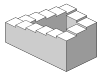

The Progress Illusion

Revisiting meta-evaluation standards of LLM evaluators

Tianruo Rose Xu[1], Vedant Gaur[2], Liu Leqi[3], Tanya Goyal[1]
[1]Cornell University [2]University of Pennsylvania [3]University of Texas at Austin
tx88@cornell.edu, tanyagoyal@cornell.edu

Abstract

LLM judges have gained popularity as an inexpensive and performant substitute for human evaluation. However, we observe that the meta-evaluation setting in which the reliability of these LLM evaluators is established is substantially different from their use in model development. To address this, we revisit meta-evaluations of LLM evaluators under a setting that more closely aligns with practice by examining evaluators' ability to distinguish test system pairs that are closer in capability. Our fine-grained approach shows that all LLM evaluator's correlations with human judgments are concerningly low when the models perform similarly, showcasing a key limitation of current norms. Equipped with this better methodology, we next analyze the impact that the choice of the reference model makes to LLM-as-a-judge evaluator performance. We show that single-reference evaluators only perform well at ranking test systems that fall within particular capability ranges, even if the standard meta-evaluation reports high overall correlation. Taken together, our analysis shows critical issues with current LLM meta-evaluation and recommend avenues for improvement.

1 Introduction

Human evaluation is broadly accepted as the gold standard for benchmarking LLMs. However, it is time-consuming and infeasible to run human evaluation on each new model iteration during development, e.g., to test out hyperparameter choices. As a result, automatic proxies of human judgments (Zheng et al., 2023; Lin et al., 2024) are widely used. For example, more than 90% of the recent works on preference optimization only report results using the automatic evaluator AlpacaEval (Li et al., 2023; Dubois et al., 2023, 2024).

How can we **meta-evaluate** an automatic evaluator, i.e. verify if it is reliable? Benchmarks like Chatbot Arena (Chiang et al., 2024) that collect human preference judgments from millions of users play a crucial role here. They provide the "ground truth" rankings of test LLMs that the automatic evaluator rankings can be validated against. In fact, the most popular automatic evaluators today, including AlpacaEval, WildBench (Lin et al., 2024), MixEval (Ni et al., 2024) and Arena-Hard (Li et al., 2024), verify their validity by reporting high correlation with Chatbot Arena judgments.

In this paper, we make two main contributions. First, we revisit the meta-evaluation norms used to evaluate automatic LLM evaluators in the community. Our analysis of 8 popular LLM judges (Li et al., 2023; Lin et al., 2024; Ni et al., 2024; Li et al., 2024) shows **notable differences in performance of LLM judges in settings they are meta-evaluated vs. used during model development**. During meta-evaluation, an LLM judge is used to rank test systems and a high correlation of this ranking with humans indicates that the judge is reliable.[1] However, the test systems ranked in meta-evaluation settings often have fairly different capability levels. Hence, ranking (or scoring) them is relatively easier compared to using these judges during model development. In the latter case, very similar models are compared, e.g. those that have the same underlying base model but different training recipes.

Based on these observations, we propose a two-pronged meta-evaluation strategy that is more fine-grained and informative. Our approach includes a **delta analysis** (borrowing from Deutsch et al. (2022)) that examines evaluator ability to distinguish similar capability models and **stratified analysis** that identifies performance ranges of test models within which a given LLM evaluator is performant. Together, these help determine if an LLM evaluator's reported improvements for a test model is meaningful in practice.

[1]All evaluators we analyzed reported > 0.8 Kendall's τ.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19033–19043

Our experiments show that the popularly used LLM judge AlpacaEval reports Kendall's $\tau = 0.86$ under standard meta-evaluation conditions but 0.19 correlation with humans when ranking models with less than 2 points difference in AlpacaEval scores; this corresponds to typical improvements reported using this metric. We observe a similar concerning trend for other automatic evaluators, including WildBench (Lin et al., 2024), Arena-Hard (Li et al., 2024), MT-bench (Zheng et al., 2023) and MixEval (Ni et al., 2024).

Next we examine how the choice of reference model in LLM-as-a-judge evaluators impacts their performance. We first conduct our analysis in the **ground truth setting** where we compare Chatbot Arena's official rankings, derived via Bradley-Terry on pairwise preferences between two randomly sampled models, against a single reference version of the same dataset. We find that the rankings of the latter version (this strategy mirrors reference-based LLM judges like AlpacaEval), differs substantially from the ground truth Chatbot Arena rankings across all reference models. Our results show that models that most closely match the test model capabilities are generally the better choice for reference models in single reference settings.

Finally, we report results in the **LLM evaluator setting** where we compare rankings of popular leaderboards like AlpacaEval (Li et al., 2023) and WildBench (Lin et al., 2024) against gold Chatbot Arena rankings. Our results here confirm that high correlations over the entire test model can hide extremely low correlations for realistic test settings. Overall, our results and analysis show critical issues with current LLM evaluators and recommend avenues for both improving these evaluators and the meta-evaluation standards themselves.

2 Our Meta-Analysis Methodology

2.1 Background and Notation

The goal of both automatic and human evaluation is to assign scores to a set of n test systems $\Pi = \{\pi_i\}_{i \in [n]}$. Below we describe the basic methodology used by reference-based evaluators.

