PeerRank: Autonomous LLM Evaluation Through Web-Grounded, Bias-Controlled Peer Review

Yanki Margalit, Erni Avram, Ran Taig, ChatGPT5.2,
Claude Opus 4.5, Gemini-3-Pro, Oded Margalit, Nurit Cohen-Inger
Caura.ai

Abstract

Evaluating large language models typically relies on human-authored benchmarks, reference answers, and human or single-model judgments—approaches that scale poorly, become quickly outdated, and mismatch open-world deployments that depend on web retrieval and synthesis. We introduce PeerRank, a fully autonomous end-to-end evaluation framework in which models generate evaluation tasks, answer them with live *web grounding*, judge peer responses *without* web access using a standardized rubric, and aggregate dense peer assessments into relative performance estimates, without human supervision or gold references. PeerRank treats evaluation as a multi-agent process where each model participates symmetrically as task designer, respondent, and evaluator, while controlled shuffling and blinding isolate systematic judge effects. In a large-scale study over 12 commercially available models and 420 autonomously generated questions spanning five categories, PeerRank produces stable, discriminative rankings and reveals measurable identity and presentation biases when controls are removed. Rankings are robust across aggregation methods, with strong agreement between mean peer scores and Elo ratings from induced pairwise comparisons. We further validate PeerRank on TruthfulQA and GSM8K, where peer scores correlate with objective accuracy. Together, these results suggest that bias-aware peer evaluation with web-grounded answering can scale open-world LLM assessment beyond benchmark-centric methods.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in domains where correctness, robustness, and up-to-date knowledge are critical. However, evaluation methodologies have lagged behind model capabilities. Most existing evaluations still rely on static benchmarks with human-authored tasks and reference answers, which are costly to maintain, prone to contamination, and quickly become outdated [13, 29]. These benchmarks often assume a closed-world setting, even as real deployments increasingly involve web access and tool use.

This study asks whether LLMs can be evaluated *by their peers*, building on LLM-as-a-judge and preference-based evaluation paradigms [4, 16, 34], without external supervision, while still yielding meaningful and interpretable performance estimates at scale.

We introduce PeerRank, a fully autonomous, multi-agent evaluation framework in which models generate evaluation tasks, answer them with live web grounding, evaluate peer responses, and aggregate results into rankings and bias measurements, without any human supervision or oracle reference answers. Closest to our framing, peer-review-inspired evaluators such as PRE [5] and other peer-based, label-free approaches [27] reduce reliance on human judgment; PeerRank extends this line by making the task distribution fully autonomous and by explicitly isolating judge biases. Related multi-agent work studies self-evolving task and benchmark generation for dynamic evaluation [28]. In our largest autonomous run, PeerRank evaluates 12 commercially available models on 420 model generated questions spanning five categories, and performs peer judging under three controlled regimes (shuffle-only, blind-only, and shuffle+blind) to quantify self bias, identity (name) bias, and position bias, consistent with known sensitivities of LLM-as-a-judge evaluators [25, 26, 33]. Figure 1 provides an overview of the end-to-end system.

Contributions

This paper makes the following contributions:

1. **Fully endogenous peer evaluation.** We introduce PeerRank, an end-to-end framework in which LLMs autonomously generate evaluation tasks, answer them, judge peers, and aggregate

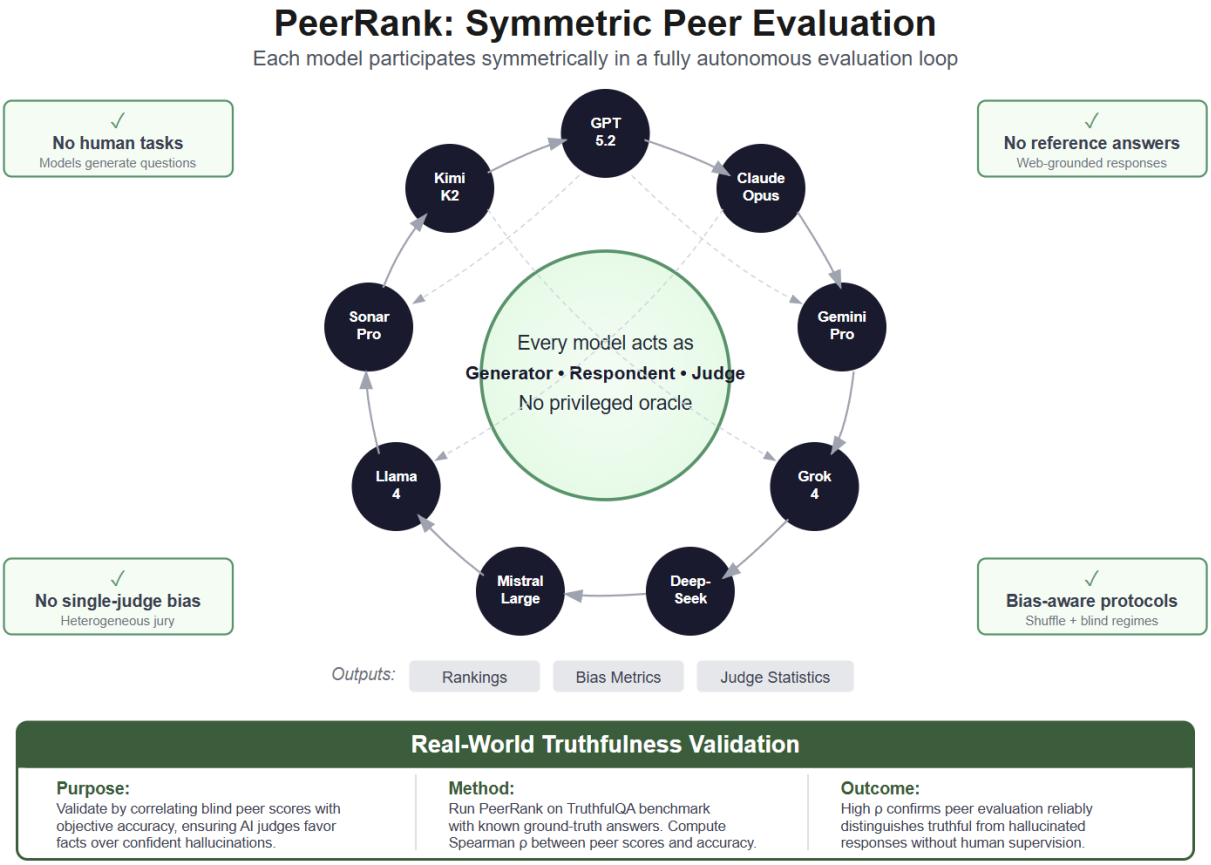

Figure 1: Fully endogenous PeerRank evaluation pipeline (no human input). Models generate questions, answer with web grounding, evaluate peers under bias-control protocols, and aggregate scores into rankings and bias metrics. We additionally validate real-world truthfulness by running PeerRank on TruthfulQA [14] and correlating peer scores with objective accuracy.

results, without human-authored prompts, reference answers, or human judgments.

2. **Open-world, web-grounded testing.** PeerRank enables live web access during answering while disabling it during judging so evaluators score only the submitted answer, not additional evidence retrieved at scoring time, keeping judgments blind and comparable across models.

3. **Bias-aware protocols.** We control and measure self bias, identity (name) bias, and position bias via systematic shuffling and blinding, and report these effects alongside final rankings.

4. **Empirical validation at scale.** In a study of 12 models and 420 autonomously generated questions, we show that peer-based aggregation yields stable and discriminative rankings.

2 Related Work

PeerRank relates to several lines of research on evaluating LLMs, including benchmark-centric evaluation, open-world and agent-based evaluation, LLM-as-a-judge approaches, and studies of bias and reliability in automated evaluation.

