On the Role of Citations in Preference Data

Yu Hou[1], Hal Daumé III[1], Rachel Rudinger[1], William Walden[2]

[1]University of Maryland, [2]Johns Hopkins University

Correspondence: houyu@umd.edu,wwalden1@jh.edu

Abstract

Many NLP tasks require systems to provide *attribution* in their outputs—i.e. citations to grounding sources. Attribution serves as a bulwark against model hallucination and as a means for users to verify the credibility of model outputs. Yet, it is unclear how humans and LLMs evaluate citations when comparing outputs, a process central to reward modeling and modern LLM post-training. This paper studies the role of citations in the preferences of human judges and LLMs within the context of scientific question answering, leveraging mixed effects models to investigate the influence of citations on pairwise judgments. Among our key findings are (1) that humans prefer more *diverse* citations but fewer citations *overall*, and (2) that LLMs show some stronger citation-related preferences compared to humans, despite lacking access to the sources, but these preferences depend on the data and specific models. Finally, we discuss the implications of our findings for preference data collection.[1]

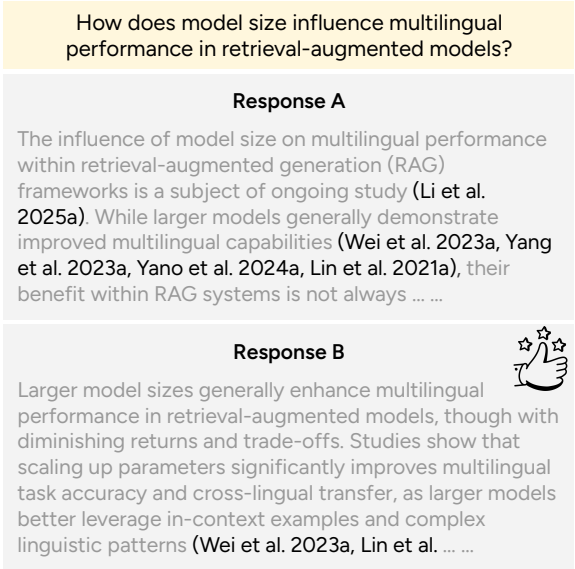

Figure 1: A pair of citation-attributed model responses to a question from SciArena (Zhao et al., 2025a). We use mixed-effects models to understand the impact of citations on human and LLM preferences on such pairs.

1 Introduction

Pairwise preference data plays an essential role in LLM development, both in post-training (Ouyang et al., 2022; Li et al., 2025) and in evaluation (Chiang et al., 2024), enabling models to learn difficult-to-specify behaviors and to be efficiently ranked and evaluated. But we pay for these benefits with interpretability: the reason for a given preference is generally not known—perhaps even to the assessors themselves. Such interpretability is vital to understanding dimensions of human and model bias, which in turn is a necessity for the development of fair and robust reward models (Dubois et al., 2024; Wu et al., 2025; Bharadwaj et al., 2025; Zhao et al., 2025b; Gehrmann, 2025).

This need has motivated a budding line of research on inferring interpretable explanations for preference judgments. Much of this work adopts a "propose-and-validate" paradigm, in which candidate explanations are *proposed*—e.g. via LLM analysis (Findeis et al., 2025a; Zhong et al., 2024) or via sparse autoencoder features (Movva et al., 2025)—and then *validated* by attempting to predict (human or model) judgments from these explanations. This work, though valuable, neglects a category of data core to most retrieval-augmented generation (RAG) and deep research applications—namely, long-form responses *with citations*.

Specifically, as more people turn to AI as scientific research assistants—for finding papers or answering scholarly queries—citations play an increasingly important role in information-seeking tasks than they once did with traditional web search. In this work, we ask: **how do citation-related features influence human preferences in scientific question answering?** Further, because LLM

[1]Code and data can be found at: https://github.com/houyu0930/citation-preference.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 40582–40599
October 24-29, 2026 ©2026 Association for Computational Linguistics

judges are often used for preference labeling as a proxy for human expert annotations, we compare: **how do LLMs weigh citations in their pairwise assessment of responses?**

We draw on *mixed-effects models* (MEMs; Gelman and Hill, 2014) to interrogate these questions on two datasets, which have pairwise judgments on citation-backed model responses to diverse scientific questions (Figure 1): SciArena (Zhao et al., 2025a) and ResearchQA (Yifei et al., 2025). We use human preferences from these datasets and set up LLM judge experiments to compare the observed effects on each dataset. Specifically, we develop a structured approach for auditing preference data: we decompose judgments into interpretable dimensions and estimate their associations with human and LLM preferences separately, allowing a side-by-side comparison. This approach is not specific to the datasets studied here, and can be applied to other preference collections.

Overall, we find that: **(1)** human judges tend to prefer responses with *more diverse* but *fewer overall* citations. **(2)** Experiments on four open-source LLMs (Llama, Gemma, Qwen, and Skywork) show that the influence of citation-related and non-citation features on LLM preferences differs from their influence on human preferences, highlighting potential LLM biases. **(3)** Among the four LLM judges, we find that they exhibit dissimilar correlational patterns and citation preferences. We discuss the implications of these findings, hoping to motivate more careful attention to the treatment of citations in preference data collection.

2 General Setup

Although both SciArena (Zhao et al., 2025a) and ResearchQA (Yifei et al., 2025) focus on scientific QA scenarios, their processes for collecting preferences differ. We discuss the high-level setup here and provide more details in Section 3 and Section 4.

Dataset Selection Criteria Our analysis requires datasets that satisfy three conditions. First, responses must contain citations.[2] Second, each query must have multiple responses so that pairwise comparisons are possible. Third, because we wish to compare LLM judge behavior with that of humans, each response pair must have a clear human preference vote, allowing us to probe LLM judge preferences on the same data. Following an expansive search of existing datasets, these requirements constrained us to only two options: SciArena, with $12,249$ comparisons, and ResearchQA, with 169 available comparisons. In both datasets, human annotators had access to the cited sources.

LLM Inference Experiment We elicit preferences on these examples from four LLMs: Llama3.3-70B (Grattafiori et al., 2024), Gemma3-27B (Team et al., 2025), Qwen3-30B (Yang et al., 2025), and Skywork-Critic-70B (Shiwen et al., 2024). Skywork is specifically trained as a generative reward model and was selected due to its top rank on RAG-RewardBench (Jin et al., 2025), while the other three were selected due to their general popularity among open-weight models, though they were not trained explicitly as reward models. To control for biases arising from presentation order, we run experiments for both possible orderings of responses in the prompt, yielding $12,249 \times 2 = 24,498$ and $169 \times 2 = 338$ judgments per LLM for SciArena and ResearchQA, respectively. We use the same judge prompt for both datasets (see below). [3]

LLM Judge Prompt

```
You are an expert in scientific
literature synthesis. Your task is to
evaluate the quality of two AI-generated
citation-attributed responses to a user's
question. Assess both responses for
relevance, accuracy, clarity, and appropriate
use of citations. Then, select the response,
Output (a) or Output (b), that best address
the user's question.

User Question:
{question}

Output (a):
{response_a}

Output (b):
{response_b}

Which is best, Output(a) or Output (b)?
Answer with only: "A" or "B"
```

Mixed-Effects Models (MEMs) are hierarchical regression models that incorporate slope and/or intercept parameters that vary by group (*random effects*) as well as parameters that are shared across groups (*fixed effects*). MEMs are widely used for interpretable data analysis (Gelman and Hill, 2007; McElreath, 2018; Federico et al., 2014; Gantt et al., 2020; White et al., 2022; Sun et al., 2023; Sandoval

[2] We found that many preference datasets that *ought* to satisfy this first condition do not.