Reference-based LLM Evaluators Let $\mathcal{E}$ be a reference-based evaluator. Broadly, $\mathcal{E}$ can be characterized by the tuple $(\mathcal{X}, \pi_{\text{ref}}, \pi_{\text{judge}})$, which refers to a prompt set, a reference model and a judge model, respectively. For each $x \in \mathcal{X}$, the judge model π_{judge} determines whether the quality of the test model output, i.e., $\pi(x)$, exceeds the quality of the reference model output, i.e. $\pi_{\text{ref}}(x)$. These evaluators then report the win-rate score for test models based on these preference judgments:

$$S(\pi; \mathcal{E}) := \mathbb{E}_{x \sim \mathcal{X}} \left[\mathbb{I}_{\mathcal{E}} \{ \pi_{\text{ref}}(x) \prec \pi(x) \} \right] \quad (1)$$

The ranking $\mathcal{R}(\Pi; \mathcal{E})$ for models $\pi \in \Pi$ is determined using this win-rate $w(\pi; \mathcal{E})$ with respect to the reference model π_{ref}.

Prior research has determined best practices for various components of this evaluation pipeline, such as selecting the prompt set $\mathcal{X}$ (Ni et al., 2024; Lin et al., 2024), the judge model π_{judge} (Kim et al., 2023) using pairwise comparisons against a reference v/s independently scoring each instance without a reference output (Kim et al., 2023), and designing instruction prompt set used by π_{judge} (Zeng et al.). We add to this body of work in this paper by systematically analyzing how the choice of reference model(s) $\pi_{\text{ref}}(x)$ impacts the performance of LLM-as-a-judge evaluators. In the next section, we discuss community norms for meta-evaluating LLM evaluators and describe the methodology we use in this paper.

Community norms for meta-evaluating automatic evaluators Human judgments remain the gold standard for evaluating the quality of free-text responses. Suppose we have access to human rankings $\mathcal{R}_{\text{gold}}$ for a set of models Π'. We can validate if an automatic LLM evaluator is reliable by comparing its rankings $\mathcal{R}_{\mathcal{E}}(\Pi)$ against these gold standard rankings $\mathcal{R}_{\text{gold}}(\Pi')$ for the common set of models:

$$\text{Corr}(\mathcal{R}_{\mathcal{E}}(\Pi \cap \Pi'), \mathcal{R}_{\text{gold}}(\Pi \cap \Pi')) \quad (2)$$

Automatic evaluators are considered reliable if they have high correlation with these ground truth human rankings. For e.g., the broadly used AlpacaEval metric validated their approach by reporting a Spearman correlation of 0.98 with Chatbot Arena rankings (Chiang et al., 2024) which is a public leaderboard collated using human judgments on an open chat platform. Similarly, WildBench (Lin et al., 2024) and Arena-Hard (Li et al., 2024) report 0.95 and 0.91 Pearson correlations against Chatbot Arena respectively.

2.2 Revisiting meta-evaluation standards

Meta-evaluation of generation metrics has been a widely studied problem in NLP in traditional

Evaluator	τ_{standard}	Δ used to compute τ_Δ					
		0.3	0.5	1	2	5	10
LC AlpacaEval 2.0	.869	.230	.263	.280	.196	.517	.730
AlpacaEval 2.0	.815	.294	.083	.116	.319	.616	.763
AlpacaEval 1.0	.583	-.547	-.617	-.087	-.140	.235	.336
WildBench **	.824	–	–	–	–	–	.440
Arena-Hard	.773	-.125	-.083	.000	.207	.325	.470
MT-bench	.788	.216	.349	.583	.746	.788	.788
MixEval-Hard	.867	-1.000	-.600	-.250	.076	.555	.739
MixEval	.808	.000	.200	.142	.384	.333	.633

Table 1: Kendall's τ correlation (both standard τ_{standard} and τ_Δ between evaluators and human rankings derived from Chatbot Arena. We highlight the cell corresponding to the approximate reported improvement, i.e. Δ, in recent works for each evaluator in gray. Our results show that correlation between automatic and human rankings on these realistic system pairs is very low. **: We omit correlation numbers for certain Δs where the number of realistic system pairs is lower than 15.

text generation fields, particularly summarization (Fabbri et al., 2020; Deutsch et al., 2021; Tang et al., 2022; Goyal et al., 2022). Multiple works in this line have discussed the importance of aligning this meta-evaluation settings (e.g. the choice of test models compared or their score ranges) to the settings in which they will be used in practice (Deutsch et al., 2022; Peyrard, 2019).

Consider a metric $\mathcal{E}$ on which research papers, on average, report improvement δ. **Is this improvement meaningful?** Deutsch et al. (2022) show that reported improvements in published works is generally much smaller than the average difference between model pairs that were used to validate the metric $\mathcal{E}$ using Equation 2. Their work proposes a modification to Equation 2 to determine the reliability of $\mathcal{E}$ in realistic improvement ranges. In this paper, we borrow their methodology, described below, to investigate this for popularly used LLM evaluators today.

