The Judge Effect in Two-Round Legal Debate on LegalBench

David Johnson, Onur Bilgin, and John Licato
Advancing Machine and Human Reasoning (AMHR) Lab
Bellini College of Artificial Intelligence, Cybersecurity and Computing
University of South Florida
{daj3, onurbilgin, licato}@usf.edu

Abstract

Large language models can produce fluent, plausible legal analysis while still misapplying rules and outputting incorrect labels. Such errors are especially problematic in legal reasoning tasks, where they can be persuasive to non-specialist observers. In this paper, we use the term "judge effect" for the paired difference between an advocates-only protocol (NoJ) and a judge-augmented protocol (Judge LLM), and we study this effect in two-round legal debate on LegalBench contracts. Compared to an advocates-only baseline, the Judge LLM improves accuracy on contracts in all three same-model setups. We also report token cost explicitly.

Introduction

Multi-agent debates (MAD) use multiple interacting agents backed by large language models (LLMs) to improve problem solving across tasks such as question answering and fact checking (Irving, Christiano, and Amodei 2018; Du et al. 2023; Chen, Saha, and Bansal 2024). Prior work shows that LLMs can exhibit systematic failure modes and biases as debating and judging agents (Fu et al. 2023; Zheng et al. 2023; Wang et al. 2023; Saito et al. 2023). These issues are especially pronounced in legal reasoning, where legal correctness depends on the correct application of rule elements to facts. We investigate whether adding a neutral Judge LLM to a multi-agent debate can improve accuracy on LegalBench contracts (Guha et al. 2023). We refer to the advocates-only condition as **NoJ** (no judge) and the judge-augmented condition as **Judge LLM**. We address two questions: **RQ1:** when model identity is held fixed, does adding a Judge LLM improve accuracy relative to NoJ? **RQ2:** does the UniEval consistency score track correctness on the final decision output? Prior work on debate and LLM-based judging shows that aggregation rules and evaluator biases can materially affect outcomes (Irving, Christiano, and Amodei 2018; Du et al. 2023; Chen, Saha, and Bansal 2024; Wynn, Satija, and Hadfield 2025; Zhang et al. 2025; Wang et al. 2023; Saito et al. 2023; Fu et al. 2023; Zheng et al. 2023).

Copyright © 2026 by the authors. Open access article published under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

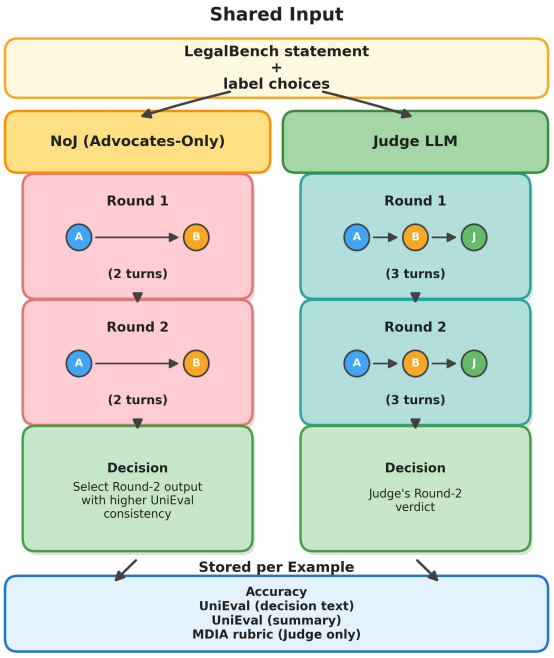

Figure 1: Two-round debate protocols for NoJ and Judge LLM.

Methodology

Debate Protocols

Figure 1 illustrates the two debate protocols. In NoJ, we score both Round 2 advocate outputs with the UniEval consistency score and use the emitted label of the higher-scoring output as the prediction, breaking exact ties uniformly at random. In Judge LLM, the final prediction is the Judge LLM's Round 2 verdict. In the main comparison, the Judge LLM uses the same base model as the advocates.