Benchmark-Centric and Holistic Evaluation Traditional LLM evaluation has relied on static benchmarks with human-authored tasks and reference answers, such as question answering datasets [6, 11, 23] and exam-style evaluations. While these benchmarks provide comparability and repeatability, they suffer from several limitations: they become stale over time [7, 20], are vulnerable to contamination through repeated exposure [2, 3], and often fail to capture real-world usage where tasks and facts evolve [24, 32].

Holistic evaluation suites, such as HELM [13], aim to address some of these limitations by broadening

task coverage and reporting multiple metrics under standardized prompts. However, these approaches still depend on human-written tasks and human-anchored evaluation criteria, and they largely assume closed-world answering without external knowledge access.

Open-World and Agent-Based Evaluation As LLMs increasingly operate as agents with tools and web access, evaluation has expanded toward open-world and interactive settings. AgentBench [15] and WebArena [35] evaluate planning, tool use, and task completion in realistic environments, emphasizing agentic behavior rather than static question answering. These benchmarks improve ecological validity but often sacrifice standardization and make it difficult to disentangle model capability from environment-specific design choices.

PeerRank adopts a complementary approach: it evaluates models on natural-language questions answered under live web grounding, while preserving a controlled and repeatable evaluation protocol through standardized peer judging.

LLM-as-a-Judge and Preference-Based Ranking Using LLMs as judges has emerged as a scalable alternative to human evaluation for open-ended generation. Frameworks such as MT-Bench [34] demonstrate that strong LLM judges can correlate well with human preferences, while preference platforms such as Chatbot Arena [4] aggregate large numbers of comparisons into relative rankings via Elo-style methods [9].

PeerRank builds on this paradigm but departs from prior work in two key ways. First, it distributes judging across multiple heterogeneous models rather than privileging a single judge. Second, it removes dependence on human-authored prompts, gold answers, or baseline models, making the evaluation distribution itself fully endogenous.

Bias and Reliability of Automated Evaluation Recent work has identified systematic biases in LLM-based evaluation, including position bias [26], verbosity bias, self-preference [30], and non-transitive preferences. These effects can significantly distort rankings when a single judge or uncontrolled protocol is treated as an oracle. Work such as [25] studies the reliability and alignment of LLM evaluators, and [12] establishes static benchmarks for judge correctness, while [33] explicitly quantifies biases in LLM-as-a-judge settings. Other approaches propose debiasing or calibration strategies (e.g., by controlling presentation effects or post-hoc score adjustment) and emphasize the need to treat judge behavior as a first-class measurement object.

Table 1: Comparison of large language model evaluation paradigms. △ indicates partial or implicit support. PeerRank is the only framework that is fully endogenous, peer-based, web-grounded at answer time, and explicitly bias-aware.

Approach	Human Tasks	Human Judging	Web-Grounded	Peer-Based	Bias-Aware
Static Benchmarks (QA, Exams)	✓	✓	×	×	×
Holistic Suites (HELM)	✓	✓	×	×	△
Agent Benchmarks (AgentBench, WebArena)	✓	×	✓	×	×
LLM-as-a-Judge (MT-Bench)	✓	×	×	△	△
Preference Platforms (Chatbot Arena)	✓	✓	×	✓	△
PeerRank (this work)	×	×	✓	✓	✓

PeerRank treats bias not merely as a nuisance to mitigate but as a measurable phenomenon. By running multiple controlled evaluation regimes (shuffle-only, blind-only, shuffle+blind), PeerRank explicitly isolates and quantifies self bias, identity (name) bias, and position bias, and reports these quantities alongside final rankings.

Summary and Positioning Table 1 summarizes how PeerRank differs from prior evaluation paradigms along key dimensions.

3 Methodology

PeerRank evaluates LLMs through a fully autonomous, multi-agent pipeline in which models generate evaluation tasks, answer them under web grounding, evaluate peer responses, and aggregate results into rankings and bias measurements. Figure 2 illustrates the end-to-end process.

Setup and Notation Let $\mathcal{M} = \{m_1, \ldots, m_K\}$ denote the evaluated models. Each model participates symmetrically as a question generator, respondent, and evaluator. Let n be the number of questions generated per model, yielding a total evaluation set $\mathcal{Q}$ with $|\mathcal{Q}| = Kn$.

For a question $q \in \mathcal{Q}$ and model $m_j \in \mathcal{M}$, let $a_{j,q}$ denote the generated answer, and let $s_{i,j,q} \in \{1, \ldots, 10\}$ be the score assigned by evaluator m_i to m_j's answer to q.

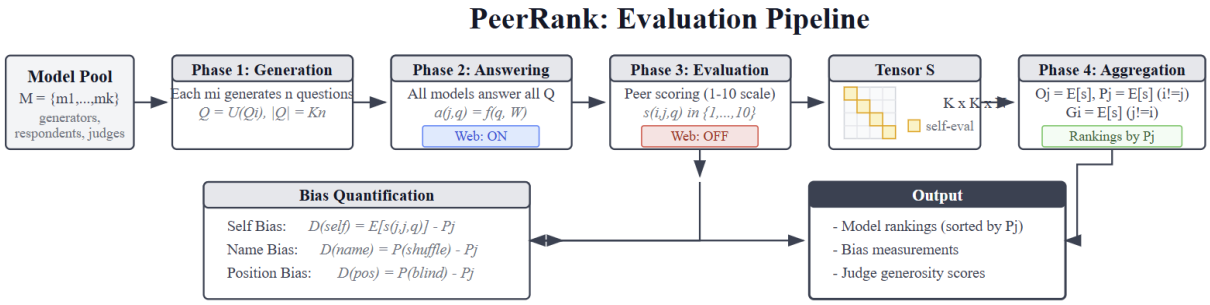

Figure 2: PeerRank evaluation pipeline. Models act symmetrically as question generators, respondents, and judges. Answers are generated with web access enabled, while evaluation is performed without web access with bias quantification. Scores are aggregated into peer rankings, bias measurements, and judge statistics.

3.1 Phase 1: Endogenous Question Generation

Each model m_i independently generates a set

$$\mathcal{Q}_i = \{q_{i,1}, \ldots, q_{i,n}\},$$

drawing from a fixed category set. The full evaluation corpus is

$$\mathcal{Q} = \bigcup_{i=1}^{K} \mathcal{Q}_i.$$

No filtering or deduplication is applied, ensuring that the task distribution is defined endogenously by participating models rather than human curation.

Home-question advantage. A natural concern is that models may perform better on questions they authored. We test this by comparing each model's mean peer score on its own questions versus questions authored by other models, excluding self-evaluations. We report per-model differences (own minus other) with effect sizes and statistical significance (Appendix A, Table A1).

3.2 Phase 2: Web-Grounded Answer Generation

All models answer all questions. For each (m_j, q) pair, the answer is generated as

$$a_{j,q} = f_{m_j}(q, \mathcal{W}),$$

where $\mathcal{W}$ denotes an external web environment accessible via search tools. Web access is enabled in this phase to evaluate retrieval and synthesis under open-world conditions. Wall-clock response times are recorded but are not used during scoring.

Web grounding is implemented using provider-native search tools when available, and otherwise via external retrieval result injection; Table 2 summarizes the mechanism used for each model family.

3.3 Phase 3: Peer Evaluation

Each model evaluates all answers to each question, producing scores

$$s_{i,j,q} \in \{1, \ldots, 10\}.$$

Judging is performed without web access, using a standardized rubric emphasizing correctness, completeness, clarity, and usefulness. Web grounding is disabled during evaluation so judges score only the submitted answer, not extra evidence retrieved at scoring time. This keeps judging blind and comparable across models. The resulting evaluation tensor

$$\mathbf{S} = \{s_{i,j,q}\}$$

has dimensions $K \times K \times N$, with diagonal entries corresponding to self-evaluations.