[3] For further experiment setup details, see Appendix A.

et al., 2023; Benito-Santos et al., 2025). Following the Bradley-Terry model (Bradley and Terry, 1952), we treat binary preferences as the dependent variable ($\boldsymbol{y}$) in a mixed-effects logistic regression:

$$\text{logit}(P(\boldsymbol{y} = 1 \mid \mathbf{u})) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

where $\mathbf{X}$ is a matrix of predictors; $\boldsymbol{\beta}$ is a vector of fixed-effects regression coefficients; $\mathbf{Z}$ is the design matrix for random effects; $\mathbf{u}$ is a vector of random effects, where $u_j \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$, with $\mathbf{G}$ as the covariance matrix; and $\boldsymbol{\beta}$ and $\mathbf{G}$ are the parameters to be inferred.

Fixed-Effects Predictors We consider three citation-related variables for each dataset: citation diversity, citation count, citation age (SciArena only), and citation mismatch (ResearchQA only). These variables are a subset of those studied in citation bias research (Wahle et al., 2025; Algaba et al., 2025; Fang et al., 2025), targeted at comparison of closed sets of citations rather than (fully open) citation selection (as responses in a pair are conditioned on the same query and retrieved documents).

Citation diversity ($\mathbf{C}_D$) captures the diversity of cited sources and is measured by the Shannon diversity index (i.e. entropy): $-\sum_{i=1}^{n} p_i \log_2(p_i)$, where n is the number of unique sources in the response and p_i represents the proportion of citations attributed to source i. Low values indicate low diversity (i.e. when a response is dominated by only a small number of sources, repeatedly cited).

Citation count ($\mathbf{C}_C$) measures the number of citations in a response. C_C and C_D are used to test hypotheses about citations' *distribution*.

Citation age ($\mathbf{C}_A$) [SciArena Only] captures the average publication date of cited sources as presented in Figure 1, calculated as $\text{avg}(\ln(2026 - y))$, where y ranges over citations' publication years.[4] Since dates are not present in ResearchQA citations, this predictor cannot be used for that dataset.

Citation mismatch ($\mathbf{C}_M$) [ResearchQA only] captures discrepancies between inline citations and the reference list in ResearchQA responses. Unlike SciArena, which uses author-year inline citations (e.g., "`Li et al. 2025`"), ResearchQA responses commonly use numbered citations (e.g., "`[1]`", "`[2]`") accompanied by a reference list at the end of the response. In a minority of cases, inline citations and reference list entries do not match: either an inline citation has no entry in the reference list or vice-versa. To account for the possibility that judges may penalize such discrepancies, we count the number of mismatched citations C_M.

To give meaningful interpretations to these citation-related effects, which are often intertwined with response content, we control for two other potential confounds apart from presentation order:

Response length ($\mathbf{R}_L$, in characters)—a common source of bias in preference data (Dubois et al., 2024; Park et al., 2024; Hu et al., 2025).

Response content ($\mathbf{R}_C$) measures the difference in unique content between two responses (after removing inline citations). We use BERTScore (Zhang et al., 2020) precision to capture semantic overlap: the precision of BERTScore(A, B) reflects how much of Response A's content is covered by Response B. We define the proportion of unique content in A relative to B as: $1 - \text{Precision}(A, B)$, where a higher value indicates more unique content.[5]

We express each continuous predictor as the z-scored difference between Response A (R_A) and Response B (R_B) for that variable. No fixed-effects predictors show multicollinearity, with variance inflation factors (VIFs) < 2.

Random-Effects Design Given the differences between the two datasets (e.g., the number of models involved in generating responses), we specify a maximal random-effects structure for each dataset separately (Barr et al., 2013). We then simplify each maximal model by iteratively eliminating random effects via AIC-based (Akaike, 1974) model selection to ensure parsimony and minimize overfitting (Matuschek et al., 2017), repeating this process for each set of human or LLM judgments.

Average Marginal Effects We report population-level average marginal effects (Williams, 2012; Arel-Bundock et al., 2024). Results are shown as ΔWin Rate in percentage points (*pp*): the change in probability that R_A is chosen per 1-SD increase in the predictor, marginalizing over random effects.

[4] We apply a log transformation as the date distribution is heavily skewed, with 79.67% of citations falling after 2019, with a median of 2023 and a mode of 2024.

[5] We use `roberta-large` for BERTScore.

Higher values indicate (for each predictor) that R_A has (a) higher citation diversity, (b) more total citations, (c) overall older citations (SciArena) or more mismatched citations (ResearchQA), (d) a longer response, or (e) more unique content.

3 Case Study: SciArena

For both SciArena (§3) and ResearchQA (§4) we first discuss details of the data, how those details inform our mixed-effects model design, and what research questions we seek to answer with these models. We then discuss observations about the influence of citation-related features on human preferences and how this compares to the influence of *non*-citation-related features. Finally, we present observations on the four LLM judges.

3.1 Setup Details

Data SciArena (Zhao et al., 2025a) contains majority preferences from 102 researchers on pairs of LLM-generated responses to scientific queries (Q). We denote an ordered pair of responses as R_{AB}=(R_A, R_B) and the ordered pair of models that generated R_{AB} as M_{AB}=(M_A, M_B). M_A and M_B generate responses given the same set of retrieved documents. We focus on examples where (i) both R_A and R_B contain ≥ 1 citation(s) and (ii) the preference label is *either* R_A or R_B, filtering cases where R_A and R_B tie or are both deemed bad. This results in $12,249$ examples, covering 23 models and 480 unique model pairs. Each Q is labeled with one of six question types (Q_{type}) and one of 25 subjects (Q_{subj}) drawn from one of the five main categories (Q_{cat}).

Mixed-Effects Model Design The rich features of SciArena—including many responses per model (M_A/M_B), question category (Q_{cat}), and question type (Q_{type})—motivate our use of a relatively complex maximal random effects structure. We include random *intercepts* for M_A, M_B, M_{AB}, Q_{type}, and Q_{subj}, and for interactions defined by $\{M_A, M_B\} \times \{Q_{\text{type}}, Q_{\text{cat}}\}$. Second, we include random *slopes* for all continuous variables (C_D, C_C, C_A, R_L, R_C) separately for each M_{AB}, Q_{subj}, and Q_{type}.[6] The continuous fixed-effects predictors are citation diversity, citation count, citation age, response length, and response content. Appendix B has details on model selection and final model fits.

[6]This random-effects design, written below (in `lme4` syntax), in the maximal structure:

$$
\begin{aligned}
&(1|M_A) + (1|M_B) \\
&+ (1|M_A : Q_{\text{type}}) + (1|M_B : Q_{\text{type}}) \\
&+ (1|M_A : Q_{\text{cat}}) + (1|M_B : Q_{\text{cat}}) \\
&+ (\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C|M_{AB}) \\
&+ (\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C|Q_{\text{subj}}) \\
&+ (\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C||Q_{\text{type}})
\end{aligned}
$$

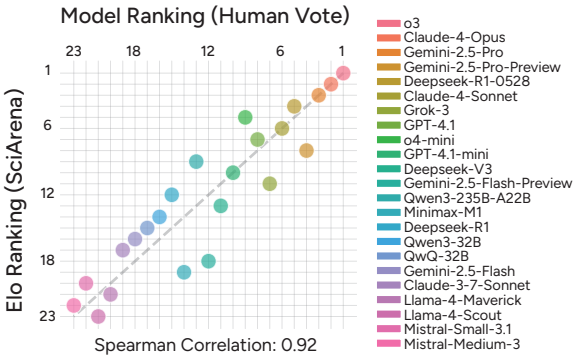

Figure 2: Model ranking correlation between Elo Ranking (SciArena) and Random-Effects Ranking (Human).

Research Questions For SciArena, we ask:

1. For both human and LLM judges, are more diverse citations preferred or is it only the total citation count that matters? How do the effect sizes for these predictors compare to those for non-citation-related predictors?
2. Do citations' publication dates correlate with preferences, as newer information may reflect more current science than older information?