Dataset, Models, and Metrics

LegalBench contracts. We evaluate the full "LegalBench ucc/common law" test split, which contains 94 short transaction descriptions labeled as "ucc/common law" (Guha et al. 2023; Institute 2026). Prompts are self-contained and fit within the model context window; agents receive only the

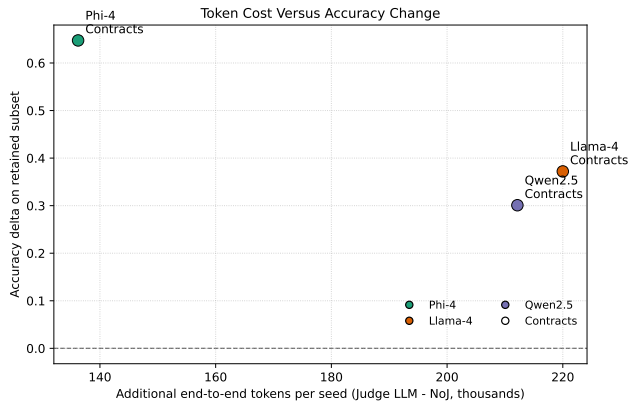

Figure 2: Token cost versus accuracy change for the same-model contracts comparison. Each point is one model setup. The x-axis is the mean additional end-to-end tokens per seed across evaluated samples (Judge LLM minus NoJ), and the y-axis is mean ΔAccuracy. Seeds are 101, 102, and 103.

short transaction description and label choices (UCC vs. common law), not an underlying full contract document. Correct classification requires applying the governing-law test for goods vs. services (including mixed transactions) under Article 2 of the Uniform Commercial Code (UCC).

We evaluate three debater setups using Phi-4, Llama-4, and Qwen2.5 (Abdin et al. 2024; Microsoft 2026; Meta 2026; Qwen et al. 2024; Qwen 2026; Qi et al. 2024; Tamirisa et al. 2024). We report results over seeds 101, 102, and 103, with n shown explicitly for each protocol as evaluated/attempted samples per seed. Here, *evaluated* means that the protocol produced a valid final label that could be scored against gold. We also report end-to-end prompt-plus-generation totals and summarize the added Judge LLM cost as the mean additional tokens per seed across evaluated samples.

Results and Discussion

RQ1. The Judge LLM improves contracts accuracy across all three same-model setups: 0.333 to 0.981 for Phi-4, 0.596 to 0.968 for Llama-4, and 0.580 to 0.880 for Qwen2.5 (Table 1). The target decision often reduces to a direct doctrinal checklist (predominant purpose / "sale of goods" under UCC Article 2 vs. common law), and prompts frequently contain direct cues about what is being exchanged (goods, services, or mixed transactions).

RQ2. In NoJ, the UniEval consistency score is the rule used to choose between the two Round 2 advocate outputs, but it should not be treated as a direct measure of legal correctness for the final decision. Because the design keeps model identity fixed between advocates and judge, the main remaining protocol differences are the judge role itself and the extra tokens required to run it.

Figure 2 shows that Phi-4 yields the largest accuracy gain at the lowest added token cost, while Llama-4 and Qwen2.5 require much larger token increases for smaller gains.

Setup	NoJ n	Judge n	NoJ Acc	Judge LLM Acc	ΔAcc
Phi-4	54/94	52/94	0.333	**0.981**	+0.647
Llama-4	94/94	93/94	0.596	**0.968**	+0.372
Qwen2.5	88/94	92/94	0.580	**0.880**	+0.301

Table 1: Contracts same-model judge-effect results over seeds 101, 102, and 103. The higher-accuracy side is bolded in each row. n reports evaluated/attempted samples per seed.

Feature	NoJ F	Judge F	Feature	NoJ F	Judge F
Prec. Align.	387.9	435.8	Coherence	226.7	244.5
Groundedness	357.8	346.8	Log. Sound.	215.2	230.1
Completeness	295.7	287.4	Clarity	184.6	189.0
Persuasiveness	235.9	259.5	Defeasibility	75.7	135.3

Table 2: Contracts Round 2 MDIA Analysis of Variance (ANOVA) summary over pooled per-output rows from seeds 101–103. Larger F indicates stronger separation between correct and incorrect advocate outputs; NoJ uses $n = 1416$, Judge LLM uses $n = 1422$, and all reported dimensions satisfy $p < 10^{-17}$.