Approach We follow Deutsch et al. (2022) and use Kendall's τ as the correlation function throughout this paper. Given a set of n test models, Kendall's τ depends on the number of model pairs out of $\binom{n}{2}$ which are ranked the same by $\mathcal{R}_\mathcal{E}(\Pi)$, i.e. the metric being validated, and $\mathcal{R}_{\text{gold}}(\Pi)$, i.e. the ground truth human rankings:

$$\tau = \frac{P - Q}{\sqrt{P + Q + T}\sqrt{P + Q + U}} \quad (3)$$

where P and Q are the number of models that are ranked similarly and different by $\mathcal{R}_\mathcal{E}$ and $\mathcal{R}_{\text{gold}}$, and T and U are the ties by the two evaluators respectively.

To meta-evaluate evaluators only on realistic system pairs, they propose re-computing the correlation using only the subset of $\binom{n}{2}$ pairs for which the score difference is below a pre-defined margin δ. This margin can be chosen for each metric independently based on the improvements reported on it in recent work. We call this modified metric $\tau_{\Delta<\delta}$. This metric can be used to quantify how reliable the evaluator $\mathcal{E}$ is when used to report small improvements in score.

Experiment Setup for Preliminary Analysis We run experiments using the leaderboard scores of Chatbot Arena (Zheng et al., 2023; Chiang et al., 2024) as the human ground truth $\mathcal{R}_{\text{gold}}$. Note that Chatbot Arena is a broadly trusted leaderboard and used to meta-evaluate all recent LLM evaluators, including Length-controlled AlpacaEval (Li et al., 2023), WildBench (Lin et al., 2024), Arena-Hard (Li et al., 2024), and others.

We benchmark the following **reference-based** LLM evaluators in this section: (a) AlpacaEval-v1 (Dubois et al., 2023), (b) Length-controlled AlpacaEval (Li et al., 2023), (c) WildBench (Lin et al., 2024), and (d) Arena-Hard (Li et al., 2024). These are the most commonly used evaluators, particularly length-controlled AlpacaEval, in recent months.[2] Finally, although our focus is reference-based evaluators, we additionally include other popular benchmarks MT-Bench (Zheng et al., 2023) and MixEval (Ni et al., 2024) in our meta-evaluation.

Automatic LLM evaluators report low correlations with ground truth on realistic system

[2]Note that we only consider evaluators that measure the generation quality of free-form text, which is an inherently subjective task and aligns with the Chatbot Arena data. We omit more objective benchmarks, such as math (Cobbe et al., 2021; Hendrycks et al.) or multiple-choice factuality (Hendrycks et al., 2020a), as these measure orthogonal capabilities where determining output correctness is more straightforward and does not require LLMs.

pairs. We report standard (τ_{standard}) and modified (τ_{Δ}) Kendall's τ correlation for the above evaluators against Chatbot Arena's human rankings in Table 1. Across all evaluators, τ_{standard} is greater than 0.7. In contrast, τ_{Δ} with Δ values equal to improvement margins reported in recent works is less than 0.4.[3] For example, on LC AlpacaEval 2.0, the observed Kendall's $\tau_{\Delta=2}$ is only 0.196. This indicates that improvements on these metrics within such small ranges are not meaningful. Overall, our analysis shows that there exists a mismatch in the settings where these metrics where validated and their usage.

Based on these observations, **we recommend that metric developers use realistic model pairs to evaluate a given metric or evaluator.** Similarly, practitioners should use the modified Kendall's τ metric to determine whether the observed difference in metric scores signify meaningful improvements over baselines.

2.3 Meta-Analysis Methodology

Let $S(\Pi; \mathcal{E})$ and $S_{gold}(\Pi)$ refer to the scores of test models Π determined by the automatic LLM evaluator $\mathcal{E}$ and the ground truth scores respectively. Guided by the insights from our preliminary experiment, we measure the correlation of automatic LLM evaluators with human gold standard using the following two correlation approaches:

i) Delta Correlation Analysis This exactly follows the approach from Deutsch et al. (2022) described above. Specifically, we report correlation $\tau_{\Delta<\delta}$ describes correlation computed on realistic system pairs (π_1, π_2) for which the score difference $|S(\pi_1; \mathcal{E}) - S(\pi_2; \mathcal{E})| < \delta$. We follow prior work and modify Kendall's τ rank correlation to report this statistic.

Intuitively, this metric helps us quantify how meaningful is an improvement of δ reported by automatic evaluator $\mathcal{E}$. Ideally, when a new training method or model is released, their reported improvement δ' using $\mathcal{E}$ should be such that $\tau_{\Delta<\delta'}$ with human judgments is high.

ii) Stratified Rank Change We stratify the test models Π into k tiers, Π_i for $i \in \{1, 2..., k\}$, based on their S_{gold} scores. For $k = 4$, Tier 1 or Π_1 includes models in the top 25 percentile, Π_2 includes models in the 25–50 percentile, Π_3 has 50–75 percentile models, and Π_4 contains the bottom 25 percentile of the models.

For stratified rank change, we compute the average rank difference between an evaluator's ranking and the gold ranking for each model within a tier. Lower values indicate closer alignment with human preferences:

$$\Delta_{\text{rank}}(\Pi_i; \mathcal{E}) = \frac{1}{|\Pi_i|} \sum_{\pi \in \Pi_i} \left| \mathcal{R}_{\mathcal{E}}(\pi) - \mathcal{R}_{\text{gold}}(\pi) \right| \tag{4}$$

where Π_i is the set of test models in tier i, $\mathcal{R}_{\mathcal{E}}(\pi)$ is the evaluator's assigned rank of model π, and $\mathcal{R}_{\text{gold}}(\pi)$ is the gold (human) rank of the same model within the tier.