To control systematic judge effects, evaluations are conducted under three regimes: (i) shuffle-only (randomized answer order, visible identities), (ii) blind-only (fixed order, hidden identities), and (iii) shuffle+blind (randomized order, hidden identities), motivated by known order/position effects in LLM judging [26]. The shuffle+blind regime serves as the least-confounded baseline for final rankings.

3.4 Phase 4: Score Aggregation and Metrics

The observed score of model m_j is

$$O_j = \mathbb{E}_{i,q}[s_{i,j,q}],$$

which includes self-evaluations. The primary ranking metric, the peer score, excludes self-ratings:

$$P_j = \mathbb{E}_{i \neq j,q}[s_{i,j,q}].$$

Judge generosity for evaluator m_i is defined as

$$G_i = \mathbb{E}_{j \neq i,q}[s_{i,j,q}],$$

capturing systematic differences in scoring strictness.

Robustness: Elo-based aggregation. To evaluate robustness, we complemented the primary mean-score aggregation with an Elo rating approach [9] based on pairwise model comparisons. We converted evaluation scores into pairwise outcomes and estimated Elo ratings with $K - factor = 32$ over $N = 253{,}080$ matches, producing a full ranking of all candidates. The resulting ordering closely matches the mean-score ranking, with very high agreement in both linear association (Pearson $r = 0.912$) and rank consistency (Spearman $\rho = 0.888$). This convergence indicates that our findings are stable across aggregation schemes and are not an artifact of absolute scoring, as the pairwise Elo formulation is less sensitive to scale differences and noise in raw ratings.

3.5 Bias Quantification

PeerRank explicitly measures evaluation bias by comparing scores across regimes. Self bias is defined as

$$\Delta_j^{\text{self}} = \mathbb{E}_q[s_{j,j,q}] - P_j.$$

Name (identity) bias is measured as

$$\Delta_j^{\text{name}} = P_j^{\text{shuffle}} - P_j,$$

and position bias as

$$\Delta_j^{\text{pos}} = P_j^{\text{blind}} - P_j,$$

where P_j corresponds to the shuffle+blind baseline. Positive values indicate score inflation due to the corresponding factor.

3.6 Ranking Stability

Final rankings are obtained by sorting models in descending order of P_j. To characterize agreement among judges, we report the inter-judge variance

$$\sigma_j^2 = \text{Var}_{i \neq j,q}(s_{i,j,q}),$$

which measures the stability of peer assessments.

Together, these outputs yield bias-aware rankings, judge behavior statistics, and uncertainty estimates under a fully endogenous, open-world evaluation protocol.

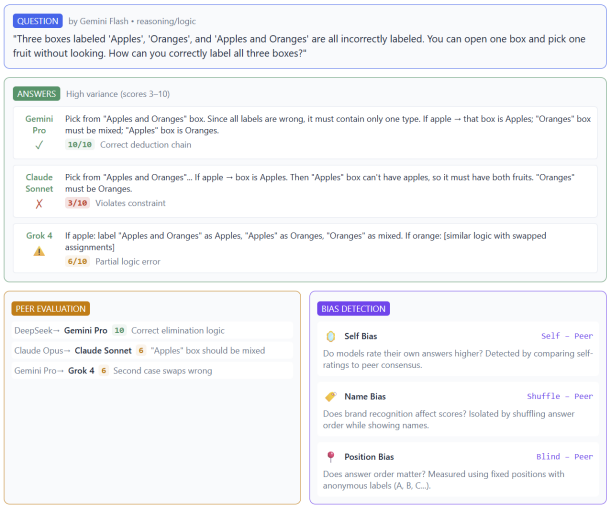

Figure 3: Sample question and its evaluation

External validity check. Because peer-based evaluation is inherently relative, we additionally test whether PeerRank peer scores track objective correctness on a benchmark with known answers (TruthfulQA). We report this validation as a result in Section 5.

4 Experimental Setup

We evaluate PeerRank in a fully autonomous, multi-model study designed to test whether peer-based evaluation yields stable and discriminative rankings under realistic, open-world answering conditions, while explicitly isolating identity and presentation biases.

Models. We run PeerRank on $K = 12$ commercially available LLMs accessed through public APIs [1, 8, 10, 17–19, 21, 22, 31]: `gpt-5.2`, `gpt-5-mini`, `claude-opus-4-5`, `claude-sonnet-4-5`, `gemini-3-pro-preview`, `gemini-3-flash-preview`, `grok-4-1-fast`, `deepseek-chat`, `llama-4-maverick`, `sonar-pro`, `kimi-k2-0905`, `mistral-large`.

All models participate symmetrically in all roles: question generation, answering, and judging.

Question generation. Each model generates $n = 35$ questions under a fixed category set: ... resulting in a combined evaluation corpus of

$$N = Kn = 12 \cdot 35 = 420$$

questions. Questions are used as-generated without filtering, deduplication, or human editing, ensuring

that the task distribution is defined endogenously by the participating models.

Datasets and task sources

PeerRank evaluates models on two complementary task sources:

Endogenous open-world task set (primary). The main study uses the fully endogenous question set $\mathcal{Q}$ generated by the participating models (Section 3), spanning five categories: `factual knowledge`; `reasoning / logic`; `current events`; `creative / open-ended`; `practical how-to`. This distribution is intentionally *not* human-curated: we apply no filtering, deduplication, or difficulty balancing, so the task population reflects the models' own priors and capabilities as question writers. The goal is to measure relative performance under a continuously refreshable, benchmark-free regime aligned with open-world deployment.

External benchmarks with ground truth. To test whether blind peer scores track objective correctness, we run PeerRank on two external benchmarks with ground truth: TruthfulQA [14], which provides multiple-choice questions with an answer key, and GSM8K, which provides open-ended math problems with exact-match numeric answers. For TruthfulQA, we sample 264 questions from the validation split (Section 5) and evaluate (i) ground-truth accuracy and (ii) peer scores under the same shuffle+blind judging protocol, providing an external validity check for peer judgment as a proxy for factual correctness. For GSM8K, we evaluate a comprehensive set of 611 questions comprising all *medium* and *hard* test questions, score each model by exact-match accuracy mapped to a 0–10 ground-truth score, and correlate this signal with shuffle+blind peer scores (Section 5).

Web-grounded answering. For every question $q \in \mathcal{Q}$, each model produces an answer with web access enabled. Models are instructed to answer directly and concisely and to avoid preambles (Appendix A). We record wall-clock response time per answer, enabling analysis of quality, latency trade-offs under web-grounded generation. Across the full run, essentially all answer calls succeeded (nearly all models answered nearly all questions, depending on provider reliability).

Web-grounding implementation by provider. While all models are evaluated under the same instruction to use web access during answering, the underlying grounding mechanism differs by provider API and integration. In our implementation, OpenAI, Anthropic, and Gemini models use provider-native web/search tools; `deepseek-chat`, `llama-4-maverick`, and `kimi-k2-0905` use external search injection (Tavily) with retrieved snippets provided as hidden context; and `sonar-pro` uses a native provider-side retrieval mechanism. Importantly, web access is enabled *only* during answering and is disabled during all judging.

Table 2: Web-grounding mechanisms used in Phase 2 (answer generation). Web access is enabled only during answering and disabled during all judging.

