Debiasing as a Measurement Intervention: Calibrated Ties and Resolution Loss in LLM-as-a-Judge Evaluation

Anonymous EMNLP submission

Abstract

LLM-as-a-judge protocols are commonly debiased by instructing judges to ignore presentation cues such as citation formatting, source labels, and evidence-display style. We show that this intervention can suppress bias while damaging the resolution of the measurement instrument. We introduce TraceJudgeBench, a diagnostic benchmark for auditing citation-like artifacts in RAG and agent-workflow evaluation, covering content-equivalent pairs, citation ablations, correctness conflicts, human-validated soft and moderate quality gaps, prompt-strength ladders, decoupled judging, and a controlled workflow-ranking probe. Across GPT-5.5, Claude Sonnet 4.6, and DeepSeek V4-Flash, stronger anti-citation prompts reduce worse-cited wins from up to 50.5% to 0%; yet some operating points already convert validated moderate-gap decisions into Tie before the strict stress-test endpoint, while correctness-conflict accuracy remains at or above 93.0%. Open-weight Qwen2.5-14B-Instruct-AWQ and Gemma-3-12B-IT runs reproduce the central moderate-gap frontier, with TRACE-style decoupling recovering 96.5–98.5% better-plain resolution. Human validation separates three meanings of Tie: correct equivalence Tie, calibrated soft-boundary Tie, and resolution-destroying Tie on validated quality gaps. We frame debiasing as a measurement intervention whose bias suppression, resolution retention, Tie cost, and protocol cost must be reported jointly. We include TraceJudgeBench as a supplementary artifact with anonymized splits, prompts, raw judge outputs, validation summaries, and analysis scripts.

1 Introduction

LLM-as-a-judge evaluation has become a standard measurement layer for open-ended generation, instruction following, retrieval-augmented generation (RAG), agent workflows, and arena-style model comparison. Its appeal is clear: LLM judges are scalable, inexpensive relative to human annotation, and applicable to outputs whose quality cannot be captured by exact match or lexical overlap. However, this scalability also makes the judge itself part of the measurement system. When a judge is sensitive to presentation, order, verbosity, model identity, or output format, the resulting score no longer reflects only the underlying quality of the candidates.

This problem becomes especially sharp in RAG and agent-workflow evaluation. Modern systems rarely produce answer-only text; they typically emit source titles, citation markers, sentence identifiers, retrieved snippets, tool traces, verification statements, and other evidence-display artifacts (Bragg et al., 2026; Chong et al., 2026; Molfese et al., 2026). These artifacts have a dual status. In some tasks, they are superficial wrappers that should not influence the judgment when the substantive answer is unchanged. In others, they are part of the evaluated behavior, because traceability, support-chain completeness, and evidence use are legitimate dimensions of quality. A judge protocol must therefore decide whether citation-like artifacts are irrelevant presentation cues or useful evidence of task performance.

A common remedy is to debias the judge with a stricter rubric: instruct it to ignore citation formatting, source labels, evidence-display style, length, and polish, and to focus only on correctness or substantive content. This intervention is often warranted. If two candidates are content-equivalent and differ only in citation-looking wrappers, a strict judge should return Tie. In this setting, Tie is not indecision; it is the desired suppression of a spurious preference.

Debiasing, however, is not merely bias removal. It is a measurement intervention that can change what the benchmark is able to distinguish. Many evaluation settings require more than separating clearly correct from clearly wrong answers. They

also require distinguishing candidates that share the same final answer but differ in support-chain completeness, reasoning specificity, evidence coverage, or workflow faithfulness. In this region, a strict anti-presentation rubric may remove useful quality resolution together with superficial presentation sensitivity. The hidden cost is not necessarily a collapse in clear correctness. Instead, the measurement signal can shift from biased preference to abstention: a judge that once over-rewarded citation-like display may become unable to select the better answer when the quality gap is real but moderate.

We study this trade-off through the lens of calibrated Tie versus resolution-destroying Tie. Tie is not a single phenomenon. On content-equivalent pairs, Tie is correct. On soft-boundary pairs where human annotators also disagree or mostly choose Tie, Tie may be a calibrated expression of uncertainty. But on human-validated moderate-quality gaps, Tie is costly: the protocol has lost benchmark resolution. This distinction is essential because debiasing methods that look successful under equivalence tests may still fail as measurement tools for quality-sensitive evaluation.

We introduce TraceJudgeBench, a diagnostic benchmark for auditing whether LLM-as-a-judge protocols preserve decision resolution after suppressing citation-like presentation cues. We run the full diagnostic matrix on three commercial judges: GPT-5.5, Claude Sonnet 4.6, and DeepSeek V4-Flash. We additionally run the central moderate-gap prompt-frontier and TRACE-recovery audits on two open-weight judges. The open-weight runs do not cover every split, but they reproduce the main protocol-level pattern: stricter anti-presentation instructions can trade superficial-win suppression for Tie inflation, while quality-preserving decoupling recovers moderate-gap resolution.

Figure 1 summarizes TraceJudgeBench as a protocol-audit framework: input pairs are organized by target relation, evaluated under prompt and protocol families, and reported through a bias–resolution frontier rather than a single aggregate score.

Our contributions are:

1. We frame debiasing in LLM-as-a-judge evaluation as a measurement intervention, not merely as bias removal.

2. We introduce TraceJudgeBench, a diagnostic benchmark for auditing citation-like artifacts and debiasing side effects in RAG and agent-workflow judging.

3. We identify debiasing-induced resolution loss, in which stricter anti-presentation rubrics suppress citation-like wins but convert human-validated moderate quality gaps into Tie.

4. We distinguish calibrated Tie from resolution-destroying Tie using content-equivalent, soft-boundary, and moderate-gap human validation tiers.

5. We show that the central bias–resolution frontier appears across commercial and open-weight judges, while reporting the full diagnostic matrix for the three commercial judges.

2 Related Work

LLM-as-a-judge evaluation and bias. Pairwise LLM judging is widely used for open-ended model evaluation, preference comparison, reward-model assessment, and arena-style ranking. MT-Bench and Chatbot Arena established LLM judges as a scalable alternative to human preference collection, while documenting limitations such as position bias, verbosity bias, self-enhancement, and constrained reasoning (Zheng et al., 2023). AlpacaEval and Length-Controlled AlpacaEval show that automatic evaluators can reward superficial correlates of quality, especially length, unless the protocol is explicitly debiased (Dubois et al., 2024). RewardBench broadens this line by evaluating reward models and LLM judges across heterogeneous preference settings (Lambert et al., 2025). Other work treats the judge itself as an object of evaluation, studying position, verbosity, self-enhancement, output-format bias, hidden preference leakage, and automatically discovered judge biases (Wang et al., 2024; Long et al., 2025; Lai et al., 2026; Li et al., 2026). Our work builds on this literature but asks a different question: after a bias is mitigated, does the protocol still preserve the resolution needed for quality-sensitive evaluation?

Evidence display and RAG evaluation. RAG evaluation has made evidence use a central evaluation target. Frameworks such as RAGAS and ARES evaluate faithfulness, answer relevance, and context quality, reflecting the fact that answer correctness alone does not fully capture grounded-generation quality (Es et al., 2024; Saad-Falcon