3.2 Human Preferences

Figure 3 plots average marginal effects on win rate (Δ Win Rate) for the three citation-related predictors and two non-citation variables.[7] Human preferences are presented on the left. Here, we observe that a 1σ increase in citation diversity difference (i.e. when R_A has more diverse citations) yields a 1.8% absolute higher probability that R_A is preferred by human judges (Figure 3a). Conversely, a 1σ increase in total citation difference yields a 3.5% *lower* probability of R_A being selected (Figure 3b). Thus, human judges prefer *more diverse* citations, but fewer of them *overall*. The effect size is similar to the response length effect (-1.2 *pp*, n.s.; Figure 3d) and the preference for semantically unique response content (4.2 *pp*, $p < 0.05$; Figure 3e). We also find that humans show slight preferences for older citations, but the result is not significant (Figure 3c).

[7]Exact estimates are reported in Table 2 (Appendix B).

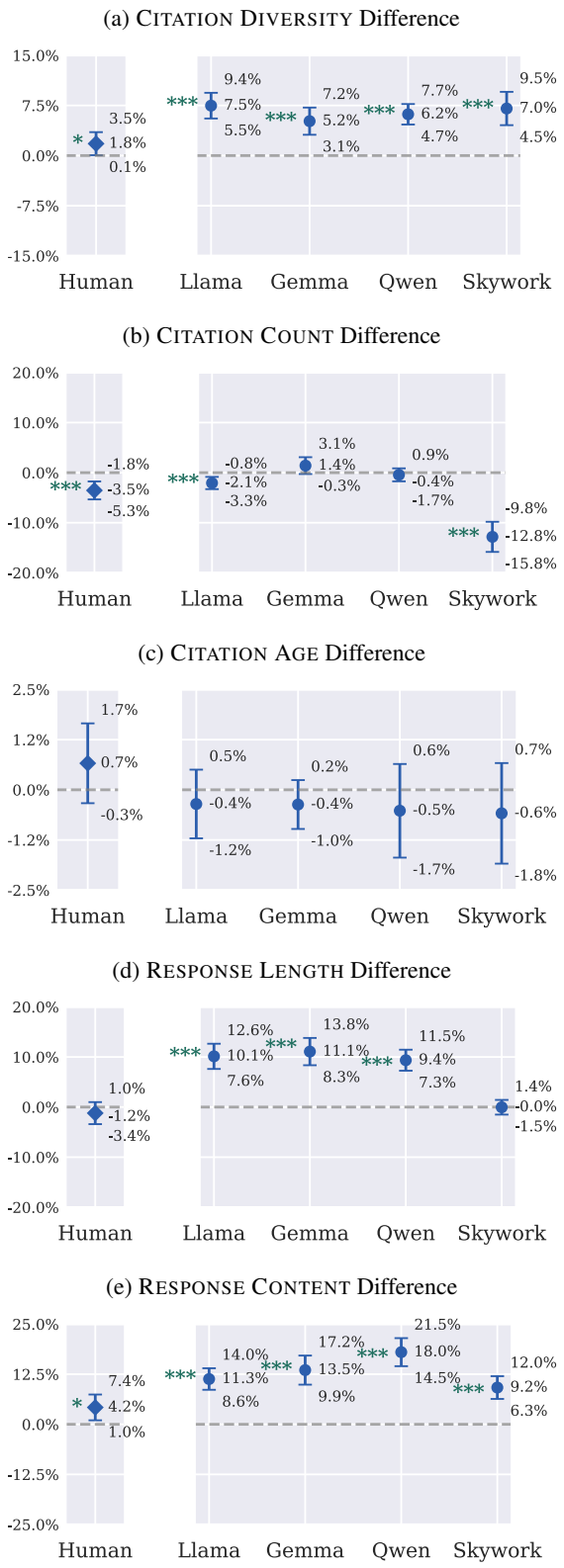

Figure 3: **SciArena**: Estimated fixed effects on human votes and LLM judgments (ΔWin Rate). Sub-figures (a)–(c) show citation-related predictors, while (d) and (e) show non-citation control variables for effect size comparison. Error bars represent 95% CIs. Significance: ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

Further, by calculating a leaderboard ranking based on both M_A and M_B random intercepts, we find a Spearman correlation of 0.92 with the Elo Ranking reported in SciArena (see Figure 2), further justifying our mixed-effects model design.

3.3 LLM Judge Preferences

Compared to humans, LLM judges show similar but stronger preferences for diverse citations (Δ Win Rate={7.5%, 5.2%, 6.2%, 7.0%} for Llama, Gemma, Qwen, and Skywork; Figure 3a). For total citation count, we observe more mixed results (Figure 3b). Llama yields a 2.1 *pp* lower probability of R_A being selected, and Skywork presents an even stronger negative correlation (-12.8 *pp*, $p < 0.001$). Results for the other two LLM judges are not significant. None of the LLM judges shows significant citation age-related preferences, though the estimated effects differ in polarity from those of human judges, showing a slight preference for newer citations (Figure 3c).

For non-citation-related variables, Llama, Gemma, and Qwen all exhibit strong preferences for longer responses (Figure 3d). By contrast, Skywork's preference is negligible, possibly because it is the only trained reward model among the four. The response content predictor (which captures response content uniqueness) shows much stronger correlations than those of human judges across all LLM judges (Figure 3e). While citations can plainly influence how humans perceive information—whether by introducing extra cognitive load or by affecting perceived trustworthiness—LLMs process this information in a fundamentally different manner. Citations constitute only a small portion of the response, and LLM judges lack the agency to search and verify the cited content (at least in the annotation protocols for SciArena and ResearchQA). Thus, it is intuitive that differences in response content are more likely to drive LLM predictions. Controlling for content-related variables thus provides greater confidence in understanding the role of citations in the preferences of LLM judges.[8]

[8]We further test prompt sensitivity using two variants: (i) a chain-of-thought (CoT) variant that asks the model to reason step by step before deciding, and (ii) a length-debiasing variant that adds an explicit instruction sentence. Decisions largely remain consistent across variants for all four judges (Cohen's $\kappa \geq 0.67$, $\geq 84\%$ raw agreement under CoT; $\kappa \geq 0.93$, $\geq 97\%$ under length debiasing). More details are in Appendix A.

4 Case Study: ResearchQA

4.1 Setup Details

Data ResearchQA (Yifei et al., 2025) is a large resource for evaluating LLM systems through curated queries and rubric items based on existing survey articles. However, only 169 of these examples feature human annotations (from a pool of 31 PhD annotators) that satisfy our criterion of having a clear preference for one response over the other. These examples span eight scientific fields (e.g., NLP) from five general domains (e.g., Engineering), and all compare responses from the same pair of models, `gpt-4.1-mini` and `gemini-2.5-flash`.

Mixed-Effects Model Design Given the limited data, we design the maximal random effects structure simply as $(1|Q_{\text{field}})$ to ensure reasonably robust parameter estimates. For the model trained on human preferences, we randomize the order of responses within each pair so that Response A has an equal chance of being the `gpt-4.1-mini` or the `gemini-2.5-flash` response. For the model trained on LLM preferences, we create two versions of each example for both orderings of responses (as in SciArena), and assess position bias through both the (global) fixed effect intercept and a separate presentation order predictor. Since all examples compare the same two models, we do not include by-model random intercepts.

As noted in §2, citations in ResearchQA lack dates and thus we eliminate the citation age fixed effect predictor. Additionally, since ResearchQA responses track citations both *inline* and as a reference list at the *end* of the text, we introduce a *citation mismatch* predictor to account for occasional discrepancies between these two lists that may impact preference judgments. See Appendix B for details on the final model fits.

Research Questions For ResearchQA, we ask:

1. For both human or LLM judges, what is the influence of citation diversity and total citation count on preference judgments?
2. How much does the inconsistency between inline citations and the reference list matter when choosing between responses?

