)	**1.43**

14	15	16	17	18	19	20	avg.
1.40 (.04/-.05)	3.07 (.11/-.11)	0.80 (.08/-.09)	1.47 (.05/-.04)	5.00 (.11/-.11)	0.96 (.08/-.09)	-	2.13
2.00 (0.00/0.00)	2.00 (0.00/0.00)	1.00 (0.00/0.00)	1.21 (0.03/-0.04)	3.00 (0.00/0.00)	0.21 (0.04/-0.03)	-	**1.57**
1.45 (.07/-.08)	4.20 (.17/-.17)	0.19 (.07/-.06)	0.08 (.03/-.02)	1.09 (.05/-.05)	1.08 (.05/-.05)	0.40 (.07/-.07)	**1.21**
1.00 (0.00/0.00)	2.00 (0.00/0.00)	2.00 (0.00/0.00)	1.00 (0.00/0.00)	1.00 (0.00/0.00)	3.00 (0.00/0.00)	0.00 (0.00/0.00)	1.43

Table 5: Binary search vs. *Anchored comparison*: Mean rank deviation ($|\Delta_{rank}|$) from ground-truth leaderboard. Result of binary search placement and anchored comparison insert by `gpt-4o[-mini]` judge are provided with bootstrapped 95% confidence interval (500 trials, 1000 samples, $|X|$=500, Arena-Hard-Auto (Li et al., 2024)).

Algorithm 2 Binary Search for Enlisting new LLM to a leaderboard

Require: Leaderboard L, new model m_{new}, test prompts X, outputs O_{ij}, assumes $|X| > |L| > n_{\text{comparisons}}$

Ensure: Updated leaderboard L' with m_{new} placed

$n_{\text{comparisons}} \leftarrow \lfloor \log_2(|L|) \rfloor$
$n_{\text{matches}} \leftarrow \lfloor |X| / n_{\text{comparisons}} \rfloor$
function BINARYSEARCHPLACEMENT(L, m_{new})
    X $\leftarrow$ Shuffle(X)
    X $\leftarrow$ concat(X;X)
    low $\leftarrow 0$
    high $\leftarrow |L| - 1$
    while low $\leq$ high **do**
        mid $\leftarrow \lfloor (\text{low} + \text{high})/2 \rfloor$
        wins $\leftarrow 0$
        for $i \leftarrow 1$ to n_{matches} **do**
            $x \leftarrow X$.pop()
            if Match($m_{\text{new}}, L[\text{mid}], x$) $= m_{\text{new}}$ **then**
                wins $\leftarrow$ wins $+1$
            end if
        end for
        if wins $> n_{\text{matches}}/2$ **then**
            high $\leftarrow$ mid -1
        else if wins $< n_{\text{matches}}/2$ **then**
            low $\leftarrow$ mid $+1$
        else if $|X| > 0$ **then**
            continue ▷ Ensure tie
        else
            return mid, tie ▷ Tie
        end if
    end while
    return low, non-tie ▷ Position found
end function
function UPDATELEADERBOARD(L, m_{new})
    position, istie $\leftarrow$ BinarySearchPlacement(L, m_{new})
    $L' \leftarrow L$.insert(position, m_{new}, istie)
    return L'
end function