Motivation for stratified analysis Our hypothesis is that, given an LLM evaluator $\mathcal{E} = (\mathcal{X}, \pi_{\text{ref}}, \pi_{\text{judge}})$, the effectiveness of $\mathcal{E}$ will depend on the capability difference between π_{ref} and models in the test tier Π_i. Consider a toy example with reference model π_{ref}'s and the two test models π_1 and π_2 that have the following capability order $\pi_{ref} >> \pi_1 > \pi_2$. Since π_{ref} is much stronger than both test models, its outputs are preferred over those of both π_1 and π_2 for all prompts $x \in \mathcal{X}$. In this scenario, the win-rate scores of both models will be 0, and the evaluator $\mathcal{E}$ cannot distinguish between them.

On the other hand, if π_{ref} and π_1 have very similar outputs, e.g. if they are derived from the same base model with only slight differences in fine-tuning, it is likely that the judge π_{judge} cannot correctly choose the better output between the two or that the judgment is inherently ambiguous (Zhao et al., 2025). In this scenario as well, the score $S(\pi_1; \mathcal{E})$ may not be reliable.

This stratified rank change methodology allows us to examine evaluator reliability in more realistic low-difference settings, similar to the delta analysis. Additionally, it provides insights into best practices for selecting the π_{ref} model(s) if the test models' capability range is known.

3 Experiment Setup

3.1 Ground Truth Source: Chatbot Arena

We use Chatbot Arena (Zheng et al., 2023; Chiang et al., 2024) to obtain the ground truth score and rankings of test models. Despite some recent reservations (Zhao et al., 2025; Singh et al., 2025), it

[3] We use the following Δ values to indicate average margin of improvement in recent works: AlpacaEval (all variants), WildBench, ArenaHard, MixEval $\rightarrow$ 2, MT-Bench $\rightarrow$ 0.5.

is broadly trusted to evaluate both newly released models[4] and meta-evaluate LLM evaluators (Li et al., 2023; Lin et al., 2024; Li et al., 2024).

Chatbot Arena collects human preferences from users visiting their web interface. A user on the platform submits a query x and is shown outputs sampled from two different models $\pi_i(x)$ and $\pi_j(x)$; the identity of these models is kept anonymous. The user submits a label $l \in \{i, j, \text{tie}\}$ to indicate their preference. Based on these preference labels, the platform estimates the pairwise win-rate of each pair of model, i.e. $P(\pi_i > \pi_j)$. Finally, this estimate is used by the Bradley-Terry model to derive scores $S_{\text{gold}}(\pi_i)$ for each test model $\pi_i \in \Pi$:

$$P(\pi_i > \pi_j) = \frac{e^{S_{\text{gold}}(\pi_i)}}{e^{S_{\text{gold}}(\pi_i)} + e^{S_{\text{gold}}(\pi_j)}} \quad (5)$$

In all our experiments, we will use these Chatbot Arena scores S_{gold} derived using Bradley-Terry as the ground truth.

3.2 Evaluation Setting 1: Meta-Evaluating Reference-based Human Evaluation

First, we analyze the effect of the choice of reference model in the ground truth setting itself, i.e. in ChatBot Arena. As described above, Chatbot Arena collects pairwise preference annotations for models π_i and π_j. However, instead of using Bradley-Terry on estimated win-rates between distinct model pairs, we can mimic the same reference-based strategy of Equation 1 to derive test models' scores and rankings. Specifically, we use the collected dataset to compute win-rates against a fixed reference model π_{ref}. Note that this uses only a subset of the entire Chatbot Arena dataset, i.e. those that have preference labels involving the π_{ref} model.

Dataset Statistics For this analysis, we require Chatbot Arena's human preference dataset, i.e. tuples (π_i, π_j, l). Although Chatbot Arena does not publicly release its data, such a subset was released in June 2023. We conduct our experiments using the largest publicly released dataset by Chatbot Arena. It consists of 55k preference annotations;[5] it includes response pairs sampled from two of 64 unique models and the corresponding pairwise preference annotation.

For this dataset, we conduct stratified rank change analysis using the 10 most commonly occurring models in the above released dataset. Basic statistics for this are included in Table 2 and full list in Table 7 of the Appendix.

3.3 Evaluation Setting 2: Meta-Evaluating Reference-based LLM Evaluation

We conduct our main experiments on two most widely used reference-based LLM evaluators:

1. **AlpacaEval** (Li et al., 2023; Dubois et al., 2023) contains a test prompt set $\mathcal{X}$ of size 805, and uses $\pi_{\text{ref}} =$ `GPT-4 Preview (11/06)`.

   For both delta and stratified analysis, we report results by varying the π_{ref} model while keeping all other specifications the same (e.g. prompt set $\mathcal{X}$, π_{judge} and evaluation prompt used to elicit judgment from π_{judge}).

   Specifically, we report results using 20 different π_{ref} models (see Table 8 in Appendix for the list). These reference models are selected so as to cover all 4 performance tiers after the stratification described in Section 2.3. We use the AlpacaEval methodology (using their released code[6]) to obtain model scores $S_{\mathcal{E}}^{\pi_{\text{ref}}}(\Pi)$ for each of the 20 π_{ref} models. This list includes the default `GPT-4 Preview (11/06)` used by official AlpacaEval.