4.2 Human Preferences

Figure 4 shows the observed effects.[9] Despite the differences in the preference data collection process between ResearchQA and SciArena, we obtain

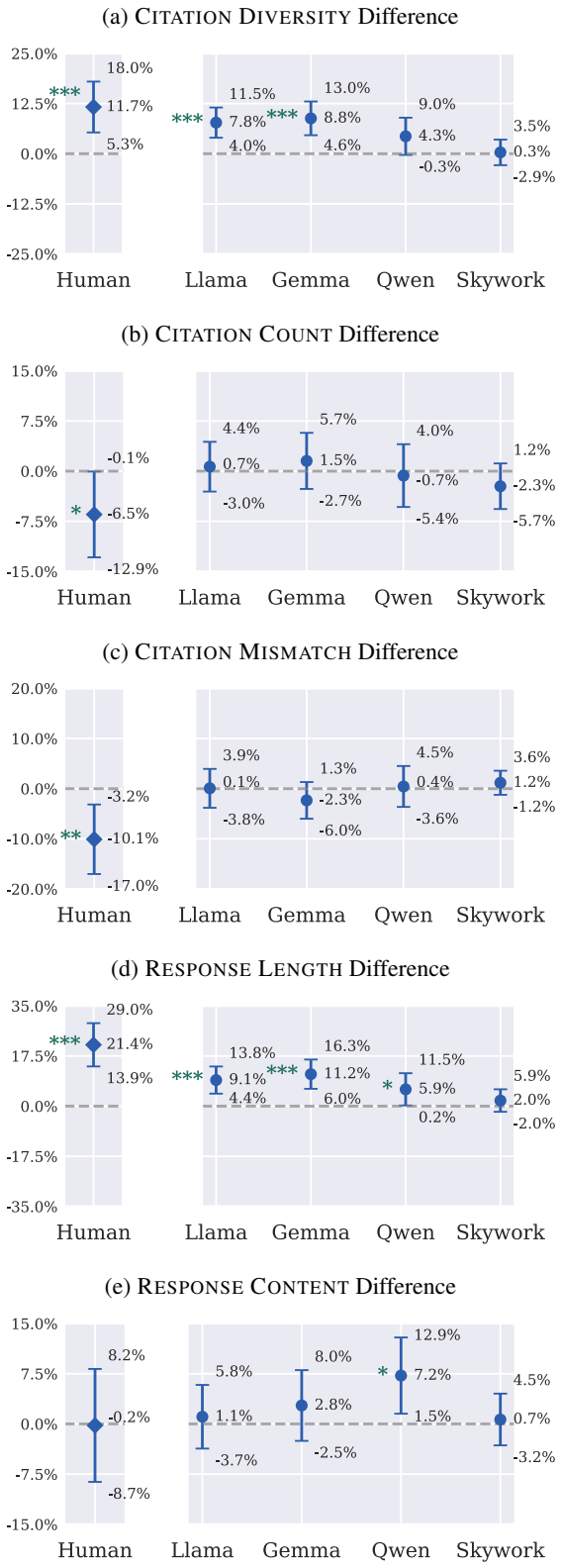

Figure 4: **ResearchQA**: Estimated fixed effects on human votes and LLM judgments (ΔWin Rate). Subfigures (a)–(c) show citation-related predictors, while (d) and (e) show non-citation control variables for effect size comparison. Error bars represent 95% CIs. Significance: ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

[9] Exact estimates are reported in Table 3 (Appendix B).

consistent results: here again, human judges prefer more diverse citations but fewer total citations (Figure 4a, Figure 4b). Furthermore, the citation mismatch predictor significantly negatively correlates with a response's probability of being chosen, suggesting that humans penalize discrepancies between inline citations and the reference list (e.g., due to hallucinated citations)—a sensible reaction (Figure 4c). The effect sizes are fairly large, with Δ Win Rate = {11.7%, -6.5%, -10.1%} for citation diversity, citation count, and citation mismatch. By comparison, Δ Win Rate = 21.4% for response length difference, while no significant effects are observed for response content difference as modeled here (Figure 4d, Figure 4e).[10]

4.3 LLM Judge Preferences

For citation-related predictors (Figure 4a, Figure 4b, Figure 4c), we find significant associations only between citation diversity and preferences from Llama and Gemma, with effect sizes of 7.8 *pp* and 8.8 *pp*, respectively. Citation diversity effects are also positive for Qwen and Skywork, though not significant. For citation count and citation mismatch, we do *not* observe the same strongly negative effects that we saw for human judges, revealing an interesting divergence in the citation-related preferences between the two judge categories.

Consistent with our SciArena results, Llama, Gemma, and Qwen all show significant bias toward longer responses (Δ Win Rate = {9.1%, 11.2%, 5.9%}), while Skywork (a trained reward model) does not (Figure 4d). For response content difference (Figure 4e), only Qwen shows a preference for unique content (7.2 *pp*, $p < 0.05$).

Sensitivity to Citation Perturbations To probe how citation *validity* and surface cues affect judge preferences, we run four citation-perturbation experiments on ResearchQA. For each item and condition, we randomly select one of the two responses and perturb its reference list, leaving the response text and inline `[N]` markers unchanged: (1) replace each publication year with a future year, (2) replace each year with a random incorrect past year, (3) shuffle the words within each title, or (4) permute URLs across the reference list, breaking the original title–URL pairings. We run four judges and report Δ, the decrease in preference for the perturbed response relative to its intact-citation baseline.[11]

All four judges show reduced preference for the perturbed response under the first three conditions ($\Delta = +5.9$ to $+20.1$ percentage points, all $p < 0.001$). Title–URL permutation produces smaller shifts ($+2.0$ to $+4.0$ points), significant only for Gemma and Skywork. Random incorrect years produce larger decreases than future years for every judge (by 2.7–9.2 points), despite future dates being overtly implausible. Among items where the selected response is preferred in both baseline presentation orders, title shuffling causes the judge to choose the alternative in at least one order for 36–46% of items; random incorrect years do so for 37–67%, depending on the judge. These results demonstrate sensitivity to citation perturbations, but leave open the extent to which LLM judges base their preferences on citation validity.

5 Discussion

5.1 Collecting Human Preference Data

A key fact about many preference data annotation protocols is that they tend to *underspecify* the criteria by which responses ought to be judged. Especially when annotators are domain experts, it is often viewed as beyond the purview of those collecting the data to prescribe a set of factors to be considered in making such judgments. Rather, these factors are to be learned *implicitly* by a reward model later trained on those judgments.

Yet, even experts are not immune from lazy or heuristic assessments of citation quality—a tendency that is especially undesirable in (ostensibly rigorous) academic settings. This reflects a broader pattern: information source preferences are influenced by internal factors such as user's cognition and affection (Zhong and Han, 2023; Łajewska et al., 2024). As such, those wanting to discourage these practices during data collection may find it valuable to provide clear citation-related rubrics or tools that assist annotators in more easily understanding cited content (e.g., reading assistants that surface key source passages (Chen et al., 2025)). Furthermore, the effects of particular protocol design choices should be at least partially observ-

[10]Note that the observed human preferences on response length and content differ from SciArena, likely due to differences in annotator populations, collection setups, and response comparisons. We leave further investigation to future work.

[11]The last condition requires at least two references on the perturbed side and covers 126 of the 169 items; the other three cover all 169 ($\times$ 2 presentation orders). We average preferences across the two presentation orders within each item and use a paired Wilcoxon signed-rank test. Positive Δ indicates reduced preference after perturbation.

able with MEMs, giving data consumers (e.g., post-training or evaluation teams) confidence that the resulting preference data reflects desired behavior.

5.2 Collecting LLM Preference Data

Recruiting human judges to produce preferences on long-form responses can be both costly and challenging. While it is much easier to set up an LLM judge to approximate human preferences, our analysis reveals that LLM judges exhibit substantially different preference patterns from humans. Moreover, these patterns are data-dependent: in some cases, citation-related factors seem to exhibit similar influences on both human and LLM judges, while in others, they may exhibit undesirable biases on superficial features (e.g., citation year, response length) or diverge from humans in other respects (e.g., in how citation inconsistencies are penalized).