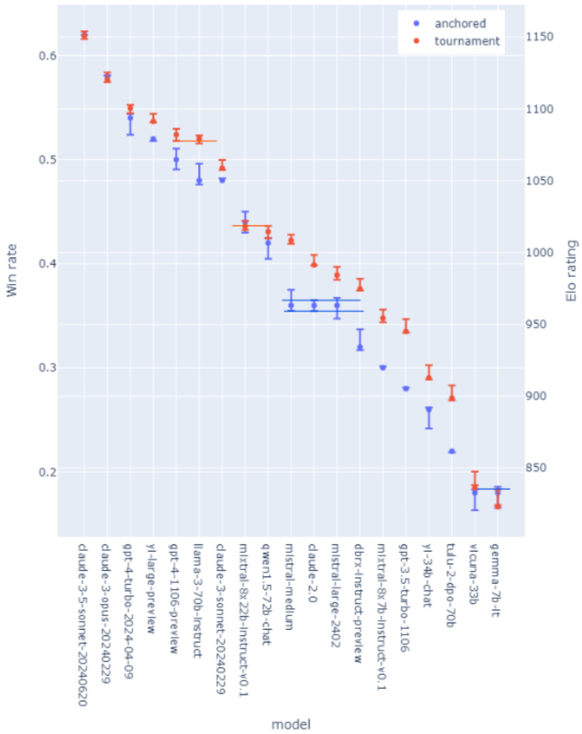

Figure 5: gpt-4o result of *anchored comparison* and tournament approach. 1000 bootstrapped median from 500 observations used for confidence interval estimation.

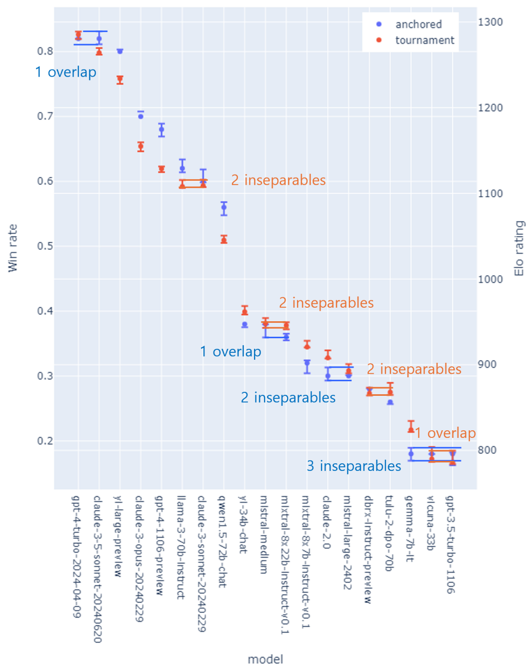

Figure 6: gpt-4o result of *anchored comparison* and tournament approach. 1000 bootstrapped median from 500 observations used for confidence interval estimation.

prompt in LLMBar github.

A.6.2 Decoding Parameters

We did not configure decoding parameters of judge LLMs (`gpt-4o[-mini]`), which its temperature defaults to 1. The only parameter we have adjusted is maximum number of tokens to be generated, which for our prompt is less than 6 (i.e. The output of our prompt is `(a)` or `(b)`). To avoid position bias, we alternated the position of the responses from a certain model across the benchmark prompt.

PROMPTS = [# metrics.txt from LLMBar

"role": "system", "content": "You are a helpful assistant in evaluating the quality of the outputs for a given instruction. Your goal is to select the best output for the given instruction.", ,

"role": "user", "content": """Select the Output (a) or Output (b) that is better for the given instruction. The two outputs are generated by two different AI chatbots respectively.

Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the output honestly/precisely/closely executes the instruction, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.
(2) Outputs should NOT contain more/less than what the instruction asks for, as such outputs do NOT precisely execute the instruction.
(3) You should avoid any potential bias and your judgment should be as objective as possible. For example, the order in which the outputs were presented should NOT affect your judgment, as Output (a) and Output (b) are **equally likely** to be the better.

Do NOT provide any explanation for your choice.
Do NOT say both / neither are good.
You should answer using ONLY "Output (a)" or "Output (b)". Do NOT output any other words.

Instruction:
instruction

Output (a):
response_a

Output (b):
response_b

Questions about Outputs:
Here are at most three questions about the outputs, which are presented from most important to least important. You can do the evaluation based on thinking about all the questions.
- Does the output well satisfy the intent of the user request?
- If applicable, is the output well-grounded in the given context information?
- Does the output itself satisfy the requirements of good writing in terms of:
1) Coherence
2) Logicality
3) Plausibility
4) Interestingness

Which is better, Output (a) or Output (b)? Your response should be either "Output (a)" or "Output (b)":""",
,]