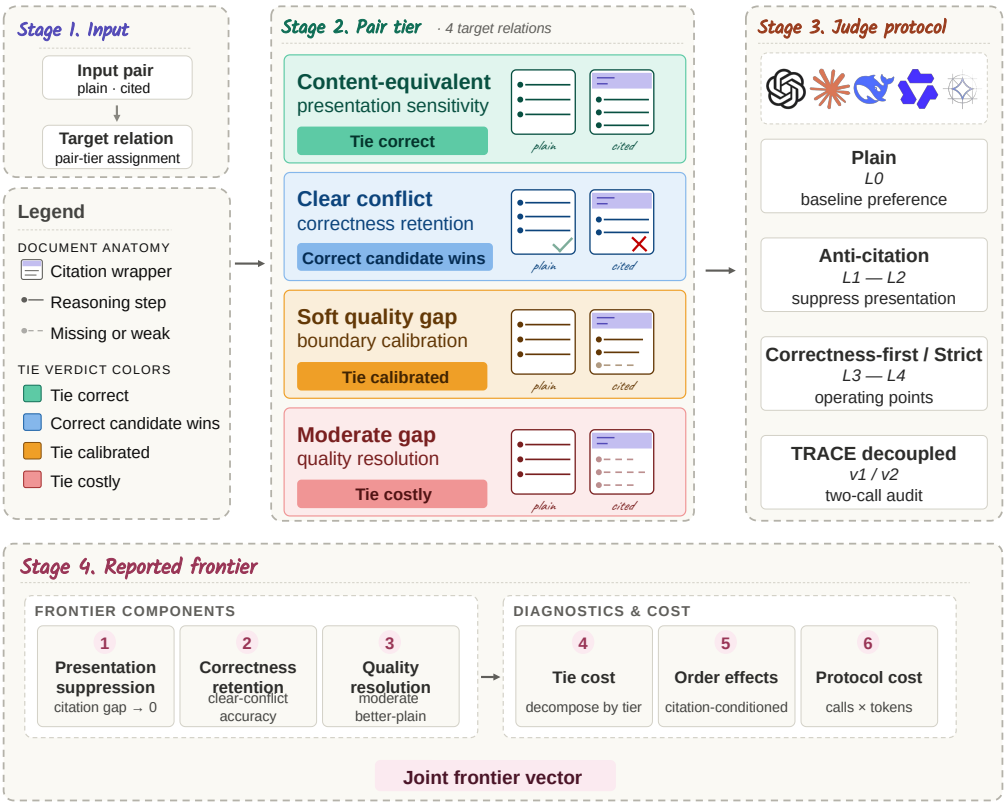

Figure 1: TraceJudgeBench protocol-audit framework

et al., 2024). Attribution work makes a related point: answer correctness and citation faithfulness are not equivalent, since an answer can be correct while its cited evidence is incomplete, irrelevant, or unsupported (Wallat et al., 2025). Citation-like artifacts therefore have a dual status. When the target is only final-answer correctness, source labels or citation formatting may be confounders. When the target includes traceability, evidence coverage, or auditability, the same artifacts may carry task-relevant information. We adopt a controlled-comparison approach in the spirit of work that holds content constant while varying representation (Zhang et al., 2026a): hold or validate substantive quality while varying citation-like presentation, then test whether debiasing preserves useful resolution.

Resolution, abstention, and benchmark separability. Reliable evaluation requires not only aggregate accuracy, but also the ability to distinguish meaningful differences. Work on micro-benchmarking, item-response theory, and rank-aware evaluation shows that benchmarks can preserve aggregate trends while failing to separate close systems or saturated item regions (Yauney et al., 2026; Zhou et al., 2026; Zhang et al., 2026b). Tie and abstention are also not inherently failures: abstention can be desirable when inputs are ambiguous or genuinely uncertain (Zhai et al., 2026; Noh et al., 2026). TraceJudgeBench connects these points at the protocol level. We ask whether a debiasing prompt suppresses spurious presentation preference while preserving validated quality gaps, and we interpret Tie relative to the target relation of each pair rather than treating all Tie as either good calibration or bad indecision.

3 Problem Setup and Metrics

We study pairwise LLM-as-a-judge evaluation. Given a task x and two candidate outputs y_A and y_B, a judge protocol P returns one of three labels: A, B, or Tie. Each comparison is run under an ordered presentation of the two candidates, so that position effects can be measured rather than assumed away.

Our focus is on pairs in which one output carries a citation-like evidence-display artifact, such as source titles, citation markers, sentence identifiers, retrieved snippets, evidence blocks, or verification-looking statements. We call this output the **cited candidate**, and the other output the **plain** or **base**

candidate. The artifact may be purely presentational, as in content-equivalent pairs, or it may coexist with real differences in support-chain completeness, reasoning specificity, or evidence coverage.

We evaluate judge protocols under three target relations. In **content-equivalent pairs**, the two candidates have the same substantive answer and differ only in citation-like presentation; the desired outcome is Tie. In **clear correctness conflicts**, one candidate is correct and the other is wrong; the desired outcome is to select the correct candidate regardless of citation display. In **graded quality conflicts**, both candidates share the correct final answer but differ in support-chain completeness or reasoning specificity; these pairs test whether a judge preserves resolution on real but moderate quality gaps.

On equivalence or near-equivalence pairs, we report the cited-candidate win rate W_{cited}, plain/base win rate W_{plain}, and Tie rate T. We also report the signed presentation effect:

$$\Delta_{\text{citation}} = W_{\text{cited}} - W_{\text{plain}}.$$

A positive value means citation-like display raises the probability of winning; a value near zero with high Tie indicates successful suppression of presentation preference. On correctness-conflict pairs, we report correct-candidate wins, wrong-candidate wins, and Tie. On moderate graded pairs, we report better-plain wins, worse-cited wins, and Tie, and summarize directional resolution with:

$$\Delta_{\text{quality}} = W_{\text{better-plain}} - W_{\text{worse-cited}}.$$

Tie is interpreted by tier. Equivalence Tie measures bias suppression, soft-boundary Tie may indicate calibration, and moderate-gap Tie measures lost resolution. We also report candidate-order gaps and protocol cost. Single-call judge protocols require one judge call per ordered comparison; TRACE-style decoupled judging requires two calls, one normalization call and one masked judging call. For signed effects and order gaps, we use pair-level bootstrap confidence intervals with 2,000 resamples; for proportions near 0% or 100%, we use Wilson score intervals. The end-to-end audit flow—pair construction, target-relation tagging, protocol selection, and frontier reporting—is summarized in Figure 1.

4 TraceJudgeBench

TraceJudgeBench is a protocol-audit benchmark, not a universal RAG leaderboard. Its goal is to test whether LLM-as-a-judge protocols suppress citation-like presentation sensitivity while preserving the resolution needed for quality-sensitive evaluation. Each split isolates one measurement question, such as presentation sensitivity, naturalistic validity, artifact triggers, correctness retention, quality resolution, order effects, or protocol cost.

Table 1 lists the seven diagnostic components, their sizes, target relations, and judge coverage. Citation-like artifacts include source titles, citation markers, sentence identifiers, retrieved snippets, evidence blocks, and verification-looking statements. In some splits, they are superficial wrappers; in others, they coexist with real support-chain differences and test whether a judge can separate presentation from evidence quality.

We construct 200 HotpotQA mechanical citation pairs and 200 FinQA mechanical citation pairs. In each pair, the cited candidate preserves the same substantive content as the plain candidate but adds citation-like presentation, such as source labels or evidence-display blocks. Any systematic preference for the cited candidate therefore indicates presentation sensitivity rather than improved answer quality. To complement these artificial but causally controlled examples, we add 1,000 natural HotpotQA citation-looking pairs whose final answers match and whose outputs differ naturally in source labeling or evidence display.

Mechanical pairs are further expanded into six ablation variants: full citation, empty shell, source-only, evidence-only, sentence-id-only, and length-control. These variants separate source-like authority, raw evidence text, formatting shell, sentence identifiers, and added length. If judges merely prefer longer answers, the length-control variant should win; if they respond to source-like authority or traceability display, the full-citation and source-only variants should be strongest.

Correctness-conflict probes test whether citation-like display overwhelms obvious correctness. In each clear conflict, the plain candidate keeps the correct answer, while the cited candidate receives citation-like presentation but has its final answer deliberately changed to an incorrect one. Reverse conflicts swap the direction, making the cited candidate correct and the plain candidate wrong. We also add position-only controls with identical plain/no-

Component	Size	Target relation	Measurement question	Judge coverage
Mechanical equivalence	200 HotpotQA + 200 FinQA pairs	content-equivalent	Does citation-like display create spurious preference?	commercial full matrix
Natural citation-looking pairs	1,000 HotpotQA pairs; 200 human-validated	near-equivalent	Does the effect persist in more natural outputs?	commercial plain full; strict on 200-pair subset
Citation ablations	six variants	content-controlled	Which artifact triggers the preference?	commercial full matrix
Position controls	1,200 Claude + 400 DeepSeek comparisons	identical candidates	Is the effect generic position bias or citation-conditioned order coupling?	targeted commercial controls
Clear and reverse conflicts	150 pairs, run across prompt paraphrases and orders	correctness conflict	Does citation display override obvious correctness?	commercial plain/strict prompt paraphrases; decoupled single prompt
Soft and moderate graded conflicts	100 soft + 100 moderate pairs	quality conflict	Does debiasing preserve validated quality gaps?	commercial full; open-weight Qwen/Gemma full moderate frontier
Workflow-ranking probe	200 HotpotQA queries, four workflow variants	system comparison	Can presentation sensitivity affect controlled tournament scores?	commercial + open-weight controlled probe, not full leaderboard

Table 1: TraceJudgeBench diagnostic components

citation candidates to separate generic position bias from citation-conditioned order coupling.

