Multi-Agent Debate for Legal Reasoning

Eryclis Silva[1], Sean Liao[1], Jared Buabeng[1], James Han[1], Arihant Singh[1], Xiaocheng Ma[1],

[1]University of Illinois at Urbana-Champaign,

Correspondence: ersilva2@illinois.edu

Abstract

Legal reasoning requires selecting among competing defensible interpretations rather than retrieving single correct answers, yet current single-agent AI systems exhibit premature commitment and confirmation bias, struggling with the interpretive complexity of legal question-answering. We propose a Multi-Agent Debate (MAD) framework that enables systematic exploration of competing legal interpretations through structured adversarial argumentation. Multiple agents defend different positions while a judge synthesizes the debate into final decisions, forcing explicit consideration of alternatives and adversarial testing of citations. We will evaluate this approach on four legal examination datasets spanning US and Brazilian legal systems, testing whether MAD improves citation quality, corrects high-confidence errors, and exhibits diminishing returns across debate rounds. This work contributes a theoretically-grounded MAD architecture for legal QA, systematic evaluation across multiple-choice and open-ended question formats, and novel Brazilian legal examination benchmarks for under-represented legal systems in NLP.

1 Introduction

Legal reasoning fundamentally differs from factual question-answering—it requires selecting among competing defensible interpretations rather than retrieving a single correct answer. A constitutional law question about free speech, for instance, limits admits multiple legitimate interpretations depending on which doctrinal framework applies. Legal professionals construct arguments by weighing these alternatives and identifying the most legally sound position given specific facts and precedents. Bar examinations and magistrate exams deliberately test this interpretive ability through questions with multiple plausible alternatives in multiple-choice format or unbounded argument space in open-ended essays.

Current single-agent AI systems struggle with this interpretive complexity. Large language models exhibit *premature commitment*—generating an initial plausible interpretation and justifying it without systematically considering alternatives. Once committed to a position, they show *confirmation bias*, selectively citing authorities that support their chosen answer while overlooking contradictory precedents. This produces high-confidence errors on questions with subtle distinctions between plausible alternatives—precisely what legal exams are designed to test.

We propose Multi-Agent Debate (MAD) as a framework for systematic exploration of competing legal interpretations. Multiple agents defend different positions, engaging in structured argumentation that forces explicit consideration of alternatives. This addresses single-agent failures through three mechanisms: (1) forced alternative consideration prevents premature commitment by guaranteeing multiple interpretations receive development; (2) adversarial citation testing surfaces whether cited authorities genuinely apply or can be distinguished; (3) comparative evaluation frames questions as "which interpretation is legally superior?" rather than "is this interpretation correct?"

Our approach applies to both multiple-choice questions—where agents defend different answer choices and debate surfaces distinctions between plausibles alternatives—and open-ended questions—where agents construct arguments from scratch, with debate enabling comprehensive issue spotting, robustness testing, and multi-perspective synthesis. We test three hypotheses: MAD improves citation quality through adversarial testing (H1), disproportionately corrects high-confidence errors via forced reconsideration (H2), and shows diminishing returns after 2-3 rounds as interpretations converge (H3).

We contribute: (1) a MAD framework designed for legal QA's interpretive challenges across ques-

tion formats; (2) evaluation on four legal examination datasets spanning US and Brazilian legal systems, common and civil law traditions, and English and Portuguese languages; (3) analysis of when adversarial reasoning outperforms single-agent approaches through systematic ablations; (4) novel Brazilian legal examination datasets for underrepresented legal systems in NLP.

2 Related Work

2.1 Multi-Agent Collaboration

Recent advances in LLMs debates (Estornell and Liu, 2024) have spurred the development of autonomous multi-agent systems, which can be broadly categorized into cooperative and adversarial interactions. Cooperative frameworks such as the **role-playing / CAMEL** approach(Li et al., 2023) allow agents to collaborate autonomously on complex tasks through role-based cooperation and initial prompting, reducing human intervention.

Adversarial frameworks use debate-style interactions to improve consistency and factuality: for example, the **Multi-Agent Debate System** framework (Wang et al., 2024a) introduces a shared retrieval knowledge pool and adaptive knowledge selection to mitigate the "cognitive islands" problem, while(Du et al., 2023) show that sparse communication topologies among debating agents can reduce computation and sometimes improve factual accuracy and reasoning diversity.

2.2 Benchmarks

The **MMLU-Pro** benchmark (Wang et al., 2024b) enhances the original MMLU by introducing more difficult, reasoning-intensive tasks and a larger answer choice set, making it a more robust test of general knowledge and reasoning ability. In the legal domain, **LawBench** (Fei et al., 2023) evaluates LLMs' legal knowledge and reasoning within the Chinese civil law framework across 20 diverse tasks, assessing their capacity for classification, extraction, and generation. Similarly, **LegalBench** (Guha et al., 2023) provides a large, collaboratively developed benchmark of 162 tasks spanning multiple areas of legal reasoning, created by legal experts to reflect both practical and theoretical challenges. Our work contributes to this landscape by introducing high-complexity Brazilian legal examination datasets and evaluating multi-agent approaches specifically designed for the interpretive challenges of legal QA.