Mode	Mechanism	Models
Tool	Provider-native web/search tool	`gpt-5.2`, `gpt-5-mini`, `claude-opus-4-5`, `claude-sonnet-4-5`, `gemini-3-pro-preview`, `gemini-3-flash-preview`, `grok-4-1-fast`
Tavily	External search injection	`deepseek-chat`, `llama-4-maverick`, `kimi-k2-0905`
Native	Provider-side retrieval	`sonar-pro`
Agents	Tool orchestration wrapper	`mistral-large`

Peer judging and bias calculation. All answers are scored by all models acting as judges, using a shared 1–10 rubric and a short written reason (8–20 words; Appendix A). Judging is performed without web access. To isolate systematic evaluator biases, we run three controlled protocols: (i) *shuffle-only* (randomized answer order, model identities visible), (ii) *blind-only* (fixed answer order, model identities hidden), and (iii) *shuffle+blind* (randomized order, identities hidden). Unless otherwise stated, we report final rankings using the shuffle+blind regime as the least-confounded baseline.

Evaluation volume and runtime. Phase runtimes for this run were approximately 1m 40.7s for Phase 1 (question generation), 41m 3.4s for Phase 2 (web-grounded answering), and 424m 19.1s for Phase 3 (peer evaluation). Phase 3 was split across shuffle-only (110m 40.6s), blind-only (156m 58.0s), and shuffle+blind (156m 38.9s).

Aggregation and reporting. We compute each model's peer score

$$P_j = \mathbb{E}_{i \neq j, q}[s_{i,j,q}],$$

defined as the mean of all peer-assigned scores excluding self-ratings. We additionally report judge generosity

$$G_i = \mathbb{E}_{j \neq i, q}[s_{i,j,q}],$$

capturing evaluator strictness differences, and bias quantities derived from cross-regime comparisons: self bias (self minus peer), name bias (shuffle-only minus shuffle+blind), and position bias (blind-only minus shuffle+blind), as defined in Section 3. Uncertainty is summarized using the standard deviation across peer judgments for each evaluated model.

5 Results

We report results from the fully autonomous PeerRank evaluation under the most bias-controlled regime (shuffle+blind), unless otherwise stated. This section summarizes (i) aggregate peer rankings, (ii) cross-model judgment structure, and (iii) quality-latency trade-offs under web grounding.

Peer-Based Model Rankings Figure 4 shows final PeerRank scores for all evaluated models. Each bar reports the mean peer score P_j, averaged over peer-assigned ratings while excluding self-evaluations. Error bars denote one standard deviation across all $(i \neq j, q)$ judgments.

Peer scores are integer 1–10 rubric ratings (Appendix A) averaged as $P_j = \mathbb{E}_{i \neq j, q}[s_{i,j,q}]$.

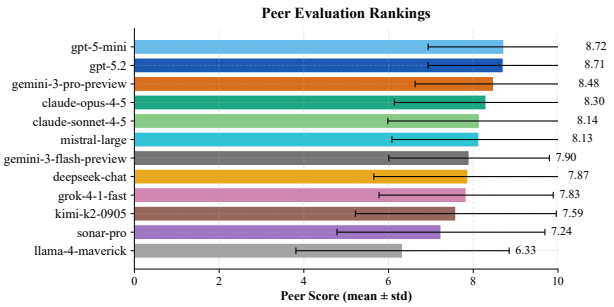

Figure 4: Peer rankings (shuffle+blind): mean peer scores ±1 SD; self-ratings excluded.

The ranking shows a tight top tier (with the top two models nearly tied) and clearer separation toward the lower tier, with non-trivial variance. This suggests the peer evaluation protocol is neither degenerate nor overly compressed. Models from different providers and architectural families interleave, suggesting PeerRank does not simply reflect shared lineage or training origin. **Provider clustering.** Peer scores nevertheless differ significantly by provider (Kruskal–Wallis $H(8) = 4135.33$, $p < 0.001$; $\eta^2 = 0.082$), indicating a modest provider-level effect.

Cross-Evaluation Structure and Self-Preference Aggregate rankings obscure individual judgment structure. Figure 5 visualizes the full cross-evaluation matrix, where rows correspond to evaluator models and columns to evaluated models.

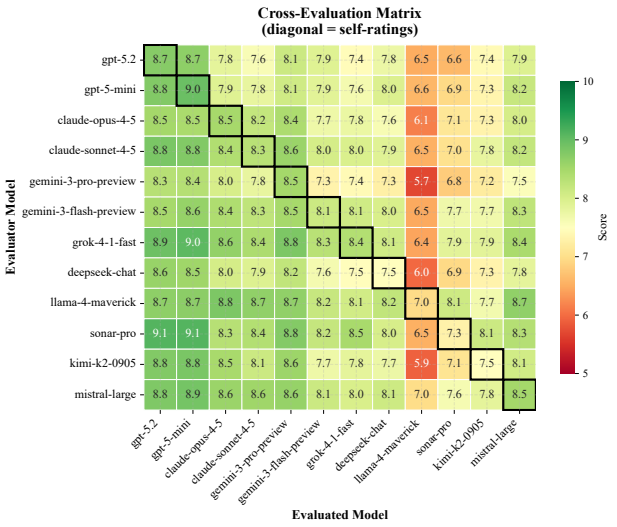

Figure 5: Cross-evaluation matrix of peer scores. Rows denote evaluator models and columns denote evaluated models. Diagonal entries correspond to self-ratings and are outlined for emphasis.

Diagonal entries are *often* higher than off-diagonal scores, indicating a common tendency toward self-preference [30]. However, this pattern is not universal: some models exhibit near-zero self-preference, and a small subset shows *negative* self-preference, rating their own answers below the scores they receive from peers. Beyond the diagonal, the matrix reveals systematic evaluator asymmetries: certain models consistently score others more generously, while others apply markedly stricter standards. These effects motivate reporting peer score, self bias, and judge generosity as separate measurements rather than relying on raw averages alone.

Despite these asymmetries, aggregation across heterogeneous judges yields stable rankings, indicating that no single evaluator dominates the signal.

Task Difficulty and Judge Disagreement To diagnose sources of model performance limits, we

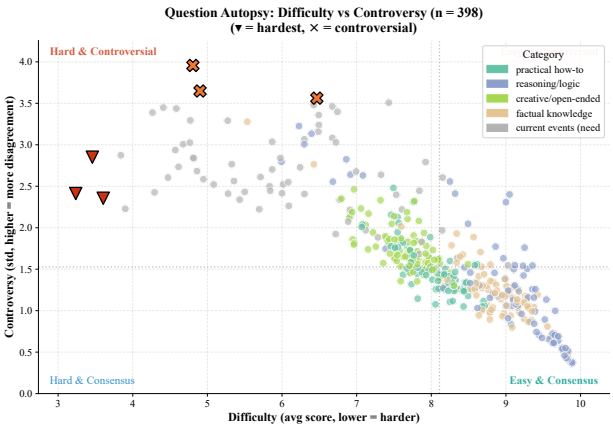

Figure 6: **Question Autopsy: Difficulty vs. Controversy ($n = 398$); questions with missing evaluations were excluded. X-axis:** Mean peer score (lower = harder). **Y-axis:** SD of peer scores across judges (higher = more disagreement). *Current Events* concentrate in the hard/controversial quadrant, while *Reasoning* and *Factual Knowledge* cluster in easy/-consensus, suggesting saturation on static tasks.