The central resolution test uses two 100-pair HotpotQA graded tiers. In the **moderate tier**, both candidates give the same correct final answer, but the plain candidate provides a complete evidence or reasoning chain while the cited candidate omits or weakens one necessary supporting hop, entity link, or evidence-chain step. In the **soft tier**, both candidates match in final answer and main evidence, but the plain candidate is somewhat more specific while the cited candidate is more generic. The soft tier is intentionally near the human decision boundary.

Two independent annotators validate the natural and graded samples. Table 2 summarizes the validation signals. The moderate tier is the main human-validated basis for the resolution-loss claim: both annotators label 91/100 pairs as better-plain, and raw agreement is 100.0%. By contrast, only 1/100 soft pairs are labeled better-plain by both annotators, confirming that soft is a boundary tier rather than a stable superiority tier. We report PABAK and Gwet's AC1 alongside raw agreement instead of Cohen's kappa, which is known to collapse on prevalence-skewed agreement (Cicchetti and Feinstein, 1990).

We compare single-call judge protocols along a prompt-strength ladder: L0 plain preference, L1 anti-citation warning, L2 quality-ignore-citation, L3 correctness-first, and L4 strict tie-required. L4 is treated as a stress-test endpoint rather than a recommended deployment rubric; the main frontier is interpreted from the full L0–L4 trajectory, especially the L2/L3 operating points that resemble common debiasing instructions. We also test TRACE-style decoupled judging. TRACE first normalizes away citation/source presentation while preserving final answer, support-chain completeness, reasoning specificity, unsupported claims, omitted support, and partial-support signals; it then performs masked pairwise judging over the normalized records. TRACE is not a new model or a deployment-ready judge. It is a two-call audit baseline that makes presentation suppression and quality-resolution retention more auditable.

We evaluate the full matrix on GPT-5.5, Claude Sonnet 4.6, and DeepSeek V4-Flash, using temperature 0 and JSON-mode outputs where supported. We additionally run the central moderate-gap prompt frontier, TRACE recovery, and controlled ranking probes on two open-weight judges, Qwen2.5-14B-Instruct-AWQ and Gemma-3-12B-IT (full identifier and serving details in Appendix D). The open-weight experiments test whether the main frontier generalizes, but they are

Validation set	n	Raw agreement	PABAK	Gwet's AC1	Main human-label signal
Natural equivalence	200	92.0%	0.840	0.918	183/200 both equivalent
Graded soft	100	96.0%	0.920	0.958	1/100 both better plain
Graded moderate	100	100.0%	1.000	1.000	91/100 both better plain
Graded overall	200	98.0%	0.960	0.976	moderate is resolution-loss evidence

Table 2: Human validation summary

not presented as a full open-weight audit.

5 Results

We organize the results around presentation sensitivity, correctness retention, quality resolution, open-weight replication, and TRACE recovery.

5.1 Citation-Like Display Creates Spurious Preference

Under plain judging, citation-like presentation can substantially affect pairwise decisions even when substantive content is unchanged. On mechanical equivalence pairs, GPT-5.5 selects the cited candidate 85.8% of the time on HotpotQA, with Claude and DeepSeek at 50.9% and 47.2% (Figure 2). The presentation effect also appears in the soft graded tier: GPT-5.5 selects the worse-cited candidate in 50.5% of cases under plain judging—a preference that anti-citation prompts later suppress to 0%. Natural citation-looking pairs show the same phenomenon under less artificial conditions: source labels, evidence blocks, and verification-looking statements can be treated as signals of support even when the underlying answer is equivalent or near-equivalent.

Citation ablations show that the trigger is not simply length. Length-control and evidence-only variants do not account for the full effect. Instead, full-citation and source-only variants are strongest, suggesting that judges respond to source-like authority or traceability display rather than to additional tokens alone. Rationale analysis (Appendix G, Table A19) supports this interpretation: citation-win rationales frequently describe the cited candidate as more supported, verifiable, or traceable, with source/citation keyword rates near 100% on citation-win decisions across all three commercial judges.

Candidate order also matters: position-only controls separate generic order bias from citation-conditioned order coupling (Appendix Table A15).

5.2 Strict Rubrics Suppress Presentation Bias While Retaining Clear-Conflict Accuracy

Stronger anti-citation prompts successfully suppress presentation preference in equivalence settings. As the prompt ladder moves from L0 plain preference to L4 strict tie-required judging, worse-cited wins fall to 0% in the commercial judges. This is the intended outcome on content-equivalent pairs: the judge should not reward citation-looking wrappers when the substantive answer is unchanged.

Importantly, this suppression does not imply a general failure to judge correctness. On clear and reverse correctness conflicts, commercial judges mostly retain high accuracy under stricter prompts. Clear-conflict probes remain at or above 93.0% across the reported commercial settings; full strict-prompt correctness-family results are included in the anonymized release artifact. This separates two failure modes. The problem is not that debiasing makes judges unable to recognize obvious wrong answers. Rather, the risk appears in a harder region: pairs where both candidates share the correct final answer but differ in support-chain completeness or reasoning specificity.

Figure 2 visualizes this first-stage success: strong anti-citation prompting suppresses high cited-candidate win rates and moves the decision mass toward Tie. If the audit stopped here, the debiasing intervention would appear successful.

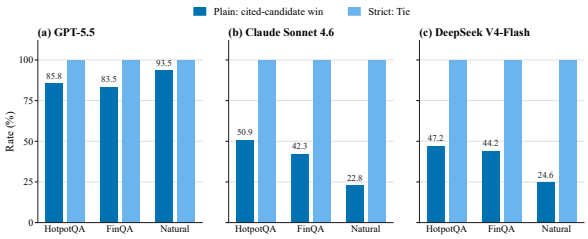

Figure 2: Equivalence suppression

5.3 Moderate Quality Gaps Reveal Resolution Loss

The moderate graded tier exposes the hidden cost of strict debiasing. In this tier, both candidates share the correct final answer, but the plain candidate has a complete support chain while the cited candidate omits or weakens a necessary supporting step. Human validation establishes that this is a real quality gap rather than a formatting preference.

For abstention-heavy judges, resolution loss begins before the strict stress-test endpoint. At L3, Claude's better-plain rate falls to 71.5% and DeepSeek's to 79.0%, while Tie rises to 28.5% and 21.0%, respectively; Gemma-3-12B-IT shows an even sharper L3 Tie jump in the open-weight replication. L4 exposes the endpoint behavior: Claude and DeepSeek drop to 11.0–15.0% better-plain decisions, and GPT-5.5 falls to 2.0%. Worse-cited wins remain near zero, so the failure is not a return of citation preference. It is resolution loss.

The soft tier explains why Tie cannot be interpreted uniformly. In soft cases, human annotators rarely agree that plain is better; Tie is plausibly calibrated. Appendix Table A18 shows worse-cited wins falling to 0% by L1 while Tie rises toward 100%. In moderate cases, by contrast, Tie erases a validated better-plain distinction. This tiered interpretation is why TraceJudgeBench reports Tie by target relation rather than only aggregate Tie rates.

5.4 The Bias–Resolution Frontier Appears Across Commercial and Open-Weight Judges

Figure 3 summarizes the central trade-off. Moving along the prompt ladder can suppress worse-cited wins, but for some judges it also shifts decision mass into Tie and reduces better-plain resolution on moderate gaps. We therefore do not rely on L4 alone: the frontier is the trajectory across L1–L3 operating points, the L4 stress test, and TRACE-style decoupling. A protocol improves the frontier only when it suppresses spurious citation preference without destroying validated quality resolution.

The open-weight judges are not run on the full diagnostic matrix, but they reproduce the central frontier. Qwen2.5-14B-Instruct-AWQ retains more moderate-gap resolution under strict prompting than the most abstention-heavy commercial judges, while Gemma-3-12B-IT shows a sharp Tie jump at the L2→L3 transition, although its L4 endpoint partially recovers better-plain decisions. These differences matter: the frontier is not a single universal curve. However, both open-weight judges support the main protocol-level claim that debiasing can trade presentation suppression for resolution loss.

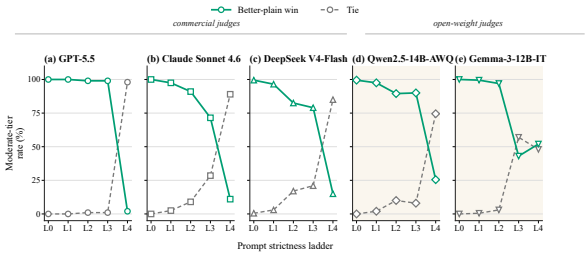

Figure 3: Bias-resolution frontier over prompt strictness

5.5 TRACE-Style Decoupling Recovers Moderate-Gap Resolution

TRACE-style decoupling shows that resolution loss is not inevitable. The correctness-only decoupled variant removes presentation artifacts but can still discard quality-relevant support-chain structure. The quality-preserving variant, `decoupled_quality_v2`, explicitly preserves final answer, support-chain completeness, reasoning specificity, unsupported claims, omitted support, and partial-support signals before masked judging.