2. **WildBench** (Lin et al., 2024) contains a carefully curated test prompt set of size 1024. WildBench reports two kinds of metrics: a reference-based pairwise metric `WB-Reward` and an individual metric `WB-Score`. In this paper, we limit our analysis to only the reference-based metric.

   Note that `WB-Reward` computes the score of each test model against three different reference models – GPT-4-Turbo-0429, Claude-3-Haiku, and Llama-2-70B-chat. The final WB-Reward score for each test model is computed as the average reward of these three individual rewards. For WildBench, we perform delta analysis for these above 3 reference models.[7] We omit strati-

[4] Example borrowed from (Zhao et al., 2025): Google's Chief Scientist used high performance on Chatbot Arena to declare the success of their recent model release: https://tinyurl.com/55xs2pz4.

[5] https://huggingface.co/datasets/lmsys/lmsys-arena-human-preference-55k

[6] https://github.com/tatsu-lab/alpaca_eval/

[7] Increasing the number of reference models requires collecting additional preference judgments using π_{ref}, as described for AlpacaEval. We found this to be prohibitively expensive, and therefore restrict our analysis to the three reference models used in the official implementation. We directly use the dataset released by WildBench for our experiments: https://github.com/allenai/WildBench/.

fied analysis due to the small number of models.

Dataset Statistics For meta-evaluating both AlpacaEval and WildBench, we use the Chatbot Arena leaderboard, as of May 15, 2024, as the ground truth.[8] For AlpacaEval, this resulted in 46 models that were common between this Chatbot Arena version and also had model-generated outputs for all 805 test prompts on the AlpacaEval github.[9] We ran AlpacaEval's LLM-based judgment on these 46 (test) x 20 (reference) model pairs. Data statistics for our study are shown in Table 2. For WildBench, we report results on the 21 common models between the leaderboards as of May 15, 2024.

4 Results and Analysis

We report results for evaluation setting 1 (described in Section 3.2) and 2 (described in Section 3.3) in Section 4.1 and Section 4.2 respectively.

4.1 Meta-evaluation of reference-based human annotation

Reference-based evaluators are suboptimal in the ground truth setting First, we report the correlation between the official Chatbot Arena rankings, computed using Bradley-Terry over the entire collected dataset, and the rankings derived from win-rates against fixed reference models. Table 3 shows our results using the 10 most common models in the released dataset as references. We find that the correlation is < 0.6 for all reference models. Surprisingly, these correlation numbers are even lower than those in Table 1 even though most of the LLM evaluators there are themselves reference-based. However, we note that this comparisons is not completely apples-to-apples, as the number of overlapping test models is different between Table 1 and Table 3.

Choice of reference model significantly impacts evaluator performance Crucially, we observe in Table 3 that different reference models report very different correlations with the official Chatbot Arena rankings. This variance even in the ground truth setting strongly motivates our investigation of reference model selection for LLM-based evaluation in Section 4.2. Finally, we highlight that the reference model reporting the lowest correlation

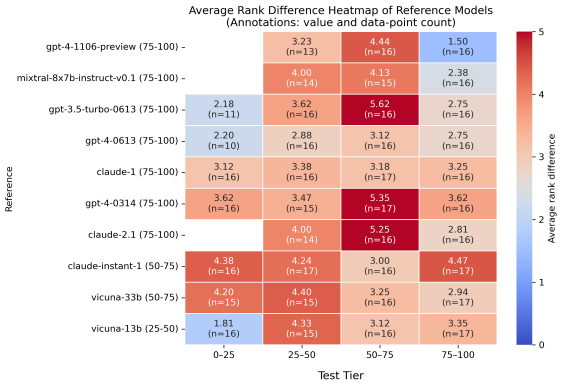

Figure 1: Average rank difference between the Bradley-Terry Chatbot Arena ranking and ranking derived form single-reference battles; each row corresponds to a different π_{ref}. Our results show that π_{ref}'s performance, i.e. its ability to distinguish between test models in a particular tier, is roughly dependent on its own tier membership.

(`gpt-4-1106-preview`; $\tau = 0.3988$) is the reference model used in the AlpacaEval leaderboard.

Next, we report results using our **stratified rank change methodology** to understand the impact of reference model choice in a more fine-grained manner. Figure 1 shows our results; each row reports rank difference between the the ground truth rankings and those derived using a particular reference model; each row corresponds to one π_{ref} and columns correspond to π_{test} that are in a particular percentile-stratified tiers. The percentile tier to which π_{ref} belongs is also in the row labels. Finally, numbers in parentheses indicate the number of models over which the rank change is computed.[10]

In the ground truth settings, reference models are generally better at ranking models in tiers closer to its own We see that the reference models capability at distinguishing between test models is roughly correlated with its performance difference with these test models. For example, we see that the reference model $\pi_{\text{ref}} =$ `gpt-4-1106-preview` reports the worser performance on the other test tiers, but outperforms all others in its own tier, i.e. the 75-100 percentile tier. Similarly, the reference model that performs the best for the 0-25 percentile group is itself a member of a close tier of 25-50 percentile.