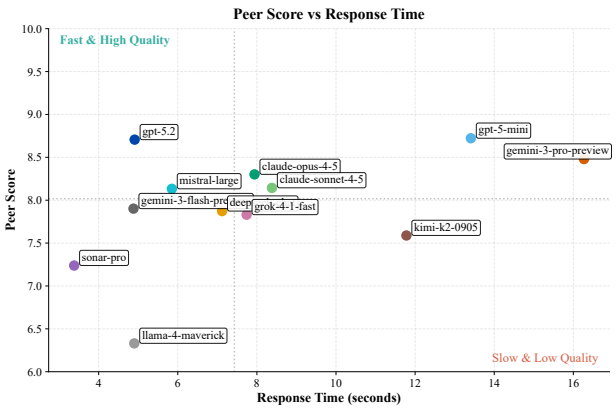

Figure 7: Quality versus speed trade-off under web grounding.

analyze the relationship between question difficulty and judge controversy across categories. We define difficulty as the mean peer score per question across non-self judgments (lower = harder for this cohort), while controversy is the standard deviation of scores (higher = more disagreement). Figure 6 reveals a clear category-based separation:

- **Retrieval drives divergence:** *Current Events* questions (grey) dominate the "Hard & Controversial" quadrant, suggesting live retrieval is the main differentiator; disagreement likely reflects conflicting retrieval or hallucinations, producing higher score variance.
- **Static reasoning saturation:** *Reasoning/Logic* and *Factual Knowledge* cluster in the "Easy & Consensus" region, suggesting static logic tasks are near-saturated for this cohort and motivating more dynamic, open-world evaluation.

Home-question advantage (results). Across models, the mean home advantage is small (+0.058 points on average; mean Cohen's $d \approx 0.02$). Home-question advantage is small on average (+0.005). Notable exceptions include a positive effect for gpt-5.2 (+0.55) and a large negative effect for kimi-k2-0905 (-1.01), with grok-4 also negative (-0.37). while seven show no significant difference (Appendix A, Table A1).

Quality-Speed Trade-offs Under Web Grounding Figure 7 plots each model's mean peer score against its average wall-clock response time during web-grounded answer generation.

Quality and latency exhibit only a weak correlation. Some models achieve above-average quality with low response times, while others trade substantial latency for modest quality gains. This decoupling suggests that inference speed is not a reliable proxy for answer quality in open-world, web-grounded settings.

Overall, PeerRank surfaces multi-dimensional performance characteristics—including quality, judgment structure, and efficiency—using a fully endogenous evaluation pipeline without human supervision.

Truthfulness validation on TruthfulQA

Peer evaluation is relative by construction: a cohort could, in principle, converge on preferences that do not track factual correctness. We therefore test whether PeerRank peer scores align with an external ground-truth signal using TruthfulQA [14], which provides multiple-choice questions with known correct options.

Protocol summary (for interpretability). We sample 264 questions from the TruthfulQA validation split and have each model answer every question with deterministic decoding (temperature = 0), treating temperature as a controlled system parameter to remove sampling variance and improve reproducibility. Each model outputs (i) a single choice letter and (ii) a 2–3 sentence justification. Judges then score these answers *without* web access using the same 1–10 rubric as in the main study, under the shuffle+blind regime (randomized answer order, identities hidden). Ground-truth accuracy A_j is computed by matching

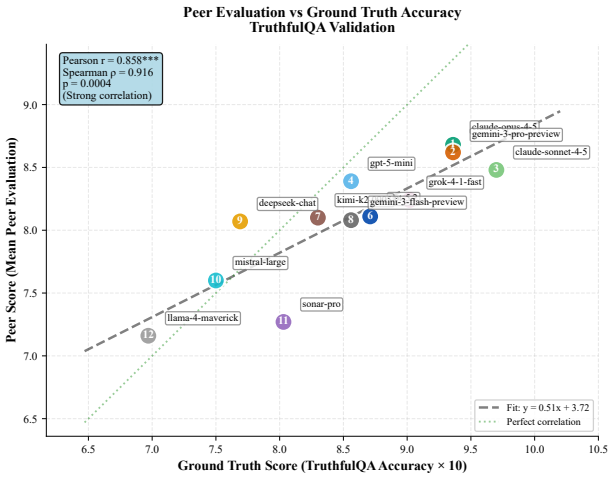

Figure 8: **PeerRank vs. TruthfulQA truth score.** Peer score (y; shuffle+blind mean) versus TruthfulQA truth score (x; 10× accuracy).

the predicted letter to the benchmark key, and we define a comparable 0–10 truth score $T_j = 10A_j$.

Peer scores track objective correctness. Across the 12-model cohort, peer scores on TruthfulQA align strongly with ground truth: Pearson $r = 0.858$ ($p = 0.0004$) and Spearman $\rho = 0.916$ ($p < 10^{-4}$). Figure 8 shows the near-linear trend; Figure 9 shows per-model deviations.

Ablation: correlation degrades without bias controls. Repeating the analysis without shuffling/blinding reduces correlation with accuracy and can remove significance (Table 3).

Metric	Bias-corrected	No correction	Δ (C–U)
Pearson r	0.641	0.573	+0.070
Spearman ρ	0.579	0.491	+0.088
p (Pearson)	0.025*	0.052	loses signif.
p (Spearman)	0.049*	0.105	loses signif.

Table 3: Ablation on TruthfulQA: removing bias controls reduces correlation with ground truth and can remove statistical significance.

Where peer judgment over- or under-ranks models. The remaining gaps are driven by rank mismatches from position/identity effects and secondary cues (e.g., explanation quality or judge calibration); Figure 10 highlights these deviations.

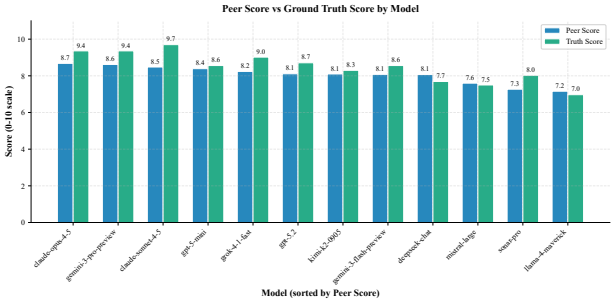

Figure 9: **Peer score vs. truth score by model (0–10).** Bars compare PeerRank peer score to 10× accuracy on TruthfulQA.

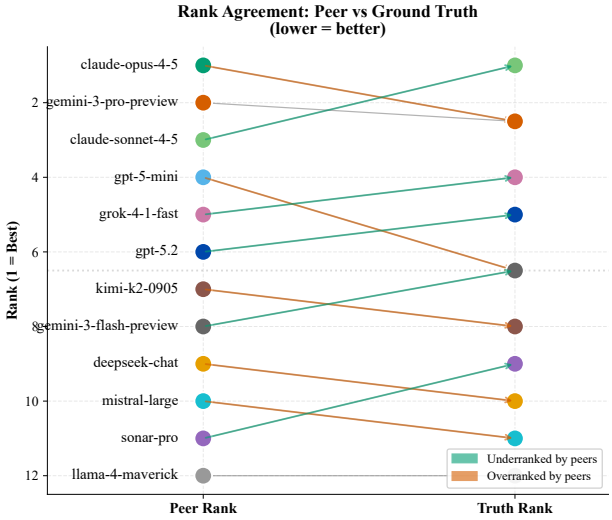

Figure 10: **Rank agreement between PeerRank and TruthfulQA.**

Truthfulness validation on GSM8K

We further test whether peer scores track objective correctness on a structured, exact-match benchmark by running PeerRank on GSM8K (math reasoning). Each model answers GSM8K questions, we compute accuracy against the gold numeric answers, and rescale accuracy to a 0–10 ground-truth score for comparison with peer scores under shuffle+blind judging.