This distinction matters. On the moderate tier, `decoupled_quality_v2` recovers or preserves better-plain resolution across all five judges included in the frontier analysis. Commercial judges recover to 97.0–100.0% better-plain decisions, and the two open-weight judges recover 96.5–98.5% better-plain resolution. Worse-cited wins remain at 0%. Thus, the moderate-gap failure is not an unavoidable consequence of ignoring citation-like presentation; it depends on whether the protocol preserves quality-relevant evidence structure while removing display cues. The full L0–L4 ladder appears in Appendix Table A11, with TRACE per-judge details in Table A12.

5.6 Controlled Ranking Shows System-Level Risk

Finally, we test whether presentation sensitivity can affect controlled system comparison. We construct a workflow-ranking probe over 200 HotpotQA queries and four workflow variants. The goal is not to produce a new RAG leaderboard, but to test whether citation-like display can alter tournament-style rankings when substantive differences are controlled.

The ranking probe confirms the system-level risk. Citation-like display can improve a variant's apparent standing under ordinary judging, while stricter or decoupled protocols reduce this advantage (full per-protocol scores and direct full-vs-full comparisons in Appendix E.8–E.9, Tables A16 and A17). This result connects the pairwise diagnostics to practical evaluation: if presentation artifacts influence pairwise wins, they can also affect aggregate rankings, model selection, and reported progress.

6 Discussion

Our results suggest that debiasing an LLM judge should be treated as a measurement intervention rather than a purely corrective prompt edit. A stricter instruction can remove an unwanted presentation preference, but it can also change the operating point of the evaluator. The key question is therefore not only whether a protocol reduces bias, but what kind of measurement signal remains after the bias is removed.

First, bias suppression is not sufficient. A protocol that suppresses citation preference on content-equivalent pairs may look successful under a narrow bias test, but still fail as a measurement instrument if it converts human-validated moderate differences into Tie. Clear-conflict accuracy alone is also insufficient as a safety check. A judge can identify obvious correct-versus-wrong cases while failing to distinguish moderate support-chain differences.

Second, Tie has multiple meanings. On content-equivalent pairs, Tie is a success because it suppresses a spurious preference. On soft-boundary pairs, Tie may be calibrated because human annotators also do not reliably prefer one candidate. On human-validated moderate gaps, however, Tie is a cost: it means the protocol failed to preserve a distinction that humans judged to be real. Debiasing studies should therefore report Tie by tier rather than only aggregate win rates.

Third, citation-like artifacts should be treated neither as automatically valid evidence nor as automatically irrelevant formatting. If the evaluation target is final-answer correctness only, citation-like display should usually be controlled or ignored. If the target includes traceability, evidence use, or workflow faithfulness, the protocol must preserve those quality dimensions while preventing source-like presentation from acting as a shortcut.

TRACE demonstrates that the failure mode is protocol-dependent. It requires additional calls and may introduce extraction errors, so it should be understood as an audit protocol rather than a universal deployment solution.

7 Conclusion

We introduced TraceJudgeBench, a diagnostic benchmark for auditing how LLM-as-a-judge protocols handle citation-like artifacts in RAG and agent-workflow evaluation. Our results show that stronger anti-citation instructions can successfully suppress spurious citation preference on content-equivalent pairs, but this success can hide resolution loss on human-validated moderate quality gaps. Tie is therefore not a single outcome: it is correct on equivalence pairs, potentially calibrated on soft-boundary pairs, and costly on validated moderate gaps. TRACE-style decoupling recovers much of the lost resolution by removing citation-like presentation while preserving quality-relevant support structure. Overall, debiasing should be reported as a frontier over presentation suppression, correctness retention, quality resolution, Tie cost, order effects, and protocol cost.

Limitations

TraceJudgeBench is a diagnostic benchmark, not a universal RAG or agent leaderboard. Its splits isolate measurement questions, which improves interpretability but does not cover the full diversity of real retrieval pipelines, agent tools, or user-facing domains (Seo et al., 2026; Shu et al., 2026; Dong et al., 2025). The full diagnostic matrix is evaluated on three commercial judges; open-weight experiments focus on the central moderate-gap frontier, TRACE recovery, and ranking probes rather than every split. Human validation is limited to 200 natural citation-looking pairs and 200 graded quality-conflict pairs with two annotators; larger multi-annotator adjudication would further strengthen the benchmark. Finally, TRACE-style decoupled judging requires additional calls and should be interpreted as an audit protocol rather than a deployment-ready judge.

Ethics Statement

LLM-as-a-judge systems are increasingly used to compare models, tune systems, select checkpoints, and report benchmark progress. If judge protocols reward superficial presentation artifacts, they may

overstate systems that produce source-looking outputs without better evidence use. If debiasing protocols collapse moderate quality differences into Tie, they may understate genuine improvements in traceability or support-chain completeness. TraceJudgeBench is intended to reduce these risks by making judge behavior more auditable. The human annotation used in this work involves judgment over model outputs rather than personal or sensitive user data. Details of annotator recruitment, workload, and compensation are provided in Appendix B.

References

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, 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. 2026. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite. In *Proceedings of the 14th International Conference on Learning Representations (ICLR 2026, Oral)*. ArXiv:2510.21652.

Penny Chong, Harshavardhan Abichandani, Jiyuan Shen, Atin Ghosh, Min Pyae Moe, Yifan Mai, and Daniel Dahlmeier. 2026. Talk, evaluate, diagnose: User-aware agent evaluation with automated error analysis. In *Proceedings of the 14th International Conference on Learning Representations (ICLR 2026)*. ArXiv:2603.15483.

Domenic V. Cicchetti and Alvan R. Feinstein. 1990. High agreement but low kappa: II. resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558.

Haoyu Dong, Pengkun Zhang, Yan Gao, Xuanyu Dong, Yilin Cheng, Mingzhe Lu, Adina Yakefu, and Shuxin Zheng. 2025. Finch: Benchmarking finance & accounting across spreadsheet-centric enterprise workflows. *Preprint*, arXiv:2512.13168.

Yann Dubois, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Proceedings of the 1st Conference on Language Modeling (COLM 2024)*. ArXiv:2404.04475.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2024)*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics. ArXiv:2309.15217.

Peng Lai, Zhihao Ou, Yong Wang, Longyue Wang, Jian Yang, Yun Chen, and Guanhua Chen. 2026. BiasScope: Towards automated detection of bias in LLM-as-a-judge evaluation. In *Proceedings of the 14th International Conference on Learning Representations (ICLR 2026)*. ArXiv:2602.09383.

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. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics. ArXiv:2403.13787.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2026. Preference leakage: A contamination problem in LLM-as-a-judge. In *Proceedings of the 14th International Conference on Learning Representations (ICLR 2026)*. ArXiv:2502.01534.

Do Xuan Long, Hai Nguyen, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F. Chen, and Min-Yen Kan. 2025. LLMs are biased towards output formats! systematically evaluating and mitigating output format bias of LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2025)*, Albuquerque, New Mexico. Association for Computational Linguistics. ArXiv:2408.08656.

Francesco Maria Molfese, Luca Moroni, Ciro Porcaro, Simone Conia, and Roberto Navigli. 2026. RetraceQA: Evaluating reasoning traces of small language models in commonsense question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*. To appear. arXiv:2510.09351.

Kangjun Noh, Seongchan Lee, Ilmun Kim, and Kyungwoo Song. 2026. Multi-LLM adaptive conformal inference for reliable LLM responses. In *Proceedings of the 14th International Conference on Learning Representations (ICLR 2026)*. ArXiv:2602.01285.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024, Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics. ArXiv:2311.09476.

Gyuhyeon Seo, Jungwoo Yang, Junseong Pyo, Nalim Kim, Jonggeun Lee, and Yohan Jo. 2026. SimuHome: A temporal- and environment-aware benchmark for smart home LLM agents. In *Proceedings of the 14th International Conference on Learning Representations (ICLR 2026, Oral)*. ArXiv:2509.24282.

Fan Shu, Yite Wang, Ruofan Wu, Boyi Liu, Zhewei Yao, Yuxiong He, and Feng Yan. 2026. DARE-bench: Evaluating modeling and instruction fidelity of LLMs in data science. In *Proceedings of the 14th International Conference on Learning Representations (ICLR 2026)*. ArXiv:2602.24288.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR 2025)*, pages 22–32, Padua, Italy. ACM. ArXiv:2412.18004.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024, Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics. ArXiv:2305.17926.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. 2026. How reliable is language model micro-benchmarking? In *Proceedings of the 14th International Conference on Learning Representations (ICLR 2026, Oral)*. ArXiv:2510.08730.