Both Table 3 and Figure 1 results demonstrate

[8] https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard

[9] https://github.com/tatsu-lab/alpaca_eval/tree/main/results

[10] Note that each tier Π_i contains $64/4 = 16$ test models. However, the released dataset may not have battles between all test models and the chosen reference.

Evaluator	$\lvert\mathcal{X}\rvert$	π_{ref}
Evaluation Setting 1: Human evaluation		
Chatbot Arena	4.9k*	Official: Varies; uses Bradley-Terry Our Analysis: 10 most common models in released dataset.
Evaluation Setting 2: LLM evaluation		
AlpacaEval	805	Official: `GPT-4 Preview (11/06)` Our Analysis: Randomly selected 20 models (See Table 8)
WildBench	1024	Official: `GPT4-Turbo-0429`, `Claude3-Haiku`, `Llama2-70B-chat` Our Analysis: same

Table 2: Data statistics for our meta-evaluation study. Note that the prompt set $\mathcal{X}$ is the same for all reference models in setting 1, but differs in setting 2. *: We report the **average** number of prompts corresponding to the 10 reference models used in our Chatbot Arena analysis. List and number of battles for each reference model is shown in Table 7.

Reference Model	τ w/ Orig. Ranking
claude-instant-1	0.5833
gpt-4-0613	0.5751
vicuna-33b	0.5465
gpt-4-0314	0.5380
claude-1	0.5173
vicuna-13b	0.5008
mixtral-8x7b-instruct-v0.1	0.4973
gpt-3.5-turbo-0613	0.4702
claude-2.1	0.4329
gpt-4-1106-preview	0.3988

Table 3: Correlation with the standard Chatbot Arena rankings (i.e. rankings computed using Bradley Terry over pairwise model battles) and reference-based rankings (i.e. rankings based on battles against a single reference model). Our results show that reference-based rankings are poorly correlated with Chatbot Arena, even when using expert labelers.

that even in the ground truth setting, the choice of reference model is critical and that choosing a single reference model is suboptimal. In the next section, we investigate the impact of reference model choice in LLM-as-a-judge settings, i.e. the actual settings that current LLMs are evaluated in.

4.2 Meta-evaluation of reference-based LLM evaluators

4.2.1 AlpacaEval

As described in Section 2.3, we simulate AlpacaEval-style win-rates for 20 different reference models. We stratify all test models (46 in total) into 4 performance tiers using the Chatbot Arena ground-truth rankings. For each tier, we report the average τ for the standard Kendall's Tau correlation and delta correlation analysis in Table 4. For stratified analysis, we report the average rank difference between the Chatbot Arena rankings and

Reference	τ_{standard}	Δ **used to compute** τ_Δ			
		2	5	10	20
Tier 1	0.745	0.123	0.221	0.424	0.627
Tier 2	0.752	-0.006	0.172	0.405	0.634
Tier 3	0.742	0.070	0.217	0.405	0.621
Tier 4	0.705	0.136	0.303	0.411	0.575

Table 4: Standard Kendall's τ and modified Kendall's τ_Δ between reference model rankings in the AlpacaEval dataset and Chatbot Arena scores across different thresholds Δ. Each row corresponds to a different reference tier. Tier 1 = top 75–100% capability, Tier 2 = 50–75%, Tier 3 = 25–50%, Tier 4 = bottom 0–25%.

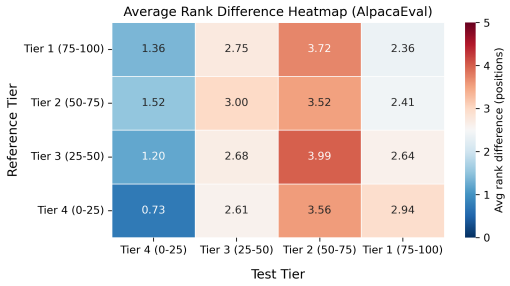

Figure 2: Average ranking difference between Chatbot Arena rankings and simulated AlpacaEval-style win-rate rankings. Rows indicate reference model tier, columns indicate test model tier. Lower values suggest better agreement with human preferences.

those using a particular reference model for each test tier. Figure 2 shows these results.

All reference tiers show notably worse correlations with humans when evaluating models with similar capabilities. Table 4 illustrates this result. This reinforces our main finding that current evaluators perform best at distinguishing models that differ substantially in capability but cannot be used to meaningfully judge differences between realistic system pairs.

Reference	τ_{standard}	Δ used to compute τ_Δ					
		2	5	10	20	30	50
WB Reward	0.900	0.455	0.579	0.538	0.713	0.790	0.867
GPT4T	0.852	-0.333	0.100	0.442	0.606	0.721	0.814
Haiku	0.890	0.000	0.368	0.543	0.676	0.742	0.845
Llama2-70B	0.852	-0.429	0.091	0.333	0.591	0.693	0.809

Table 5: Standard Kendall's τ and modified Kendall's τ_Δ between WildBench's reference model rankings and Chatbot Arena scores across different thresholds Δ.

Tier 4 reference models perform better when ranking models from other tiers We observe that reference models from Tier 4 consistently achieve the lowest, i.e. best, rank difference when evaluating models within a particular tier. As expected, it performs better at ranking models within its own capability tier.