We also report minimally defeasible interpretive argumentation (MDIA), a framework for choosing the interpretation of a rule best supported by interpretive arguments that are least vulnerable to defeat (Licato 2021). Here we use MDIA as a structured secondary signal: we score the two Round 2 advocate outputs on eight 1–10 dimensions — defeasibility, persuasiveness, completeness, groundedness, precedent alignment, coherence, clarity, and logical soundness — and compute the MDIA gap as the correct advocate's mean score minus the incorrect advocate's mean score on decisive output pairs. Pooling seeds 101, 102, and 103 with evaluators matched to each model family, the contracts gap is 1.154 [1.069, 1.242] for NoJ ($n = 708$) and 1.265 [1.174, 1.361] for Judge LLM ($n = 711$). Here, n counts decisive pairs and brackets denote 95% bootstrap intervals. Table 2 shows that precedent alignment, groundedness, and completeness are the strongest per-dimension separators between correct and incorrect contracts outputs in both protocols.

Conclusion

Across three debater setups on LegalBench contracts, we quantify how introducing a Judge LLM changes two-round debate outcomes relative to NoJ. Judge LLM improves accuracy across all three same-model contracts setups, but the compute tradeoff remains material, so token usage should be reported alongside accuracy rather than treated as hidden overhead. Systems that assist with contract interpretation or other legal analysis can influence how people understand obligations, allocate risk, and make decisions with real financial or legal consequences. Our results suggest that judge-style deliberation may improve accuracy in rule-structured settings, but it should be used as decision support with human oversight and task-specific validation rather than treated as a universal replacement for legal judgment.

References

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Chen, J.; Saha, S.; and Bansal, M. 2024. ReConcile: Roundtable conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7066–7085.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Fu, X.-Y.; Laskar, M. T. R.; Chen, C.; and Tn, S. B. 2023. Are large language models reliable judges? a study on the factuality evaluation capabilities of LLMs. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, 310–316. arXiv:2311.00681.

Guha, N.; Nyarko, J.; Ho, D. E.; Ré, C.; Chilton, A.; et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Institute, L. I. 2026. Uniform commercial code (UCC). `https://www.law.cornell.edu/ucc`. Accessed: 2026-01-07.

Irving, G.; Christiano, P.; and Amodei, D. 2018. Ai safety via debate. *arXiv preprint arXiv:1805.00899*.

Licato, J. 2021. How should ai interpret rules? a defense of minimally defeasible interpretive argumentation. *arXiv preprint arXiv:2110.13341*.

Meta. 2026. Llama-4-scout-17b-16e-instruct. `https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct`. Accessed: 2026-03-16.

Microsoft. 2026. Phi-4 model card. `https://huggingface.co/microsoft/phi-4`. Accessed: 2026-01-06.

Qi, X.; Wei, B.; Carlini, N.; Huang, Y.; Xie, T.; He, L.; Jagielski, M.; Nasr, M.; Mittal, P.; and Henderson, P. 2024. On evaluating the durability of safeguards for open-weight LLMs. *arXiv preprint arXiv:2412.07097*.

Qwen; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen. 2026. Qwen2.5-14b-instruct model card. `https://huggingface.co/Qwen/Qwen2.5-14B-Instruct`. Accessed: 2026-01-06.

Saito, K.; Wachi, A.; Wataoka, K.; and Akimoto, Y. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Tamirisa, R.; Bharathi, B.; Phan, L.; Zhou, A.; Gatti, A.; Suresh, T.; Lin, M.; Wang, J.; Wang, R.; Arel, R.; Zou, A.; Song, D.; Li, B.; Hendrycks, D.; and Mazeika, M. 2024. Tamper-resistant safeguards for open-weight LLMs. *arXiv preprint arXiv:2408.00761*.

Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Wynn, A.; Satija, H.; and Hadfield, G. 2025. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*.

Zhang, H.; Cui, Z.; Chen, J.; Wang, X.; Zhang, Q.; Wang, Z.; Wu, D.; and Hu, S. 2025. Stop overvaluing multi-agent debate – we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; et al. 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.