Skylar Zhai, Jingcheng Liang, and Dongyeop Kang. 2026. Abstain-R1: Calibrated abstention and post-refusal clarification via verifiable RL. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*. To appear. arXiv:2604.17073.

Yue Zhang, Seiji Maekawa, and Nikita Bhutani. 2026a. Same content, different representations: A controlled study for table QA. In *Proceedings of the 14th International Conference on Learning Representations (ICLR 2026)*. ArXiv:2509.22983.

Ziqian Zhang, Xinyu Hu, Yuxuan Huang, Kun Zhang, Rui Chen, Yang Liu, Qingsong Wen, Ke Xu, Xinyi Zhang, Neil Zhenqiang Gong, and Lichao Sun. 2026b. RankLLM: Weighted ranking of LLMs by quantifying question difficulty. In *Proceedings of the 14th International Conference on Learning Representations (ICLR 2026)*. ArXiv:2602.12424.

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. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*. ArXiv:2306.05685.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. 2026. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI 2026, Oral)*. ArXiv:2505.15055.

Appendix

This appendix provides construction details, prompt templates, extended results, and reproducibility notes for TraceJudgeBench. The main paper reports only the compact diagnostic matrix and the central bias-resolution frontier; the appendix expands the audit so that each result can be traced back to a split, protocol, and model setting.

Cross-reference to the main paper. The table below maps each main-paper float to the appendix material that supports it.

A Benchmark Construction Details

This section supports the *Mechanical equivalence*, *Natural citation-looking pairs*, *Citation ablations*, *Clear and reverse conflicts*, and *Soft and moderate graded conflicts* rows of main-paper Table 1. The audit pipeline is shown in main-paper Figure 1; this appendix provides the construction details behind that diagram.

A.1 Mechanical Equivalence Pairs

Mechanical equivalence pairs isolate presentation sensitivity under maximal content control. For each source item we construct two candidates:

- **Plain candidate.** Contains the task-relevant answer, evidence, reasoning, and final response without citation-like wrappers.
- **Cited candidate.** Preserves the same substantive answer content but adds citation-like evidence-display artifacts: source titles, citation markers, sentence identifiers, retrieved snippets, or verification-looking blocks.

The cited candidate introduces no new task-relevant facts. A systematic cited-candidate win on this split therefore reflects presentation sensitivity rather than improved answer quality. We construct 200 HotpotQA mechanical pairs and 200 FinQA mechanical pairs to test whether the effect appears in both open-domain multi-hop QA and financial/numerical reasoning settings.

A.2 Natural Citation-Looking Pairs

Mechanical pairs provide causal control but can look artificial. We therefore add 1,000 natural HotpotQA citation-looking pairs. These pairs preserve final-answer agreement but allow naturally occurring differences in source labels, evidence display, or verification-like wording. A 200-pair subset receives human validation.

The natural split is used in two ways. The full 1,000-pair set supports the plain-judge presentation-sensitivity audit. The 200-pair validated subset supports stricter protocol comparisons and human-equivalence checks. Thus, the natural split is not a full strict-protocol matrix over all 1,000 pairs.

A.3 Citation Ablation Variants

To identify which artifact types trigger presentation sensitivity, we derive six ablation variants from the HotpotQA mechanical pairs.

If length alone explained the effect, the length-control variant would dominate. If raw evidence explained the effect, the evidence-only variant would dominate. If source-like authority or traceability display explains the effect, full-citation and source-only variants should be strongest. Appendix E.1 reports the resulting rates.

A.4 Correctness Conflicts

Clear-conflict probes test whether citation-like display overwhelms obvious correctness. Each clear conflict contains one correct candidate and one incorrect candidate. In the main direction, the plain candidate is correct and the cited candidate is wrong; reverse conflicts swap the roles.

The split contains 150 pairs sampled from three domains—open-domain multi-hop QA, financial reasoning, and short code-style tasks—run across three prompt paraphrases and both candidate orders. Decoupled correctness judging uses a single prompt version and both orders; it is therefore a targeted probe rather than a full prompt-paraphrase matrix.

A.5 Graded Soft and Moderate Quality Conflicts

The graded splits test quality resolution when both candidates share the same correct final answer.

In the **moderate tier**, the plain candidate provides a complete evidence or reasoning chain. The cited candidate includes citation-like display but omits or weakens one necessary supporting hop, entity link, or evidence-chain step. The target relation is better-plain.

In the **soft tier**, both candidates match in final answer and main evidence. The plain candidate is somewhat more specific in reasoning, while the cited candidate is more generic. The target relation is intentionally near the human decision boundary.

Main paper	Appendix material	Topic
Figure 1 (audit pipeline)	Benchmark construction details (Appendix A)	Per-split and per-protocol design
Table 1 (diagnostic components)	Benchmark construction details (Appendix A)	Per-split design
Table 2 (human validation)	Table A4; kappa note (B.3)	Inter-annotator reliability
§5.5 compact frontier + TRACE	Tables A11 and A12	Per-judge full ladder and recovery
Figure 2 (equivalence suppression)	Tables A9 and A13	Per-judge × per-split breakdown
Figure 3 (bias–resolution frontier)	Tables A11, A12, A18	Moderate and soft prompt-ladder behavior
§5.1 rationale analysis	Table A19	Source/citation keyword rates
§5.6 controlled ranking	Tables A16 and A17	Workflow probe details

Table A1: Cross-reference from each main-paper float to the supporting appendix material.

Variant	Description	Diagnostic purpose
Full citation	Source labels, evidence display, and citation-like formatting	Tests the complete artifact bundle
Empty shell	Citation-like structure without informative evidence content	Tests formatting shell effects
Source-only	Source titles or source labels without full evidence blocks	Tests source-like authority cues
Evidence-only	Evidence text without source/citation presentation	Tests raw evidence content
Sentence-id-only	Sentence or passage identifiers without source display	Tests identifier artifacts
Length-control	Added non-substantive length on the perturbed/cited side without citation structure	Tests length as an alternative explanation

Table A2: Citation ablation variants.

Construction proceeds in three steps. (1) Lock the shared final answer and supporting facts from locally checked HotpotQA traces. (2) Generate either a moderate support-chain omission or a soft specificity reduction using fixed templates. (3) Validate the resulting pairs with blind human annotation (Appendix B).

A.6 Representative Examples

These examples illustrate the construction patterns of Appendices A.1, A.4, and A.5.

Mechanical equivalence example. *Task:* Were Scott Derrickson and Ed Wood of the same nationality? The plain candidate gives the final answer "yes" using evidence that Scott Derrickson is American and Ed Wood was an American filmmaker. The cited candidate preserves the same answer and evidence, then appends a citation block reusing the same source sentences. The target relation is content-equivalent; the gold relation is Tie.

Clear correctness conflict example. For the same task, the plain candidate keeps the correct answer "yes". The cited candidate displays source-like citation markers over the same evidence but changes the final answer to "no". The target relation is correctness conflict; the gold relation is better-plain.

Moderate quality-gap example. *Task:* Where did the descendants of the Black Seminoles settle? The plain candidate answers "Coahuila, Mexico" and includes both the Mascogos evidence and the Black Seminoles bridge. The cited candidate gives the same final answer and citation-like display but includes only the Mascogos evidence, omitting part of the support chain. The target relation is better-plain; a Tie on this item type is treated as resolution loss.

Soft boundary example. In the soft tier, both candidates share the same final answer and main evidence, but the plain candidate gives a somewhat more explicit reasoning sentence while the cited candidate gives a more generic support statement. Human validation shows that these items sit near the decision boundary; Tie is therefore interpreted as possible calibration rather than automatic failure.

B Human Validation

B.1 Annotation Tasks

Annotators see the task and two candidate outputs without knowing which side carries citation-like display or which label is expected. For natural equivalence validation, annotators use one primary label.

Annotators additionally check whether the citation-looking candidate introduces new factual content, whether the citation/source display could affect presentation preference without changing correctness, and whether any mismatch exists in final answer, numeric answer, named entity, date, or yes/no polarity.

For graded soft and moderate validation, annotators answer: (i) whether the final answers are equivalent; (ii) whether one candidate has more complete support; (iii) whether one candidate has more specific reasoning; (iv) whether citation/source display adds substantive information; and (v) the final preference label—Candidate A, Candidate B, or Tie.