One straightforward way to reduce these divergences is to ensure that LLM judges have access to cited documents through external tools (Findeis et al., 2025b)—just as human judges may search a cited source on the web or preview them in the annotation interface itself. Additionally, for domains with clearly established best practices for citations, domain-specific rubrics (of the kind discussed above) may be employed to ensure that LLM judges consider the extent to which a response emulates those standards (Farzi and Dietz, 2025). Such factors demand careful consideration when designing LLM judge systems.

5.3 Choosing an LLM Judge

Finally, we consider the implications of the differences observed among LLM judges.

First, it is evident that LLM judges *do* exhibit diverse preferences. For instance, models show quite different degrees of preference for unique response content. In some cases, these differences may indicate substantial differences in training—as in the lack of response length bias in Skywork compared to the other three models.

Second, and as a consequence of this first observation, assessing LLM judge quality simply via accuracy or correlation relative to human preferences risks obscuring idiosyncrasies of individual judge models. Analysis with MEMs of the sort conducted here is one tool that can aid in measuring biases of interest and in facilitating more nuanced and principled LLM judge selection. When collecting preference data, it is thus imperative to consider not only *whether* to use an LLM judge, but also *which* LLM judge to use and what biases are or are not acceptable for one's use case.

6 Related Work

Human/LLM Preferences Pairwise comparisons have been widely used to understand human preferences. Among these efforts, a first line of work focuses on evaluation using human preferences for a variety of complex tasks, such as information-seeking via search and coding (Chiang et al., 2024; Miroyan et al., 2025; Chi et al., 2025; Bianchi et al., 2026). A second line of work proposes different approaches to mine or uncover latent preferences from aggregated human pairwise votes (Zhong et al., 2024; Findeis et al., 2025a; Movva et al., 2025). Further, given the difficulty of collecting high-quality human preference labels, existing works use LLM judges to obtain similar pairwise votes. Despite the greater convenience and scalability of LLM judgments, much of this work has also discussed potential LLM or LLM judge biases (Bharadwaj et al., 2025; Shi et al., 2025; Tan et al., 2025; Gu et al., 2026; Dunlap et al., 2025) and how to better align LLM judges to humans (Chen et al., 2024; Liu et al., 2024). While a few works investigate citation-related preferences—focusing either on biases in how LLMs use citations compared to humans (Ando and Harada, 2026; Algaba et al., 2025) or on building systems to embed citations in long-form responses (Huang et al., 2024; Zhang et al., 2025)—to the best of our knowledge, no prior work has attempted to explain the role of citations in preference data for both human and LLM judges.

Deep Research Applications While our analysis focuses on understanding preferences to better inform preference data collection, citations are also central to deep research applications, where complex systems produce long-form responses with citations for challenging tasks (Singh et al., 2025; Shao et al., 2026). Many deep research benchmarks have been proposed, including DeepResearch Bench (Du et al., 2026), ResearchRubrics (Sharma et al., 2026), SurGE (Survey Generation Evaluation) (Su et al., 2026), and AstaBench (Bragg et al., 2025), among others. Hwang et al. (2026) study the challenges in deep research system evaluation and provide practical suggestions for designing future meta-evaluations. Survey generation and literature review are prominent applications of deep research, presenting both

significant opportunities and challenges (Agarwal et al., 2025; Yang et al., 2026; Sahu et al., 2026). Understanding citation preferences and the underlying human preferences or LLM behaviors is therefore essential for enabling more meaningful evaluation of these systems.

7 Conclusion

This work has studied how human and LLM judges weigh citations in their assessment of response quality, controlling for other potential confounds. By studying human and LLM preferences on the same data, we are able to compare predictors' effects both between these two categories of judge and across different types of LLM judge.

Broadly, we find that human judges prefer diverse citations but fewer total citations, while LLM judges exhibit different citation-related preferences depending on the data and model. We observe cases where all LLM judges exhibit an identifiable bias (e.g., response length on SciArena), while humans show no reliable preference along the same dimension. Further, we observe cases where humans show certain rational behaviors (e.g., penalizing mismatched citations on ResearchQA) that LLM judges do not. For LLM judges, different training processes also seem to engender different preferences for certain features (e.g., response length).

Taken together, these observations contribute to our understanding of the role of citations in preference data from human and LLM judges. Based on these observatins, we also provide some suggestions for preference data collection: both human and LLM judges may benefit from more detailed instructions for assessing citation quality before and during the preference collection process—with potential aids including clear annotation rubrics and external tool access. Mixed-effects models can also be integrated into the pipeline to help ensure data quality. More generally, we hope this work will motivate more careful attention to cited sources throughout the preference data collection pipeline.

Limitations

Correlation vs. Causality We present results from the final mixed-effects models and emphasize that these results reflect correlations, not causal relationships. We do not claim that these features *causally* affect human or LLM preference. Importantly, however, features need not be causal in order for LLM judges to learn to rely on them.

Dataset. We focus exclusively on SciArena and ResearchQA to study human and LLM preferences, as these are the only two datasets we could find that satisfied our research needs (§2). As Movva et al. (2025) note, different datasets can yield different observations about human preferences, often depending on the specific queries used and annotators involved. However, we believe that our mixed-effects modeling approach is widely applicable and that, given the broad coverage of the datasets, many of our findings likely generalize.

Prompting. While we adopt the same prompt setup as SciArena, different prompts could lead to variations in observed model preferences. We test sensitivity to prompt variation in Appendix A, where results suggest that LLM judge decisions largely remain unchanged. Substantially different prompts could still produce different behavior, which we leave to future work.

Linearity Assumption. To enable interpretable analysis of the impact of citations on preferences, we include the fixed-effects predictors assuming linearity. Given the complex nature of human and AI decision-making, some variables may relate to the outcome in more complex ways—for example, a longer response may be preferred only up to a point, or the effect of citation diversity may depend on response length. To probe the latter possibility, we re-run the mixed-effects analysis on all final models with an added interaction between response length and citation diversity. Across all 10 final models (1 human and 4 LLM judges on 2 datasets), the interaction is significant only for the SciArena Llama judge. We encourage future work to explore potential nonlinear or interaction effects.

Ethics Statement

Since this work focuses on the interpretability of preference data and relies on existing public, anonymized datasets, we do not believe it raises any substantive ethical concerns. Both datasets used in our analysis are publicly available and contain no personally identifiable information. The LLM judges evaluated in this work are open-weight models accessed through HuggingFace, and our analysis is observational—aimed only at understanding citation-related preferences.

Acknowledgments

We are grateful to collaborators and visiting researchers at the HLTCOE at Johns Hopkins University for feedback on this work, including Dayeon (Zoey) Ki, Maxime Dassen, Roxana Petcu, Jia-Huei Ju, Siddharth Singh, Hannah Recknor, James Mayfield, Dawn Lawrie, Laura Dietz, Eugene Yang, Daniel Khashabi, and Kevin Duh—with particular thanks to Bryan Li and Gabrielle Kaili-May Liu for especially detailed input and comments. We would like to extend our thanks to the anonymous reviewers for their valuable feedback, and to Navita Goyal, Nishant Balepur, and Christabel Acquaye for their suggestions on the draft.

References

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy Dj Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025. Litllms, llms for literature review: Are we there yet? *Trans. Mach. Learn. Res.*, 2025.

H. Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. 2025. Large language models reflect human citation patterns with a heightened citation bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6829–6864, Albuquerque, New Mexico. Association for Computational Linguistics.

Kenichiro Ando and Tatsuya Harada. 2026. Aligning large language model behavior with human citation preferences. *Preprint*, arXiv:2602.05205.

Vincent Arel-Bundock, Noah Greifer, and Andrew Heiss. 2024. How to interpret statistical models using marginaleffects for R and Python. *Journal of Statistical Software*, 111(9):1–32.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Alejandro Benito-Santos, Adrian Ghajari, and Víctor Fresno. 2025. Robust estimation of population-level effects in repeated-measures NLP experimental designs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33076–33089, Vienna, Austria. Association for Computational Linguistics.

Anirudh Bharadwaj, Chaitanya Malaviya, Nitish Joshi, and Mark Yatskar. 2025. Flattery, fluff, and fog: Diagnosing and mitigating idiosyncratic biases in preference models. *Preprint*, arXiv:2506.05339.