We evaluated GSM8K using a comprehensive test of 611 questions comprising all items from the *medium* and *hard* categories. Each model answered all questions, we computed exact-match accuracy against the gold numeric answers, and rescaled accuracy to a 0–10 ground-truth score (10× accuracy) for comparison with shuffle+blind peer scores. Across the 12-model cohort, peer scores strongly correlate with GSM8K ground-truth performance: Pearson correlation between peer score and ground-truth score is

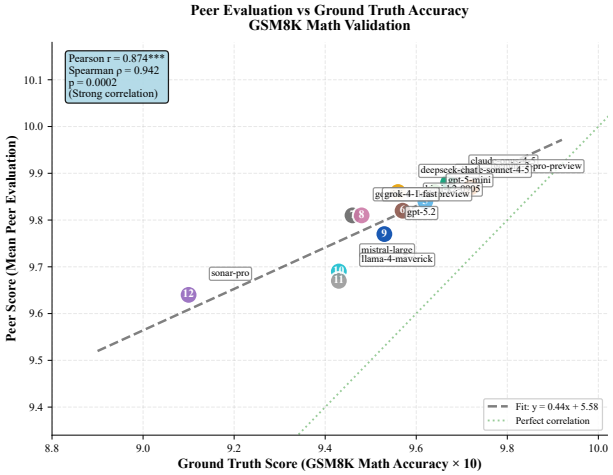

Figure 11: **PeerRank vs. GSM8K ground-truth score.** Peer score (y; shuffle+blind mean) versus GSM8K ground-truth score (x; 10× accuracy).

$r = 0.874$ ($p = 0.0002$), and rank correlation is Spearman $\rho = 0.942$.

Peer scores track objective math correctness. Across the 12-model cohort, peer scores correlate with GSM8K ground-truth performance (Figure 11). This extends the external validity check beyond factual multiple-choice QA to open-ended math problems with exact-match scoring.

Ceiling-effect caveat (compressed variance). GSM8K accuracy in this cohort is high (top-end clustering), which compresses the ground-truth range and can attenuate linear correlation. Despite this ceiling effect, the rank association remains strong, indicating that peer judgment still preserves meaningful ordering even when absolute differences are small.

6 Discussion

We interpret PeerRank as a peer-based alternative to fixed-reference evaluation, emphasizing measurable judge effects.

Bias as a First-Class Measurement Object A central finding of this work is that bias in LLM-based evaluation is structural rather than incidental. Figures 12–14 summarize measured self, name, and position biases across models.

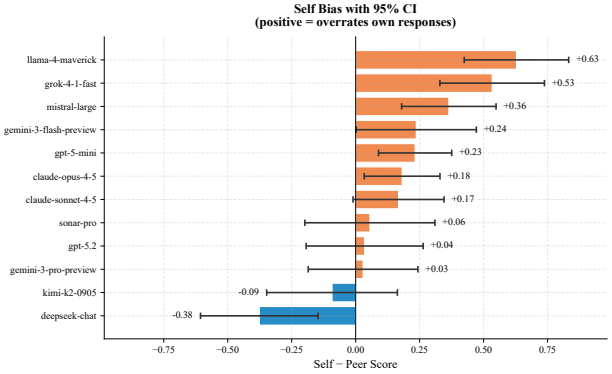

Figure 12: Self bias (self minus peer). Positive indicates self-overrating.

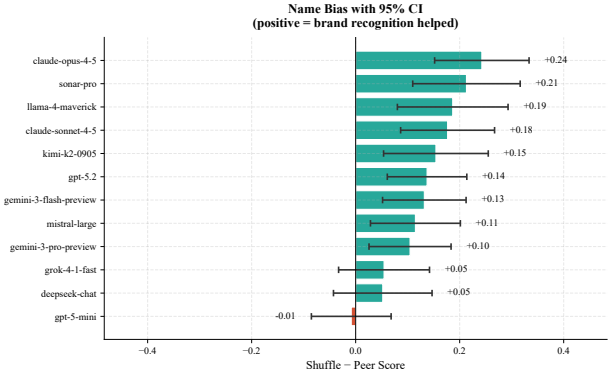

Figure 13: Name bias (visible identity effect). Positive indicates score inflation.

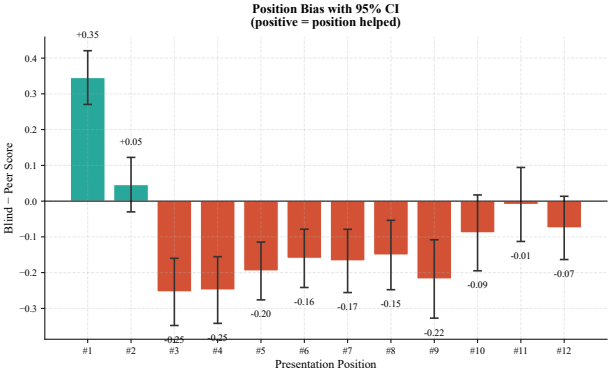

Figure 14: Position bias (answer order effect). Positive indicates position inflation.

Self bias is *typically* positive in our study, indicating that most models exhibit self-overrating when evaluating their own answers, consistent with documented self-preference effects in LLM-based judging [30]. Importantly, this behavior is heterogeneous rather than

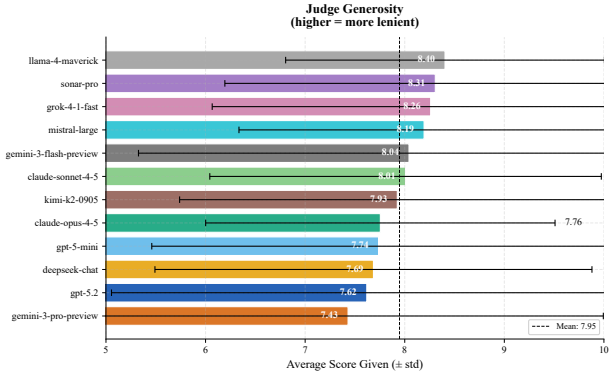

Figure 15: Judge generosity (strictness): mean score each model assigns to peers.

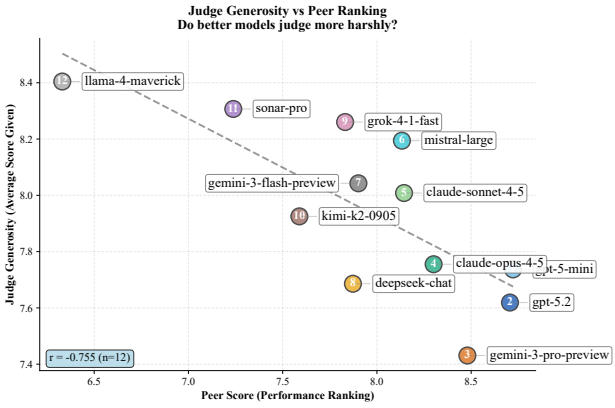

Figure 16: Generosity vs. peer score: stronger models tend to judge more harshly (negative correlation).

universal: a small subset of models shows neutral-to-negative self bias, meaning they rate their own responses at or below peer-assigned levels. Name bias is also non-trivial, with more recognizable model identities receiving higher scores when visible. Position bias is measurable: answers shown first receive a +0.35 score lift on average, while later positions (e.g., position 9) are penalized by -0.22, motivating shuffling as a control. Left uncontrolled, these effects alter rankings, motivating the shuffle+blind regime as a baseline rather than a cosmetic adjustment. By reporting bias quantities alongside final rankings, PeerRank treats evaluation bias as an explicit measurement target rather than a hidden confounder.

Judge Generosity and Heterogeneous Evaluation Behavior We characterize evaluator behavior in two ways: (i) *judge agreement*—how similarly models score the same answers—and (ii) *judge generosity*—the mean score a model assigns to others when acting as an evaluator.