The final preference label is interpreted relative to the target relation of each pair. A Tie on equivalence pairs supports the desired equivalence relation. A Tie on soft-boundary pairs may be calibrated. A Tie on moderate pairs is treated as resolution loss when human consensus establishes a better-plain direction. Natural-pair validation is reported separately from mechanical minimal-pair validation; the two are not merged in main-paper rates.

Scope of human validation. The current release uses two independent annotators for the human-validated subsets. Two annotators were recruited from graduate students; both were fluent English speakers familiar with NLP evaluation tasks. Each annotator labeled 400 pairs, consisting of 200 natural-equivalence pairs and 200 graded quality-gap pairs. The annotation workload was estimated at approximately 8 hours per annotator, and annotators were compensated at an hourly rate consistent with local minimum-wage requirements. We chose blind pairwise annotation because the target relation is local to a pair rather than an absolute quality score. The moderate-tier claim relies on high-consensus items: 91/100 pairs are labeled better-plain by both annotators, with 100.0% raw agreement over the tier. We do not claim that two annotators are sufficient for a universal preference benchmark; rather, this validation establishes that the moderate tier contains real support-chain gaps for the diagnostic audit. The annotation workbook released with the artifact includes the guidelines, item identifiers, blind candidate order, per-item labels, and adjudication flags for items that do not support the target relation.

B.2 Validation Summary

We repeat the main-paper validation table here for completeness, then add the kappa note that does not fit in the main 8 pages.

B.3 Why Cohen's Kappa Is Not Reported

The main-paper table omits Cohen's kappa because these validation sets are intentionally prevalence-skewed. In the natural-equivalence set, most examples are expected to be equivalent; in the graded-moderate set, most items are expected to be better-plain. Under such marginal skew, kappa is known to collapse even when raw agreement is high—the "high agreement but low kappa" paradox documented by Cicchetti and Feinstein (1990). We therefore report raw agreement, PABAK, and Gwet's AC1 as the primary reliability signals. Per-item kappa values can be released as supplementary data if needed; they do not change the qualitative conclusion that annotator consensus is strong on the moderate tier.

C Judge Protocols

C.1 Single-Call Prompt Ladder

TraceJudgeBench uses a prompt-strength ladder to study the operating point of a judge protocol. The ladder is not intended as an exhaustive prompt search; it provides interpretable settings from ordinary preference judging to strict anti-presentation judging.

Representative instruction excerpts for the moderate-tier ladder are shown in Table A6.

All single-call prompts request JSON with `winner` in `{A, B, Tie}` and a short rationale.

Interpreting L4. L4 is a deliberately strict stress-test endpoint. It is useful because it reveals how far a judge can move toward abstention when presentation cues are aggressively discounted, but it is not treated as the only or recommended debiasing protocol. The main paper therefore interprets the full ladder, with special attention to L2 and L3: these levels resemble common evaluation instructions that ask judges to ignore formatting, polish, citation display, length, and non-substantive

Label	Meaning
`equivalent`	Both candidates give the same final answer and are equally correct for the task. Citation/source formatting introduces no correctness-relevant difference.
`base_more_correct`	The base/plain candidate is more correct, more faithful to evidence, or less misleading.
`citation_more_correct`	The citation-looking candidate is more correct, more faithful to evidence, or less misleading.
`unclear`	The item cannot be confidently judged from the provided traces.

Table A3: Human annotation labels for natural citation-looking validation.

Validation set	n	Raw agreement	PABAK	Gwet's AC1	Main human-label signal
Natural equivalence	200	92.0%	0.840	0.918	183/200 both equivalent
Graded soft	100	96.0%	0.920	0.958	1/100 both better plain
Graded moderate	100	100.0%	1.000	1.000	91/100 both better plain
Graded overall	200	98.0%	0.960	0.976	moderate is resolution-loss evidence

Table A4: Human validation summary (repeated from main-paper Table 2 for self-contained discussion of reliability metrics).

Level	Name	Intended behavior
L0	Plain preference	Choose the better overall answer, allowing Tie
L1	Anti-citation warning	Warn that citation-like display may be superficial
L2	Quality-ignore-citation	Evaluate answer quality while ignoring citation formatting unless it adds substantive information
L3	Correctness-first	Prioritize correctness and substantive support over presentation, length, or polish
L4	Strict tie-required	Return Tie when candidates are equivalent in final-answer correctness, even if one appears more cited or polished

Table A5: Single-call prompt ladder.

Level	Prompt excerpt
L0	Choose which candidate response is better overall for the task. If both candidates are equally good, choose Tie.
L1	Choose which candidate response is better overall, but do not prefer a response merely because it looks cited, sourced, or workflow-like.
L2	Choose the response that is better overall, but ignore citation formatting, source labels, evidence-block style, response length, and polish by themselves. Consider those features only if they change task correctness or genuinely necessary information.
L3	First judge task correctness. If both answers are equally correct, use genuinely task-relevant support completeness or reasoning specificity as a tiebreaker, but do not use citation-like formatting, source titles, evidence display style, length, or polish as advantages by themselves.
L4	Prefer a candidate only when it is more correct for the task. If both candidates give the same correct final answer and differ only in citation formatting, source labels, evidence display, length, style, verification wording, support-chain detail, or reasoning specificity, choose Tie.

Table A6: Moderate-tier prompt excerpts. Full prompt templates are released with the artifact.

presentation while still selecting the better answer when substantive support differs. Resolution loss at L2/L3 is therefore more informative than the L4 endpoint alone.

C.2 Generic Pairwise Judge Output Schema

All protocols request a structured label:

```
{
  "winner": "A | B | Tie",
  "reason": "brief explanation"
}
```

When structured output is not natively enforced by the endpoint, responses are parsed with a conservative label extractor. Ambiguous or unparsable outputs are logged for manual inspection rather than silently coerced.

C.3 TRACE-Style Decoupled Audit

TRACE-style decoupling uses two calls per ordered comparison in our implementation:

1. **Joint normalization call.** Normalize both candidates in one call, removing or masking citation/source presentation while preserving final answer, support-chain completeness, reasoning specificity, unsupported claims, omitted support, and partial-support signals.

2. **Masked judging call.** Compare the normalized records without seeing the original citation-like display.

The quality-preserving version used for the main frontier is `decoupled_quality_v2`. The correctness-only version, `decoupled_v1`, is useful as a diagnostic contrast because it can suppress presentation cues while failing to preserve quality-relevant support-chain structure.

The Stage 1 normalization prompt instructs the judge to remove citation formatting, source labels, source titles, bracketed references, appended citation blocks, evidence-display style, surface polish, and wording differences that do not alter task-relevant content. It explicitly preserves final answer, task-relevant factual claims, reasoning needed to connect claims to the final answer, support-chain completeness, explanation specificity, missing support, and whether the response merely states an answer without support.

To reduce the risk that TRACE merely moves the cue into another representation, the normalized record uses fixed fields rather than free-form rewritten answers: final answer, support facts, reasoning links, missing or weakened support, unsupported claims, and residual uncertainty. Source titles, bracketed citations, document labels, passage identifiers, and source-count features are masked or omitted. The masked judging call sees only these normalized fields and the task, not the original candidate text or the original side labels. Released logs include both the original pair and normalized record so that reviewers can audit whether presentation cues were removed and whether quality-relevant support was preserved.

The Stage 2 masked comparison prompt receives only the normalized records. It judges task correctness, support-chain completeness, and reasoning specificity. It instructs the judge not to infer quality from citations, source labels, formatting, length, polish, or evidence-display style, and to choose Tie when the normalized content is task-quality equivalent.

TRACE is an audit protocol, not a deployment-ready solution. Its role is to test whether resolution loss is inevitable under debiasing; the answer is no, provided the protocol preserves support-chain quality during normalization.

D Model and Serving Details

D.1 Commercial Judges

The full diagnostic matrix is evaluated on three commercial judges.

All runs use temperature 0. Structured JSON output is used where supported. Provider-side model snapshots may not be fully exposed by all endpoints; released logs include call dates, endpoint aliases, endpoint-reported snapshots when available, response metadata, prompt templates, retry rules, and raw A/B/Tie decisions.

D.2 Open-Weight Judges

The open-weight experiments focus on robustness replication of the central moderate-gap frontier rather than full TraceJudgeBench coverage.

Both open-weight judges are served via a local OpenAI-compatible endpoint. Released logs include the exact `MODEL_NAME` returned by the endpoint, base URL type, quantization/runtime, call dates, and inference parameters. The main paper does not claim full open-weight coverage of equivalence, ablation, correctness-conflict, position-control, or paraphrase-audit splits.