Overall, these results indicate that there does not exist a clear strategy for selecting reference models given a test tier, and that reference model choices must be constantly re-evaluated as newer and better models are released.

4.3 WildBench

WildBench already uses multiple references; How does this strategy fare? We find that the 3 chosen reference models of this leaderboard – GPT-4-Turbo-0429 (GPT4T), Claude-3-Haiku (Haiku), and Llama-2-70B-chat (Llama) fall in Tier 1 (75-100 percentile), tier 2 (50-75 percentile) and tier 3 (25-50 percentile) respectively.

High overall correlations over entire test model can hide extremely low correlations for certain test model ranges We find that all reference models and the aggregate WB Reward metric report $> .85$ rank correlation when considering the entire test model set. Interestingly, although individual reference models report low correlations under the delta-analysis, we find that the overall WB-Reward has a decent correlation (0.455) for $\Delta = 2$. This shows that the aggregation strategy helps overcome the shortcomings of individual reference models.

Single v/s Multiple references under the same budget Aggregated WB Reward outperforms single reference models in Table 6. However, the cost of WB Reward is 3 times that of the single reference models. We ask: can WildBench's multiple reference strategy improve AlpacaEval's single-reference while maintaining the same cost as the original single reference strategy.

Reference Tiers	τ_{standard}	Δ used to compute τ_Δ		
		2	5	10
Tier 1,2,3,4	0.6864	0.0764	0.1835	0.3638
Tier 1,2,3	0.6807	0.0677	0.1800	0.3615
Tier 2,3,4	0.7071	0.0449	0.1670	0.3598
Tier 3,4	0.7172	0.0802	0.2016	0.3752
Tier 1,4	0.6812	0.0766	0.1938	0.3814
Tier 1,3	0.6720	0.0588	0.1769	0.3602

Table 6: Randomized evaluation results under different subsets of reference model tiers. Each row reports the standard Kendall's τ and modified Kendall's τ_Δ between test model rankings (based on simulated comparisons against sampled reference models) and Chatbot Arena scores. Our strategy maintains a fixed total comparison budget across tiers. In contrast, WildBench's approach increases the budget linearly with the number of reference models, making it more costly.

Setup: For each $x \in \mathcal{X}$, we randomly select a reference model π^x_{ref} to compare the test model's output against. Next, this preference data is used to estimate $P(\pi_{\text{test}} > \pi_{\text{ref}})$ for all test and reference model pairs. Similar to Chatbot Arena, we use Bradley Terry to obtain relative scores and rankings for all test models.

Results: Table 6 shows our results. The leftmost column shows the tier membership from which the reference models are sampled with equal probability. **We find that no reference tier combination improves performance consistently over single-reference models for AlpacaEval under the same budget.**

5 Related Work

Traditional benchmarks for evaluating large language models (LLMs) have often relied on close-ended, multiple-choice datasets such as MMLU (Hendrycks et al., 2020b). These benchmarks are advantageous because of clear ground truth, but do not evaluate LLMs in real use settings. LLM evaluators evaluating free-form generations (Li et al., 2024; Lin et al., 2024; Li et al., 2023) provide a way of evaluation beyond these limited formats.

Evaluation methods have shifted toward LLM-

as-a-judge frameworks (Li et al., 2024; Lin et al., 2024; Li et al., 2023; Kim et al., 2023) to rate outputs of other models. These methods typically employ either individual scoring or pairwise comparison (Kim et al., 2023). We discuss, in Section 2.3, meta-evaluation standards used to establish reliability of these evaluators.

6 Conclusion

Our work shows that commonly used reference-based LLM evaluators report high correlations in easy, wide-gap settings but fail to reliably distinguish between similarly capable models. We introduce delta and stratified rank change analysis to address these issues with meta-evaluation and find a substantial different in evaluator performance depending on the choice and capability tier of the reference model. Our findings underscore the need for a more nuanced meta-evaluation framework.

7 Limitations

Our work follows the LLM community's standard practice and uses Chatbot Arena as the source for ground truth annotations. However, recent work (Zhao et al., 2025; Singh et al., 2025) have cast doubts on the reliability of their annotations and rankings. We believe that, since we base the majority of our analysis on Chatbot Arena dataset collected before June 2023, some sources of errors outlined in these papers will not contaminate our analysis. We hope that future work can re-run similar studies as ours with more reliable human data.

Adding to the earlier point, Chatbot Arena only publicly released a subset of its human annotations. The full dataset, including the user prompts, model outputs and battle results for more models would allow us to perform a more exhaustive analysis. Finally, in order to limit the cost of from API calls for eliciting LLM judgments from closed models, we restricted our analysis to two benchmarks.

References

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Preprint*, arXiv:2305.14387.

A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, Richard Socher, and Dragomir R. Radev. 2020. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020b. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *Sort*, 2(4):0–6.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

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.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *arXiv preprint arXiv:2406.06565*.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, et al. 2025. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

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

Wenting Zhao, Alexander M Rush, and Tanya Goyal. 2025. Challenges in trustworthy human evaluation of chatbots. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3359–3365, Albuquerque, New Mexico. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Number of Battles per Model

In Section 3.2, we mentioned that we used a subset of the Chatbot Arena human preference dataset to simulate reference-based evaluation. Table 7 below shows the number of battles each model participated in.