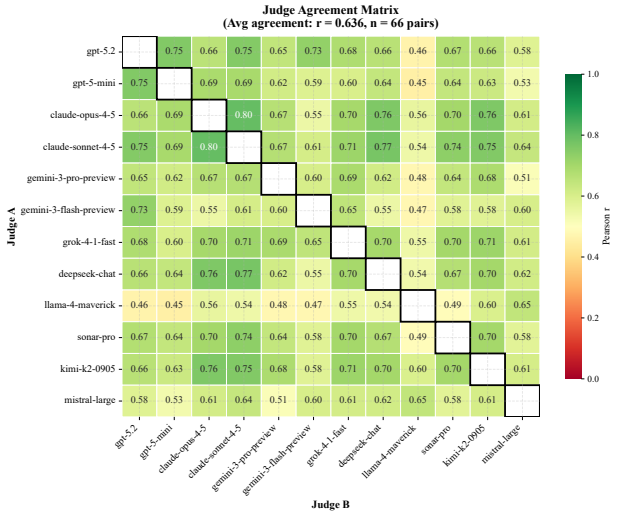

Figure 17: Judge agreement matrix: pairwise Pearson correlations between judges' scoring patterns.

Figure 16 shows that judge generosity is negatively correlated with peer performance across models (Pearson $r = -0.755$, $n = 12$), indicating that higher-ranked models tend to assign lower scores on average. This pattern suggests that evaluator strictness is not merely noise: stronger models may apply tighter standards for correctness, completeness, or rigor when judging peers. At the same time, the correlation is far from perfect, implying that judging style remains partially independent from answer quality—supporting PeerRank 's design choice to aggregate across many heterogeneous judges rather than relying on a single evaluator model.

Judge generosity varies substantially across models, reflecting differences in calibration and evaluative style. Consistent with the negative generosity–performance correlation (Figure 16, right), higher-performing models tend to judge more harshly on average, though individual deviations remain substantial. This heterogeneity highlights a limitation of single-judge evaluation paradigms and motivates aggregation across multiple, diverse judges as employed in PeerRank.

Judge agreement. To measure how consistently models apply the rubric when scoring peers, we compute pairwise Pearson correlations between judges' score vectors across the full evaluated answer set (Figure 17). Average agreement is moderate ($\bar{r} = 0.636$, $n = 66$ judge pairs), with substantial variation across judge pairs, motivating aggregation across a diverse judge pool.

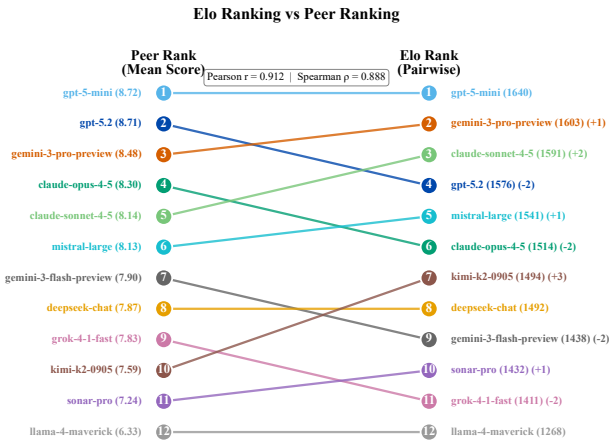

Figure 18: **Elo vs. peer ranking.** Elo ratings computed from pairwise outcomes closely match the mean-score leaderboard (see correlations on plot).

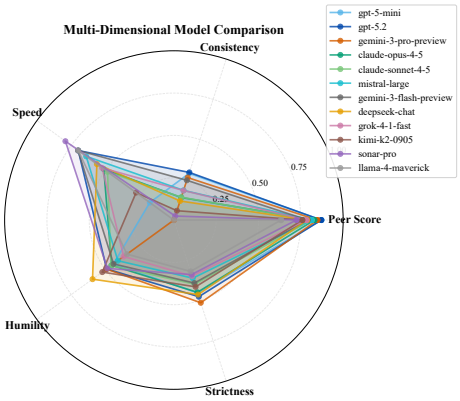

Figure 19: Multi-dimensional comparison of models across normalized evaluation dimensions, including quality, speed, consistency, humility, and strictness. No single model dominates all dimensions, highlighting trade-offs hidden by scalar rankings.

Robustness: Elo vs. mean peer score. As a robustness check, we also compute an Elo ranking from pairwise outcomes induced by peer evaluations. Figure 18 shows that Elo and mean peer scores yield essentially the same ordering, suggesting our conclusions are not an artifact of the absolute 1–10 scale or judge calibration. At the same time, this agreement does not resolve application-specific trade-offs, motivating a multi-dimensional view of model behavior.

Multi-Dimensional Model Characterization Scalar rankings obscure important trade-offs between model attributes. Figure 19 presents a multi-dimensional comparison across normalized quality, speed, consistency, humility (inverse self bias), and strictness (inverse generosity).

No single model optimizes all dimensions simultaneously. Models with high peer scores may exhibit higher latency or lower humility, while faster models may trade off consistency or strictness. These results caution against over-reliance on single-number leaderboards and motivate richer evaluation artifacts that expose behavioral profiles.

Broader Implications and Conclusion Taken together, these findings suggest several implications for LLM evaluation. First, peer-based evaluation can serve as a scalable complement—and in some settings an alternative—to human-anchored benchmarks, provided that bias is explicitly measured and controlled. Second, evaluation outcomes are sensitive to presentation, identity, and judge-specific effects, implying that naïve LLM-as-a-judge pipelines risk systematic distortion. Third, open-world, web-grounded evaluation introduces realistic variance that static, closed-world benchmarks tend to suppress.

PeerRank does not aim to replace human evaluation or correctness-based benchmarks. Instead, it offers a continuously updatable evaluation paradigm aligned with deployment conditions, where models retrieve information, synthesize across sources, and operate without fixed reference answers. By making the evaluation distribution endogenous and treating bias as a measurable quantity, PeerRank shifts evaluation from matching static answer keys toward understanding relative performance under a shared, bias-aware protocol.

Limitations (1) No absolute ground truth —scores are relative to this population; a uniformly weak cohort could produce high absolute scores. PeerRank evaluates models relative to the participating population; absolute scores should not be compared across disjoint runs without calibration. (2) Task distribution bias —questions reflect generator capabilities and may underrepresent certain domains. (3) Temporal confounds —API latency reflects server load, not purely model computation. (4) Moderate scale —while the study covers 12 models and 420 questions, statistical power may still be limited for fine-grained subgroup analyses (e.g., per-category or per-judge effects). (5) Rubric subjectivity —judges may weight criteria differently despite standardization.

Future Work We plan to extend PeerRank in four concrete directions. First, we will study sensitivity to prompt design by systematically ranking prompt templates themselves—varying rubric wording, in-

struction framing, and judge context—and quantifying how these choices shift model ordering and score distributions. Second, we will quantify the impact of real-world web grounding tools and practices (e.g., retrieval settings, citation requirements, and browsing strategies) on both peer scores and ranking stability. Third, we will analyze how rankings and judge behaviors differ between *reasoning* models and *non-reasoning* models, testing whether explicit deliberation changes answer quality, calibration, bias sensitivity, or evaluation consistency under the same peer-review protocol. Together, these extensions aim to strengthen the validity and reproducibility of peer-based evaluation by anchoring results to verifiable truth checks and making prompt and grounding effects explicit.

Conclusion PeerRank reframes large language model evaluation as an emergent, social process among models rather than alignment to a fixed human-authored ground truth. Our results show that fully endogenous peer evaluation can produce stable and discriminative rankings, provided that systematic judge effects are explicitly measured and controlled.

References

[1] Anthropic. Models overview. https://docs.anthropic.com/en/docs/about-claude/models, 2025. Accessed: 2026-01-23.