E Extended Results

E.1 Equivalence and Natural Protocol Summary

This expanded view supports main-paper Figure 2 (equivalence suppression) and the first claim of §5.1.

Here n is the number of judge rows after applying the protocol-specific pair subset, prompt paraphrases, and candidate-order settings, not simply the number of source questions. Plain and strict single-call rows include prompt paraphrases and candidate-order presentation; decoupled rows use

Judge	Identifier used in logs	Additional endpoint metadata	Coverage
GPT-5.5	`gpt-5.5`	endpoint-reported snapshot `gpt-5.5-2026-04-23`	full diagnostic matrix
Claude Sonnet 4.6	`claude-sonnet-4-6`	provider snapshot not fully exposed by the endpoint	full diagnostic matrix
DeepSeek	`deepseek-v4-flash`	endpoint-reported alias used for all DeepSeek runs	full diagnostic matrix

Table A7: Commercial judge identifiers and coverage.

Judge	Checkpoint / serving identifier	Coverage	Role
Qwen2.5-14B-Instruct-AWQ	`Qwen2.5-14B-Instruct-AWQ`	full moderate prompt frontier, TRACE recovery, controlled ranking probe	open-weight robustness replication
Gemma-3-12B-IT	`Gemma-3-12B-IT`	full moderate prompt frontier, TRACE recovery, controlled ranking probe	open-weight robustness replication

Table A8: Open-weight judge identifiers and coverage.

Model	Protocol	n	Tie	Citation win	Base win	Signed effect
GPT-5.5	Mechanical Plain	1200	14.2	85.8	0.0	85.8
GPT-5.5	Mechanical Strict	1200	100.0	0.0	0.0	0.0
GPT-5.5	Decoupled Mechanical	400	100.0	0.0	0.0	0.0
GPT-5.5	Natural Plain	6000	6.4	93.5	0.1	93.4
GPT-5.5	Natural Strict	1200	99.9	0.1	0.0	0.1
Claude Sonnet 4.6	Mechanical Plain	1200	48.8	50.9	0.2	50.7
Claude Sonnet 4.6	Mechanical Strict	1200	100.0	0.0	0.0	0.0
Claude Sonnet 4.6	Decoupled Mechanical	400	100.0	0.0	0.0	0.0
Claude Sonnet 4.6	Natural Plain	6000	77.1	22.8	0.1	22.7
Claude Sonnet 4.6	Natural Strict	1200	100.0	0.0	0.0	0.0
DeepSeek	Mechanical Plain	1200	52.5	47.2	0.3	46.8
DeepSeek	Mechanical Strict	1200	100.0	0.0	0.0	0.0
DeepSeek	Decoupled Mechanical	400	100.0	0.0	0.0	0.0
DeepSeek	Natural Plain	6000	74.4	24.6	1.0	23.7
DeepSeek	Natural Strict	1200	99.8	0.0	0.2	-0.2

Table A9: Equivalence and natural protocol summary.

a smaller audited subset and one protocol setting. The released manifest records the exact pair subset, prompt family, and order expansion for each row count.

This table supports the first-stage claim in the main paper: stricter and decoupled protocols suppress citation wins on equivalence-like pairs, mostly by moving decisions to Tie.

E.2 Citation Ablation Results

The citation-ablation split identifies which artifact types activate citation preference. Rates are percentages over 400 comparisons per variant.

The strongest positive effects are full-citation and source-only variants. Length-control does not explain the effect.

E.3 Moderate Prompt Frontier (Full Ladder)

The main paper reports L0 and L4 endpoints in compact form. The full prompt ladder shows how better-plain resolution changes across stricter operating points; the pattern is model-specific and not assumed to be monotonic.

The table reports better-plain and Tie rates; the omitted residual is the worse-cited win rate. Worse-cited rates are near zero across the moderate ladder and never drive the frontier. For example, Qwen L0 is 99.5% better-plain, 0.5% worse-cited, and 0.0% Tie; the largest omitted residual in this compact table is 2.0%, from Qwen at L3. The frontier is therefore driven primarily by better-plain decisions being converted into Tie, not by renewed prefer-

Model	Variant	Tie	Citation win	Base win	Signed effect
GPT-5.5	full citation	25.2	74.5	0.2	74.2
GPT-5.5	source-only	41.2	58.8	0.0	58.8
GPT-5.5	empty shell	52.0	0.0	48.0	-48.0
GPT-5.5	evidence-only	16.2	0.0	83.8	-83.8
GPT-5.5	sentence-id-only	95.5	0.0	4.5	-4.5
GPT-5.5	length-control	0.5	0.0	99.5	-99.5
Claude Sonnet 4.6	full citation	48.5	51.5	0.0	51.5
Claude Sonnet 4.6	source-only	71.8	28.2	0.0	28.2
Claude Sonnet 4.6	empty shell	98.0	0.0	2.0	-2.0
Claude Sonnet 4.6	evidence-only	98.8	0.0	1.2	-1.2
Claude Sonnet 4.6	sentence-id-only	100.0	0.0	0.0	0.0
Claude Sonnet 4.6	length-control	0.2	0.0	99.8	-99.8
DeepSeek	full citation	53.2	46.8	0.0	46.8
DeepSeek	source-only	90.5	9.5	0.0	9.5
DeepSeek	empty shell	87.2	12.2	0.5	11.8
DeepSeek	evidence-only	83.5	0.0	16.5	-16.5
DeepSeek	sentence-id-only	98.8	1.0	0.2	0.8
DeepSeek	length-control	7.8	0.0	92.2	-92.2

Table A10: Citation ablation results.

Judge	L0		L1		L2		L3		L4	
	better	Tie	better	Tie	better	Tie	better	Tie	better	Tie
GPT-5.5	100.0	0.0	100.0	0.0	99.0	1.0	99.0	1.0	2.0	98.0
Claude Sonnet 4.6	100.0	0.0	97.5	2.5	91.0	9.0	71.5	28.5	11.0	89.0
DeepSeek	99.5	0.5	96.5	3.0	82.5	17.0	79.0	21.0	15.0	85.0
Qwen2.5-14B-Instruct-AWQ	99.5	0.0	97.5	2.0	89.5	10.0	90.0	8.0	25.5	74.5
Gemma-3-12B-IT	100.0	0.0	99.5	0.5	97.0	3.0	43.0	57.0	52.0	48.0

Table A11: Moderate prompt frontier (full ladder).

ence for the weaker cited candidate.

E.4 TRACE Recovery on the Moderate Tier

Table A12 reports the TRACE recovery rates on the moderate tier.

TRACE recovery supports the interpretation that resolution loss is protocol-dependent. Removing citation-like display is not sufficient; the normalization step must preserve evidence-chain quality.

E.5 Equivalence Suppression by Split

Main-paper Figure 2 (equivalence suppression) condenses these numbers into per-judge bars. See Table A13 for the full breakdown.

These results motivate the first half of the paper: strict anti-citation prompting can indeed suppress presentation preference in equivalence settings.

E.6 Correctness Retention

Clear and reverse correctness conflicts show that resolution loss is not simply correctness collapse. The decoupled correctness-retention probe remains at or above 93.0% correct-base wins across commercial judges, supporting the claim that the main failure mode lies in moderate support-chain gaps rather than obvious correct-versus-wrong cases. L4 strict-prompt correctness-family results are included in the anonymized release artifact; Table A14 is a targeted decoupled probe and should not be read as the L4 strict-prompt endpoint.

The decoupled correctness-retention probe rates appear in Table A14.

For full plain/strict correctness-family reporting, the anonymized release artifact includes per-judge, per-prompt-family, per-paraphrase, and per-order rates.

E.7 Position Controls

Position-only controls use *identical* plain/no-citation candidates in both orders. Any non-Tie decision in this setting would indicate generic position bias. Table A15 shows that the controls return all Tie.

Why this does not contradict the citation-order gaps reported in §5.1. The 100% Tie above holds when *both* candidates are identical plain text. The citation-order gaps in the main paper arise on pairs where one candidate carries citation-like display and the other does not, presented in different orders. Position-only controls therefore confirm

Judge	TRACE better-plain	TRACE worse-cited	TRACE Tie
GPT-5.5	99.0	0.0	1.0
Claude Sonnet 4.6	100.0	0.0	0.0
DeepSeek	97.0	0.0	3.0
Qwen2.5-14B-Instruct-AWQ	98.5	0.0	1.5
Gemma-3-12B-IT	96.5	0.0	3.5

Table A12: TRACE recovery on the moderate tier.