Owen Bianchi, Maya Willey, Chelsea X Alvarado, Benjamin Danek, Marzieh Khani, Nicole Kuznetsov, Anant Dadu, Syed Shah, Mathew J Koretsky, Mary B Makarious, Cory Weller, Kristin S Levine, Sungwon Kim, Paige Jarreau, Dan Vitale, Elise Marsan, Hirotaka Iwaki, Hampton Leonard, Sara Bandres-Ciga, and 5 others. 2026. Cardbiomedbench: a benchmark for evaluating the performance of large language models in biomedical research. *The Lancet Digital Health*, 8(1):100943.

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

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, and 20 others. 2025. Astabench: Rigorous benchmarking of ai agents with a scientific research suite. *Preprint*, arXiv:2510.21652.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Shiping Chen, Duncan Brumby, and Anna Cox. 2025. Envisioning the future of peer review: Investigating llm-assisted reviewing using chatgpt as a case study. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, CHIWORK '25, New York, NY, USA. Association for Computing Machinery.

Wayne Chi, Valerie Chen, Anastasios Nikolas Angelopoulos, Wei-Lin Chiang, Aditya Mittal, Naman Jain, Tianjun Zhang, Ion Stoica, Chris Donahue, and Ameet Talwalkar. 2025. Copilot arena: A platform for code LLM evaluation in the wild. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. 2026. Deepresearch bench: A comprehensive benchmark for deep research agents. In *The Fourteenth International Conference on Learning Representations*.

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

Lisa Dunlap, Krishna Mandal, Trevor Darrell, Jacob Steinhardt, and Joseph E Gonzalez. 2025. Vibecheck: Discover and quantify qualitative differences in large language models. *International Conference on Learning Representations*.

Hanpei Fang, Sijie Tao, Nuo Chen, Kai-Xin Chang, and Tetsuya Sakai. 2025. Do large language models favor recent content? a study on recency bias in llm-based reranking. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2025, page 85–94, New York, NY, USA. Association for Computing Machinery.

Naghmeh Farzi and Laura Dietz. 2025. Criteria-based llm relevance judgments. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 254–263, New York, NY, USA. Association for Computing Machinery.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653, Doha, Qatar. Association for Computational Linguistics.

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert D. Mullins. 2025a. Inverse Constitutional AI: Compressing Preferences into Principles. In *The Thirteenth International Conference on Learning Representations*.

Arduin Findeis, Floris Weers, Guoli Yin, Ke Ye, Ruoming Pang, and Tom Gunter. 2025b. Can external validation tools improve annotation quality for LLM-as-a-judge? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15997–16020, Vienna, Austria. Association for Computational Linguistics.

William Gantt, Benjamin Kane, and Aaron Steven White. 2020. Natural language inference with mixed effects. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 81–87, Barcelona, Spain (Online). Association for Computational Linguistics.

Sebastian Gehrmann. 2025. Reward models are metrics in a trench coat. *Preprint*, arXiv:2510.03231.

Andrew Gelman and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*, volume Analytical methods for social research. Cambridge University Press, New York.

Andrew Gelman and Jennifer Hill. 2014. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York City.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on llm-as-a-judge. *The Innovation*, 7(6):101253.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. 2025. Explaining length bias in LLM-based preference evaluations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6763–6794, Suzhou, China. Association for Computational Linguistics.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024. Training language models to generate text with citations via fine-grained rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2926–2949, Bangkok, Thailand. Association for Computational Linguistics.

Jena D. Hwang, Varsha Kishore, Amanpreet Singh, Dany Haddad, Aakanksha Naik, Malachi Hamada, Jonathan Bragg, Mike D'Arcy, Daniel S. Weld, Lucy Lu Wang, Doug Downey, and Sergey Feldman. 2026. Deep research, shallow evaluation: A case study in meta-evaluation for long-form qa benchmarks. *Preprint*, arXiv:2603.06942.

Zhuoran Jin, Hongbang Yuan, Tianyi Men, Pengfei Cao, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2025. RAG-RewardBench: Benchmarking reward models in retrieval augmented generation for preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17061–17090, Vienna, Austria. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Weronika Łajewska, Krisztian Balog, Damiano Spina, and Johanne Trippas. 2024. Can users detect biases or factual errors in generated responses in conversational information-seeking? In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, page 92–102, New York, NY, USA. Association for Computing Machinery.

Shuyue Stella Li, Melanie Sclar, Hunter Lang, Ansong Ni, Jacqueline He, Puxin Xu, Andrew Cohen, Chan Young Park, Yulia Tsvetkov, and Asli Celikyilmaz. 2025. Prefpalette: Personalized preference modeling with latent attributes. In *Second Conference on Language Modeling*.

Gabrielle Kaili-May Liu, Gal Yona, Avi Caciularu, Idan Szpektor, Tim G. J. Rudner, and Arman Cohan. 2025. MetaFaith: Faithful natural language uncertainty expression in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29600–29644, Suzhou, China. Association for Computational Linguistics.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*.

Hannes Matuschek, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates. 2017. Balancing type i error and power in linear mixed models. *Journal of Memory and Language*, 94:305–315.

Richard McElreath. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.

Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N. Angelopoulos, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. 2025. Search arena: Analyzing search-augmented llms. *Preprint*, arXiv:2506.05334.

Rajiv Movva, Smitha Milli, Sewon Min, and Emma Pierson. 2025. What's in my human feedback? learning interpretable descriptions of preference data. *Preprint*, arXiv:2510.26202.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

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

Gaurav Sahu, Laurent Charlin, and Christopher Pal. 2026. Rethinking literature search evaluation: Deep research helps, and human citation lists are not a ground truth. *Preprint*, arXiv:2605.29234.

Sandra Sandoval, Jieyu Zhao, Marine Carpuat, and Hal Daumé III. 2023. A rose by any other name would not smell as sweet: Social bias in names mistranslation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3945, Singapore. Association for Computational Linguistics.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, and 2 others. 2026. DR tulu: Reinforcement learning with evolving rubrics for deep research. In *Forty-third International Conference on Machine Learning*.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. 2026. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. In *The Fourteenth International Conference on Learning Representations*.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Tu Shiwen, Zhao Liang, Chris Yuhao Liu, Liang Zeng, and Yang Liu. 2024. Skywork critic model series. `https://huggingface.co/Skywork`.

Amanpreet Singh, Joseph Chee Chang, Dany Haddad, Aakanksha Naik, Jena D. Hwang, Rodney Kinney, Daniel S Weld, Doug Downey, and Sergey Feldman. 2025. Ai2 scholar QA: Organized literature synthesis with attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 513–523, Vienna, Austria. Association for Computational Linguistics.

Weihang Su, Anzhe Xie, Qingyao Ai, Jianming Long, Xuanyi Chen, Jiaxin Mao, Ziyi Ye, and Yiqun Liu. 2026. Surge: A benchmark and evaluation framework for scientific survey generation. *Preprint*, arXiv:2508.15658.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca A. Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Jan Philip Wahle, Terry Lima Ruas, Mohamed Abdalla, Bela Gipp, and Saif M. Mohammad. 2025. Citation amnesia: On the recency bias of NLP and other academic fields. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1027–1044, Abu Dhabi, UAE. Association for Computational Linguistics.

Julia White, Noah Goodman, and Robert Hawkins. 2022. Mixed-effects transformers for hierarchical adaptation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3944–3954, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Richard Williams. 2012. Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal*, 12(2):308–331.

Zhaofeng Wu, Michihiro Yasunaga, Andrew Cohen, Yoon Kim, Asli Celikyilmaz, and Marjan Ghazvininejad. 2025. reWordBench: Benchmarking and improving the robustness of reward models with transformed inputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3383–3409, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ziyue Yang, Da Ma, Hanqi Li, Zijian Wang, Tiancheng Huang, Zijian Hu, Chenrun Wang, Yunzhe Zhang, Xiaobao Wu, Kai Yu, and Lu Chen. 2026. Deepsurvey: Enhancing analytical depth and citation reliability in automated survey generation. *Preprint*, arXiv:2605.29522.