[2] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[3] Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, et al. Benchmarking large language models under data contamination: A survey from static to dynamic evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10091–10109, 2025.

[4] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Angelopoulos, Zhuohan Li, Dacheng Li, Banghua Zhu, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference, 2024.

[5] Xiuying Chu, Yongsheng Zhang, Bing Qin, and Ting Liu. PRE: A peer review based large language model evaluator. In Yuki Miyao, Ani Nenkova, and Dan Roth, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4648–4670. Association for Computational Linguistics, 2024.

[6] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[7] Nurit Cohen-Inger, Yehonatan Elisha, Bracha Shapira, Lior Rokach, and Seffi Cohen. Forget what you know about llms evaluations–llms are like a chameleon. *arXiv preprint arXiv:2502.07445*, 2025.

[8] DeepSeek. Deepseek api documentation. https://platform.deepseek.com/docs, 2025. Accessed: 2026-01-23.

[9] Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, 1978.

[10] Google. Gemini models documentation. https://ai.google.dev/gemini-api/docs/models, 2025. Accessed: 2026-01-23.

[11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[12] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[13] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. HELM: Holistic evaluation of language models, 2022.

[14] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods, 2021.

[15] Xiao Liu, Lianmin Zheng, Yu Du, Xiaowei Huang, Yan Sun, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. AgentBench: Evaluating LLMs as agents, 2023.

[16] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522. Association for Computational Linguistics, 2023.

[17] Meta. Llama model documentation. https://www.llama.com/docs/, 2025. Accessed: 2026-01-23.

[18] Mistral AI. Mistral models documentation. https://docs.mistral.ai/getting-started/models/, 2025. Accessed: 2026-01-23.

[19] Moonshot AI. Kimi api documentation. https://platform.moonshot.cn/docs/, 2025. Accessed: 2026-01-23.

[20] Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. A survey on large language model benchmarks. *arXiv preprint arXiv:2508.15361*, 2025.

[21] OpenAI. Models documentation. https://platform.openai.com/docs/models, 2025. Accessed: 2026-01-23.

[22] Perplexity AI. Sonar api documentation. https://docs.perplexity.ai/, 2025. Accessed: 2026-01-23.

[23] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics, 2016.

[24] Dimitrios Rontogiannis, Maxime Peyrard, Nicolas Baldwin, Martin Josifoski, Robert West, and Dimitrios Gunopulos. Interactive evaluation of large language models for multi-requirement software engineering tasks. *arXiv preprint arXiv:2508.18905*, 2025.

[25] Lin Shi, Caroline de Langhe, Eve Fleisig, Dan Sun, Aayush Sharma, Yejin Choi, and Noah A. Smith. Judging the judges: Evaluating alignment and reliability of LLM evaluators. In *International Conference on Learning Representations*, 2025.

[26] Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of IJCNLP-AACL (Long Papers)*, 2025.

[27] Jiaqi Wang et al. Unsupervised peer model evaluation (UPME): Evaluating large language models without human labels, 2024.

[28] Yiming Wang, X Long, Y Fan, J Wei, and X Huang. Benchmark self-evolving: A multi-agent framework for dynamic LLM evaluation, 2024.

[29] Yubo Wang et al. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

[30] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

[31] xAI. xai api documentation. https://docs.x.ai/, 2025. Accessed: 2026-01-23.

[32] Jiaxuan You, Mingjie Liu, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. LLM-evolve: Evaluation for LLM's evolving capability on benchmarks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16937–16942. Association for Computational Linguistics, 2024.

[33] Yuezhan Zhao et al. Justice or prejudice? quantifying biases in LLM-as-a-judge, 2024.

[34] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena, 2023.

[35] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Kyle Lo, Zichao Liu, Zhuohan Chen, Yonatan Bisk, Daniel Fried, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents, 2023.

A Prompts and constants

This appendix lists the exact prompt templates and constants used for question generation, web-grounded answering, and peer evaluation.

A.1 Constants

Name	Value
Questions per model	`NUM_QUESTIONS = 35`
Categories	`factual knowledge; reasoning / logic; current events; creative / open-ended; practical how-to`
Max answer length	`MAX_ANSWER_WORDS = 200`
Score range	$\{1, \ldots, 10\}$
Judge reason length	8–20 words
Judge flags	fixed vocabulary array (see Peer Evaluation prompt)

A.2 Prompt: Question Generation

```
Generate exactly {NUM_QUESTIONS} diverse questions for testing AI capabilities.

Use ONLY these exact category values: {category_list}

Return as JSON object:
{"questions": [{"category": "factual knowledge", "question": "Your question"}]}
```

A.3 Prompt: Web-Grounded Answer Generation

```
Answer this question directly and concisely in {MAX_ANSWER_WORDS} words or less.
Do not start with "Based on..." or similar preambles.

{question}
```

A.4 Prompt: Peer Evaluation

```
You are grading responses for a benchmark. Score EACH response independently.

DO NOT try to identify authorship. Ignore writing style and focus on quality.

Scoring rubric (overall 1-10 integer):
- 10: Correct + complete + well-justified; directly answers; no hallucinations.
- 7-9: Mostly correct; minor omissions/imprecision; reasoning mostly sound.
- 4-6: Mixed/partial correctness, unclear reasoning, or misses key constraints.
- 1-3: Mostly incorrect, misleading, evasive, or hallucinated/unsupported.

Priority rules:
- Prioritize correctness/faithfulness over eloquence.
- Penalize confident-sounding unsupported specifics (made-up numbers, names, dates, citations).
- Citations [1][2], source mentions, or "I searched..." phrasing are NEUTRAL - do not reward or penalize.
- If the question is subjective/creative: score instruction-following, coherence, and usefulness; do not
    mark "incorrect" unless it violates constraints or is nonsensical.

Calibration / score discipline:
- Use the full range when justified; avoid clustering 7-8.
- If there are 3+ responses and quality differs, use at least 3 distinct scores.
```

```
Question:
{question}

Responses:
{responses}

Output format (STRICT):
- Return ONLY a single JSON object (no markdown, no extra text).
- You MUST include an entry for EVERY label present in Responses, exactly once.
- Each entry MUST contain keys: "score", "reason", "flags" (no other keys).
- "score" MUST be an integer 1--10.
- "reason" MUST be 8--20 words and cite a specific strength or flaw.
- "flags" MUST be an array using only:
  "hallucination", "unsupported_specifics", "evasive", "incorrect",
  "good_uncertainty", "clear_correct"
  Use [] if none apply.

Example:
{"{label_example}": {"score": 8, "reason": "Correct core claim, minor omission on edge case; clear and
     grounded.", "flags": ["clear_correct"]}}
```

Derived metrics example: home advantage

Model	Own Qs	Other Qs	Diff	n_{own}	n_{other}	Cohen's d	Sig
kimi	6.66	7.68	-1.01	365	3742	-0.43	***
gem-3-pro	8.96	8.39	+0.57	343	3764	+0.29	***
mistral	8.51	8.09	+0.42	380	3790	+0.19	***
deepseek	7.74	8.06	-0.32	324	3780	-0.14	*
gemini-3-fla	7.48	7.83	-0.35	320	3784	-0.17	**

Table A1: Models with statistically significant differences between performance on own-authored vs. other-authored questions (peer scores exclude self-evaluation). Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Average home advantage across models is +0.005 points.

A.5 Ethics statement

This work evaluates commercial AI systems via public APIs within terms of service. No human subjects were involved. Rankings may influence deployment decisions; readers should consider task-specific requirements beyond aggregate scores.

A.6 Reproducibility

API calls were made January 25, 2026; results may vary with model updates. Code, prompts, and raw data are available at