Model	Number of Battles
gpt-4-1106-preview	7387
gpt-3.5-turbo-0613	7083
gpt-4-0613	6165
claude-2.1	5583
claude-instant-1	4136
gpt-4-0314	4122
claude-1	3978
vicuna-33b	3720
mixtral-8x7b-instruct-v0.1	3545
vicuna-13b	3448

Table 7: Number of battles each model participated in within the Arena Human Preference dataset.

B Reference Model Capability Tiers

In Section 4.2.1, we stratified reference models by their Chatbot Arena performance into four tiers to study how evaluator effectiveness varies with reference model in different tiers. Table 8 presents the models, their win rates, Arena scores, and Kendall's τ (computed against gold across all test tiers).

C Model-Level Reward Breakdown in WildBench

In Section 4.3, we further analyzed WildBench performance by breaking down how each reference model contributed to the final reward. Table 9 shows scores and reward components across 21 models.

This illustrates that model rankings can change significantly depending on the choice of reference, especially for middle- and lower-tier models.

Reference Model	Winrate (%)	Arena Score	Kendall's τ (All tiers)
Tier 1			
Meta-Llama-3-70B-Instruct	33.18	1206	0.7706
claude-3-5-sonnet-20240620	40.56	1268	0.7457
gpt-4o-mini-2024-07-18	44.65	1272	0.7309
Qwen2-72B-Instruct	29.85	1187	0.7309
gpt4_0314	22.07	1186	0.7471
Tier 2			
Yi-34B-Chat	29.66	1111	0.7949
mistral-large-2402	21.44	1157	0.7275
Mixtral-8x7B-Instruct-v0.1	18.26	1114	0.7484
mistral-medium	21.86	1148	0.7389
Tier 3			
dbrx-instruct	18.45	1103	0.7275
Qwen1.5-7B-Chat	11.77	1070	0.7485
Starling-LM-7B-alpha	14.25	1088	0.7551
llama-2-70b-chat-hf	13.89	1093	0.7383
wizardlm-70b	14.38	1106	0.7497
OpenHermes-2.5-Mistral-7B	10.34	1074	0.7508
gpt-3.5-turbo-1106	9.18	1068	0.7457
llama-2-13b-chat-hf	7.70	1063	0.7234
Tier 4			
Qwen-14B-Chat	7.50	1035	0.7247
chatglm2-6b	2.76	924	0.6728
vicuna-13b	5.83	1042	0.7167

Table 8: Reference models grouped by capability tier (Chatbot Arena), with Kendall's τ versus gold rankings computed across all test tiers

.

Model	WB Elo	Arena Score	Tier	gpt4t	haiku	llama	reward
Mistral-7B-Instruct-v0.2	1095.81	1072	Tier 4	-54.74	-19.34	3.61	-23.49
mistral-large-2402	1158.02	1157	Tier 2	-46.39	-2.98	18.43	-10.31
gemma-2b-it	996.57	990	Tier 4	-86.08	-69.63	-58.84	-71.52
gpt-3.5-turbo-0125	1124.51	1106	Tier 3	-64.84	-27.44	-4.35	-32.21
Nous-Hermes-2-Mixtral-8x7B-DPO	1080.54	1084	Tier 4	-54.74	-16.28	2.44	-22.86
gpt-4-0125-preview	1208.30	1245	Tier 1	-4.25	37.84	51.32	28.30
tulu-2-dpo-70b	1114.00	1099	Tier 3	-54.93	-16.99	3.22	-22.90
nemotron-4-340b-instruct	1184.38	1209	Tier 1	-21.04	28.86	42.58	16.80
Yi-1.5-34B-Chat	1162.79	1157	Tier 2	-17.29	24.76	40.57	16.01
gpt-4o-2024-05-13	1239.47	1285	Tier 1	1.66	43.26	52.54	32.49
Qwen2-72B-Instruct	1176.37	1187	Tier 2	-34.08	13.04	31.84	3.60
Mixtral-8x7B-Instruct-v0.1	1122.78	1114	Tier 3	-50.15	-11.62	9.67	-17.37
gemma-7b-it	1058.18	1037	Tier 4	-77.25	-52.44	-35.87	-55.19
Starling-LM-7B-beta	1119.33	1119	Tier 3	-47.02	-3.22	15.87	-11.46
deepseek-coder-v2	1183.53	1178	Tier 2	-26.44	20.21	36.43	10.07
claude-3-5-sonnet-20240620	1226.31	1268	Tier 1	-4.39	40.92	50.15	28.89
gemini-1.5-flash	1196.80	1227	Tier 1	-11.28	27.15	40.09	18.65
gemini-1.5-pro	1221.13	1260	Tier 1	-1.32	39.06	47.95	28.56
Llama-2-7b-chat-hf	1033.37	1037	Tier 4	-66.60	-39.05	-25.20	-43.62
Meta-Llama-3-70B-Instruct	1194.65	1206	Tier 2	-18.43	30.06	45.80	19.14
Meta-Llama-3-8B-Instruct	1134.24	1152	Tier 3	-46.34	-7.28	14.45	-13.05

Table 9: List of 21 models shared between WildBench and Chatbot Arena leaderboards, including their WildBench Elo, Arena score, tier, and reward contributions from individual reference models (gpt4t, haiku, llama).