Li S. Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. 2025. Researchqa: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *Preprint*, arXiv:2509.00496.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5098–5122, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Taira Anderson, Jonathan Bragg, Joseph Chee Chang, Jesse Dodge, Matt Latzke, Yixin Liu, Charles McGrady, Xiangru Tang, Zihang Wang, Chen Zhao, Hannaneh Hajishirzi, Doug Downey, and Arman Cohan. 2025a. Sciarena: An open evaluation platform for foundation models in scientific literature tasks. *Preprint*, arXiv:2507.01001.

Yulai Zhao, Haolin Liu, Dian Yu, Sunyuan Kung, Meijia Chen, Haitao Mi, and Dong Yu. 2025b. One token to fool llm-as-a-judge. *Preprint*, arXiv:2507.08794.

Huan Zhong and Zhengbiao Han. 2023. A systematic review of information source preference research. *Journal of Documentation*, 80(1):255–274.

Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. 2024. Explaining datasets in words: Statistical models with natural language parameters. In *Advances in Neural Information Processing Systems*, volume 37, pages 79350–79380. Curran Associates, Inc.

A Experiment Details

A.1 LLM Judge Experiment Setup

We set up the zero-shot experiments using the same prompt as in SciArena (Zhao et al., 2025a). Hugging Face model names are: `meta-llama/Llama-3.3-70B-Instruct`, `google/gemma-3-27b-it`, `Qwen/Qwen3-30B-A3B-Instruct-2507`, and `Skywork/Skywork-Critic-Llama-3.1-70B`.

We serve the LLMs using vLLM (Kwon et al., 2023), with 4 NVIDIA A100 GPUs for Llama and Skywork and 1 A100 for Gemma and Qwen. The model inference implementation is based on the SciArena public codebase (with temperature 0).[12]

A.2 Flipped Response Order Experiment

We include a response presentation order variable to control for potential positional bias in citation preferences. We use the same prompt template for the flipped-order experiment but switch the order of R_A and R_B as shown in the prompt below.

LLM Judge Prompt (Flipped Order)

```
You are an expert in scientific literature synthesis. Your task is to evaluate the quality of two AI-generated citation-attributed responses to a user's question. Assess both responses for relevance, accuracy, clarity, and appropriate use of citations. Then, select the response, Output (a) or Output (b), that best address the user's question.

User Question:
{question}

Output (a):
{response_b}

Output (b):
{response_a}

Which is best, Output(a) or Output (b)?
Answer with only: "A" or "B"
```

A.3 Prompt Sensitivity Analysis

To assess whether our findings depend on the specific judge prompt, we re-run all LLM judges (Qwen, Gemma, Llama, and Skywork) on SciArena using two variants of the original prompt: a chain-of-thought (CoT) variant, which asks the model to reason step by step before committing to a decision, and a length-debiasing variant (following Liu et al. (2025)), which adds an explicit instruction not to let response length influence the evaluation. Both variants are otherwise identical to the original prompt. As in the main experiments, each pair is judged under both presentation orders. We quantify stability as the agreement between each variant's decisions and those of the original prompt, using Cohen's κ alongside raw agreement.

LLM Judge Prompt (CoT)

```
You are an expert in scientific literature synthesis. Your task is to evaluate the quality of two AI-generated citation-attributed responses to a user's question. Assess both responses for relevance, accuracy, clarity, and appropriate use of citations. Then, select the response, Output (a) or Output (b), that best addresses the user's question.

User Question:
{question}

Output (a):
{response_a}

Output (b):
{response_b}

Which is best, Output (a) or Output (b)?
Let's think step by step, then give your final answer.
End your response with a line formatted exactly as:
Answer: <A or B>
```

LLM Judge Prompt (Length Debiasing)

```
You are an expert in scientific literature synthesis. Your task is to evaluate the quality of two AI-generated citation-attributed responses to a user's question. Assess both responses for relevance, accuracy, clarity, and appropriate use of citations. Then, select the response, Output (a) or Output (b), that best addresses the user's question.

User Question:
{question}

Output (a):
{response_a}

Output (b):
{response_b}

Which is best, Output (a) or Output (b)?
Do NOT allow the length of the responses to influence your evaluation.
Answer with only: "A" or "B"
```

Table 1 below reports the results. Overall, these results suggest that our judges' decisions are largely stable under prompt variation.

[12] https://github.com/yale-nlp/SciArena

	Qwen	Gemma	Llama	Skywork
CoT	.80 (90%)	.67 (84%)	.72 (88%)	.89 (95%)
LD	.96 (98%)	.93 (97%)	.93 (97%)	.98 (99%)

Table 1: Agreement between each prompt variant and the original prompt, reported as Cohen's κ with raw agreement in parentheses. LD: Length Debiasing.

B MEM Details & Model Selection

Although there is not yet consensus on how to select the random-effects design matrix, we follow the recommendation to "*keep it maximal*". This leverages the advantage of mixed-effects models over fixed-effects models: the ability to isolate complex group-level variance. We start with the maximal random effects structure feasible given our data groups, removing terms to ensure a non-singular fit, and further simplifying the random structure based on AIC.[13] Note that this approach is relatively costly because fitting the complex model requires more computation and likely over-specifies the possible data generation process.

Specifically, data groups in **SciArena** include: (1) models (M_A and M_B; $N = 23$); (2) query types ($|Q_{\text{type}}| = 6$); (3) query subjects ($|Q_{\text{subj}}| = 25$, which further belong to 5 main categories, Q_{cat}); and (4) model pairs ($|M_{AB}| = 480$). The six question types are: Conceptual Explanation, Challenges & Limitations, State-of-the-Art Assessment, Methodology Inquiry, Paper Finding, and Others. The five main categories include: Engineering, Natural Science, Healthcare, Humanities & Social Science, and Others. Further, for example, Computer Science, Electrical Engineering, and Civil Engineering are subjects under Engineering category. Readers can check SciArena (Zhao et al., 2025a) for full details.

When designing **random intercepts**, we start by considering the number of groups. It is usually suggested to include variables as random effects when the groups are drawn from a larger population with more than 5 levels. (Therefore, during this step, we select Q_{subj} rather than Q_{cat}.) For **interactions**, we hypothesize that different models might show different response quality based on the queries. To balance the number of observations available to estimate a group, we define interactions between $\{M_A, M_B\}$ and $\{Q_{\text{type}}, Q_{\text{cat}}\}$. For example, using $M_A : Q_{\text{subject}}$ leads to at most 23×25 groups, resulting in fewer observations per group, whereas using Q_{cat} results in at most 23×5 groups. As a result, $M : Q_{\text{type}}$ leads to 138 groups and $M : Q_{\text{cat}}$ leads to 115 groups. Furthermore, regarding **random slopes**, we include all continuous fixed-effects predictors where we suspect possible group-level variance. The maximal form includes intercepts, slopes, and their interactions (denoted with | in §2). The exception in our case is the Q_{type} groups, which have too few levels to estimate the parameters reliably. Therefore, we have to assume there is no interaction between the question type intercept and slopes in the maximal design, using the syntax ||. On the fixed-effects side, we retain all predictors during the process of simplifying the random effects, as these are the main effects we hope to observe. The final mixed-effects models:

$$
\begin{aligned}
&\text{(1) Human vote} \sim \\
&\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C \\
&+ (1|M_A) + (1|M_B) \\
&+ (1|M_B : Q_{\text{type}}) \\
&+ (1|M_A : Q_{\text{cat}}) + (1|M_B : Q_{\text{cat}}) \\
&+ (\Delta\text{C}_D + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C||M_{AB}) \\
&+ (0 + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C||Q_{\text{subj}}) \\
&+ (\Delta\text{C}_D + \Delta\text{R}_C||Q_{\text{type}})
\end{aligned}
$$