Judge	Split	Plain cited win	Strict Tie
GPT-5.5	HotpotQA mechanical	85.8	100.0
GPT-5.5	FinQA mechanical	83.5	100.0
GPT-5.5	Natural citation-looking	93.5	99.9
Claude Sonnet 4.6	HotpotQA mechanical	50.9	100.0
Claude Sonnet 4.6	FinQA mechanical	42.3	100.0
Claude Sonnet 4.6	Natural citation-looking	22.8	100.0
DeepSeek	HotpotQA mechanical	47.2	100.0
DeepSeek	FinQA mechanical	44.2	100.0
DeepSeek	Natural citation-looking	24.6	99.8

Table A13: Equivalence suppression under strict prompting.

Model	n	Correct base win	Wrong cited win	Tie
GPT-5.5	300	97.7	1.0	1.3
Claude Sonnet 4.6	300	98.0	1.0	1.0
DeepSeek	300	93.0	2.0	5.0

Table A14: Decoupled correctness-retention probe.

Model	n	First-position win	Second-position win	Tie
Claude Sonnet 4.6	1200	0.0	0.0	100.0
DeepSeek	400	0.0	0.0	100.0

Table A15: Position-only controls.

that the order coupling is *citation-conditioned*, not generic—i.e., position bias does not emerge until citation-like display is introduced.

GPT-5.5 is not included in this targeted position-only control table because the added controls were run for judges that showed the largest citation-conditioned order gaps in the commercial audit. In the mechanical plain split, GPT-5.5's cited-win rate differs by 2.7 percentage points across candidate orders, whereas DeepSeek and Claude required additional controls to rule out generic position-only behavior. The absence of a GPT-5.5 row should therefore be read as scoped control coverage, not as a claim that GPT-5.5 has no order sensitivity in general.

E.8 Workflow-Ranking Probe

The ranking probe uses four workflow variants—`answer_only`, `plain_full`, `cited_full`, and `cited_partial`. It contains 200 HotpotQA queries, all six unordered system pairs per query, and both candidate orders, yielding 2,400 judge rows per protocol/model setting (each system appears in 1,200 ordered comparisons). Scores count one point for a win and one-half point for a Tie, normalized to percentages.

E.9 Direct Workflow-Pair Comparisons

Direct full-workflow comparisons show the presentation effect more explicitly:

The direct-comparison table uses 400 ordered rows per judge/protocol, corresponding to 200 queries in both candidate orders. It should therefore not be compared directly to order-collapsed 200-row summaries.

F Soft Tier: Calibrated Tie Counterpart

Main-paper §5.3 argues that Tie is calibrated on soft-boundary pairs but resolution-destroying on moderate-gap pairs. The soft-tier prompt ladder,

Judge	Protocol	answer_only	plain_full	cited_full	cited_partial
GPT-5.5	plain preference	19.38	66.62	98.17	15.83
GPT-5.5	TRACE	2.62	83.00	83.17	31.21
Claude Sonnet 4.6	plain preference	2.54	76.83	89.29	31.33
Claude Sonnet 4.6	TRACE	2.38	83.12	83.12	31.37
DeepSeek	plain preference	1.71	80.12	82.50	35.67
DeepSeek	TRACE	2.04	80.33	82.75	34.88

Table A16: Workflow-ranking probe system scores.

Judge	Protocol	`plain_full` wins	`cited_full` wins	Tie
GPT-5.5	plain preference	0	371	29
GPT-5.5	correctness-first	0	2	398
GPT-5.5	TRACE	0	0	400
Claude Sonnet 4.6	plain preference	0	147	253
Claude Sonnet 4.6	correctness-first	0	0	400
Claude Sonnet 4.6	TRACE	0	1	399
DeepSeek	plain preference	1	4	395
DeepSeek	correctness-first	0	0	400
DeepSeek	TRACE	5	23	372

Table A17: Direct `plain_full` versus `cited_full` comparisons.

reported here, supports the calibrated half of that claim.

Figure A1: Soft-tier prompt-ladder behavior

The contrast between Table A18 (soft) and Table A11 (moderate) is the core empirical evidence behind the three-meanings-of-Tie distinction in §5.3: when humans agree there is a direction, judges that abstain are losing resolution; when humans themselves do not commit to a direction, judges that abstain are calibrated.

G Rationale Keyword Analysis

This section supports the main paper's claim (§5.1) that judges respond to source-like authority or traceability display rather than to additional tokens alone. Rationale analysis is used only as supporting evidence; it does not reveal hidden mechanisms and is not treated as causal proof.

We analyze public judge rationales for source/citation-related keywords and compare citation-win rationales with Tie rationales.

Citation-win rationales consistently describe cited outputs as more supported, traceable, or verifiable. Tie rationales can still mention citation-like display, especially for GPT-5.5 and natural-pair Claude results. We interpret this as a distinction between *cue perception* and *cue use*: a judge may notice citation-like display while declining to convert it into a preference.

Table A19 reports the rationale subsets used for the keyword audit rather than a complete rationale matrix over every split and model. Missing cells indicate that the corresponding rationale subset was not part of the released keyword-rate audit or was too small for stable reporting.

H Reproducibility Checklist

Released artifacts include:

- `splits/`: benchmark split files with stable item identifiers, source dataset tags, target relation labels, candidate-order fields, and artifact-side metadata;
- `prompts/`: full L0–L4 prompt templates, prompt paraphrases, TRACE normalization prompts, and masked judging prompts;
- `raw_outputs/`: raw A/B/Tie outputs, rationales, parse status, retry status, and unparsed responses;
- `metadata/`: model identifiers, endpoint aliases, endpoint-returned snapshots when

Judge	L0		L1		L2		L3		L4	
	worse	Tie	worse	Tie	worse	Tie	worse	Tie	worse	Tie
GPT-5.5	50.5	32.2	0.0	78.5	0.0	99.5	0.0	99.0	0.0	100.0
Claude Sonnet 4.6	10.0	51.5	0.0	80.5	0.0	93.0	0.0	98.5	0.0	100.0
DeepSeek	24.5	70.0	0.0	97.5	0.0	98.0	0.0	99.0	0.0	99.5

Table A18: Soft-tier prompt frontier. "worse" is the worse-cited win rate, "Tie" is the Tie rate; the omitted residual is the better-plain rate, which is small throughout because the soft tier is intentionally constructed near the human decision boundary.

Judge	Split	Decision type	Keyword rate
GPT-5.5	HotpotQA mechanical	citation win	100.0
GPT-5.5	FinQA mechanical	citation win	100.0
GPT-5.5	HotpotQA mechanical	Tie	96.5
Claude Sonnet 4.6	HotpotQA mechanical	citation win	100.0
Claude Sonnet 4.6	FinQA mechanical	citation win	100.0
Claude Sonnet 4.6	HotpotQA mechanical	Tie	40.1
Claude Sonnet 4.6	FinQA mechanical	Tie	30.9
Claude Sonnet 4.6	Natural citation-looking	Tie	97.1
DeepSeek	HotpotQA mechanical	citation win	100.0
DeepSeek	HotpotQA mechanical	Tie	15.6

Table A19: Source/citation keyword rates in judge rationales (%).

available, call dates, inference parameters, and provider response metadata;

- `human_validation/`: anonymized annotation workbooks, guidelines, blind candidate order, per-item labels, and validation summaries;
- `scripts/`: parsers and analysis scripts for win rates, Tie rates, signed effects, order gaps, PABAK, Gwet's AC1, Wilson intervals, and bootstrap confidence intervals;
- `requirements.txt`: pinned non-standard Python runtime dependency for annotation-workbook processing; the remaining scripts use the Python standard library;
- `figures/`: scripts and source tables used to generate the main-paper figures and appendix tables;
- `logs/`: amendment logs for any model-name, endpoint-name, parsing, or data-cleaning changes.

The artifact is anonymized. It does not include private API keys, account identifiers, or non-anonymized annotator information.

Caveat on endpoint version pinning. Some commercial endpoints expose limited version information. For DeepSeek the endpoint-returned alias (`deepseek-v4-flash`) does not always reveal the deeper provider-side snapshot; for Claude Sonnet 4.6, the snapshot is not exposed by the gateway. Released logs therefore preserve all available metadata returned by the endpoint and report the call-date window, but exact provider-side snapshots may not be recoverable. We treat this as a reproducibility limitation rather than a defect of the audit design.

Prompt paraphrase coverage. The prompt ladder uses one paraphrase per level for soft and most moderate analyses, with three paraphrases reserved for clear and reverse correctness conflicts. A single L3 paraphrase may not capture the full variance of correctness-first rubric behavior; the anonymized release artifact lists all paraphrases used so the analysis can be re-run with extended coverage.