$$
\begin{aligned}
&\text{(2) Llama vote} \sim \\
&\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C + \text{order} \\
&+ (1|M_A) + (1|M_B) \\
&+ (1|M_A : Q_{\text{type}}) + (1|M_B : Q_{\text{type}}) \\
&+ (1|M_A : Q_{\text{cat}}) + (1|M_B : Q_{\text{cat}}) \\
&+ (\Delta\text{C}_D + \Delta\text{R}_L||M_{AB}) \\
&+ (\Delta\text{C}_C + \Delta\text{C}_D + \Delta\text{R}_C||Q_{\text{subj}}) \\
&+ (\Delta\text{C}_A + \Delta\text{R}_L||Q_{\text{type}})
\end{aligned}
$$

$$
\begin{aligned}
&\text{(3) Gemma vote} \sim \\
&\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C + \text{order} \\
&+ (1|M_A) + (1|M_B) \\
&+ (1|M_A : Q_{\text{type}}) \\
&+ (1|M_B : Q_{\text{cat}}) \\
&+ (\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L||M_{AB}) \\
&+ (\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{R}_L + \Delta\text{R}_C||Q_{\text{subj}}) \\
&+ (\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{R}_L + \Delta\text{R}_C||Q_{\text{type}})
\end{aligned}
$$

[13] https://www.rdocumentation.org/packages/lme4/versions/1.1-38/topics/isSingular

(4) Qwen vote $\sim$

$$\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C + \text{order}$$
$$+ (1|M_A) + (1|M_B)$$
$$+ (1|M_A : Q_{\text{type}}) + (1|M_B : Q_{\text{type}})$$
$$+ (1|M_A : Q_{\text{cat}}) + (1|M_B : Q_{\text{cat}})$$
$$+ (\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{R}_L + \Delta\text{R}_C || M_{AB})$$
$$+ (\Delta\text{C}_D + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C || Q_{\text{subj}})$$
$$+ (\Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C || Q_{\text{type}})$$

(5) Skywork vote $\sim$

$$\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C + \text{order}$$
$$+ (1|M_A) + (1|M_B)$$
$$+ (1|M_A : Q_{\text{type}}) + (1|M_B : Q_{\text{type}})$$
$$+ (1|M_A : Q_{\text{cat}}) + (1|M_B : Q_{\text{cat}})$$
$$+ (\Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_C || M_{AB})$$
$$+ (\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_A + \Delta\text{R}_L + \Delta\text{R}_C$$
$$|| Q_{\text{subj}})$$
$$+ (\Delta\text{C}_D + \Delta\text{C}_A + \Delta\text{R}_C || Q_{\text{type}})$$

For **ResearchQA**, because we have less data groups (e.g., responses are only from two models: `gpt-4.1-mini` and `gemini-2.5-flash` and only 169 observations, the maximal form is: $(1|Q_{\text{field}})$, where they are in total 8 fields. The final mixed-effects model for human preference is:

Human vote $\sim$

$$\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_M + \Delta\text{R}_L + \Delta\text{R}_C + (1|Q_{\text{field}})$$

The final MEM for LLM judge preference is:

Llama / Gemma / Qwen / Skywork vote $\sim$

$$\Delta\text{C}_D + \Delta\text{C}_C + \Delta\text{C}_M + \Delta\text{R}_L + \Delta\text{R}_C + \text{order}$$

The estimated fixed effects of the final model on human votes and LLM judgments are reported in Table 2 (SciArena) and Table 3 (ResearchQA).

C Usage of Large Language Models

We use LLMs to support and refine the writing of our work. We do not rely on them to generate content or sentences from scratch and take full responsibility for the final outcome. In addition, we use AI to help brainstorm figure visualizations and assist with coding when necessary.

Judge	Predictor	ΔWin Rate	95% CI	p
Human	Citation Diversity	+1.78*	[+0.06, +3.51]	0.043
	Citation Count	-3.55***	[-5.34, -1.75]	<0.001
	Citation Age	+0.66	[-0.33, +1.66]	0.193
	Response Length	-1.21	[-3.40, +0.99]	0.283
	Response Content	+4.21*	[+1.00, +7.43]	0.010
Llama	Citation Diversity	+7.47***	[+5.55, +9.39]	<0.001
	Citation Count	-2.07***	[-3.29, -0.85]	<0.001
	Citation Age	-0.35	[-1.21, +0.50]	0.417
	Response Length	+10.13***	[+7.61, +12.64]	<0.001
	Response Content	+11.30***	[+8.61, +13.98]	<0.001
Gemma	Citation Diversity	+5.15***	[+3.12, +7.19]	<0.001
	Citation Count	+1.41	[-0.27, +3.09]	0.100
	Citation Age	-0.37	[-0.98, +0.24]	0.237
	Response Length	+11.07***	[+8.35, +13.80]	<0.001
	Response Content	+13.52***	[+9.89, +17.15]	<0.001
Qwen	Citation Diversity	+6.18***	[+4.65, +7.71]	<0.001
	Citation Count	-0.44	[-1.73, +0.86]	0.510
	Citation Age	-0.52	[-1.69, +0.64]	0.380
	Response Length	+9.35***	[+7.26, +11.45]	<0.001
	Response Content	+18.00***	[+14.50, +21.50]	<0.001
Skywork	Citation Diversity	+7.04***	[+4.55, +9.54]	<0.001
	Citation Count	-12.82***	[-15.83, -9.81]	<0.001
	Citation Age	-0.59	[-1.84, +0.67]	0.359
	Response Length	-0.04	[-1.50, +1.43]	0.960
	Response Content	+9.17***	[+6.33, +12.01]	<0.001

Table 2: Estimated fixed effects on human votes and LLM judgments for **SciArena**, reported as changes in win rate (percentage points) for a one standard deviation increase in each standardized predictor. This table reports the exact estimates underlying Figure 3. Significance: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Judge	Predictor	ΔWin Rate	95% CI	p
Human	Citation Diversity	+11.65***	[+5.31, +17.99]	<0.001
	Citation Count	-6.47*	[-12.88, -0.05]	0.048
	Citation Mismatch	-10.08**	[-17.01, -3.15]	0.004
	Response Length	+21.41***	[+13.86, +28.96]	<0.001
	Response Content	-0.22	[-8.66, +8.22]	0.959
Llama	Citation Diversity	+7.76***	[+3.99, +11.54]	<0.001
	Citation Count	+0.67	[-3.05, +4.39]	0.722
	Citation Mismatch	+0.07	[-3.82, +3.95]	0.973
	Response Length	+9.10***	[+4.36, +13.84]	<0.001
	Response Content	+1.07	[-3.68, +5.83]	0.659
Gemma	Citation Diversity	+8.82***	[+4.62, +13.03]	<0.001
	Citation Count	+1.53	[-2.66, +5.73]	0.474
	Citation Mismatch	-2.33	[-5.99, +1.32]	0.211
	Response Length	+11.16***	[+6.05, +16.27]	<0.001
	Response Content	+2.76	[-2.53, +8.04]	0.307
Qwen	Citation Diversity	+4.35	[-0.30, +8.99]	0.067
	Citation Count	-0.66	[-5.35, +4.03]	0.781
	Citation Mismatch	+0.44	[-3.63, +4.52]	0.832
	Response Length	+5.87*	[+0.24, +11.49]	0.041
	Response Content	+7.23*	[+1.53, +12.92]	0.013
Skywork	Citation Diversity	+0.34	[-2.85, +3.53]	0.833
	Citation Count	-2.25	[-5.66, +1.16]	0.195
	Citation Mismatch	+1.18	[-1.22, +3.58]	0.336
	Response Length	+1.95	[-1.97, +5.88]	0.330
	Response Content	+0.66	[-3.20, +4.53]	0.737

Table 3: Estimated fixed effects on human votes and LLM judgments for **ResearchQA**, reported as changes in win rate (percentage points) for a one standard deviation increase in each standardized predictor. This table reports the exact estimates underlying Figure 4. Significance: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.