2.3 Legal Reasoning

LAR-ECHR (Chlapanis et al., 2024) introduces a Legal Argument Reasoning (LAR) task over European Court of Human Rights cases, where models must choose the correct next argument in a reasoning chain—offering a fine-grained evaluation of legal reasoning.

Yuan et al. (2024) propose a multi-agent system that decomposes complex legal reasoning, teaching LLMs to grasp legal theories and improve reasoning in tasks like confusing charge prediction.

LEXam (Fan et al., 2025) (Fan et al., 2025) is a benchmark based on 340 law exams (4,886 questions) that tests long-form and multi-step legal reasoning in both English and German, pairing open-ended questions with structured guidance and demonstrating the challenge LLMs still face in legal reasoning.

3 Proposed Approach

Goal. We test whether explicitly modeling the adversarial structure of legal reasoning via a multi-agent debate (MAD) architecture improves over a single-agent baseline on objective, exam-style legal QA. Our design follows the following frameworks, where multiple LLM agents propose answers, exchange rebuttals over a small number of rounds, and a judge model decides the final answer.

3.1 System Roles

- **Multiple Debaters**: argue for competing answers/interpretations.
- **Retriever**: a shared retrieval model (RAG) returning candidate authorities. Agents may selectively use or ignore retrieved evidence each round.
- **Judge**: reads the parties' briefs and decides on a final answer with a short rationale, explicitly checking (i) doctrinal relevance, (ii) citation correctness, (iii) factual application sufficiency.
- **(Optional) Cross-Examiner**: prompts each side with one targeted question to probe gaps or shared misconceptions before judgment.

3.2 Debate Rounds

Rather than fixing the number of debate rounds, we will **systematically vary the rounds** to measure

their effect on answer quality and citation correctness. Prior work on multi-agent debate (MAD) shows that allowing multiple rounds can improve performance, but also points to *diminishing returns* as rounds increase, motivating an explicit ablation over rounds.

Round schedule

1. **Round 0 (Retrieval Prep).** The Retreiver returns top-k authorities (short excerpts, source, jurisdiction).
2. **Round 1 (Opening).** Each Debater submits tokens and *must* cite any authority used.
3. **Round r (Rebuttal/Closing).** Each Debater (a) rebuts the opponent's strongest point; (b) may request one new snippet from the Retreiver (c) revises application.
4. **Judgment.** The Judge selects the final answer and publishes a 3-bullet rationale plus a 5-item checklist (doctrine named, on-point authority, accurate use, logical application, opponent addressed).

Rounds-as-a-factor. We evaluate different numbers of rounds to test whether more (or fewer) rounds yield better answers under our legal QA setting.

3.3 Communication Topology

Rather than fully connecting all agents each round, we will adopt a **sparse** visibility: in rebuttal/closing, each Debater only sees the opponent's immediately preceding brief; the Judge sees all. Sparse topologies can match or exceed fully connected debate while reducing compute and avoiding information overload.

3.4 Models and Role Assignment

We will systematically test how assigning different models to different roles affects answer quality.

Model assignments

1. **Homogeneous vs. heterogeneous debaters.**
   - *Same-model debaters* test whether stylistic alignment yields faster consensus but risks echoing errors.
   - *Mixed-model debaters* test whether *model heterogeneity* injects diversity that counteracts majority lock-in and improves factuality.
2. **Judge strength and identity.**
   - *Uniform Judge*: same model family as Debaters.
   - *Stronger Judge*: larger or higher-context model to improve aggregation.
   - *Cross-family Judge*: distinct family to reduce *self-preference* toward a particular generator's style.
3. **Judge awareness.**
   - *Blind judging*: Judge sees only anonymized briefs.
   - *Labeled judging*: Judge is told which model wrote which brief to stress-test bias detection.

3.5 Retrieval (RAG) Design

We implement a **shared evidence pool** indexed over the dataset's gold passages. Agents may *adaptively* cite or abstain each round; the Retreiver never fabricates text. Knowledge-enhanced debate with adaptive selection improves consistency and correctness across agents.

Hyperparameters:

- Debaters
- Rounds
- Topology
- Models & Roles
- Retrieval
- Interventions (optional)

3.6 Baselines

Baselines. (i) Single-agent with retrieval (no debate), (ii) Single-agent self-consistency with retrieval (vote among m samples).

4 Experimental Setup

4.1 Datasets

We evaluate our approach using both existing benchmarks and newly constructed datasets derived from official legal certification examinations. Unlike synthetic benchmarks, these assessments are developed and validated by legal authorities, ensuring authenticity and legal soundness. The progression from MMLU-Pro through bar examinations to magistrate examinations provides systematic complexity scaling across different legal traditions (common law vs. civil law) and languages

(English and Portuguese), enabling fine-grained analysis of multi-agent reasoning advantages.

MMLU-Pro We utilize the law section of MMLU-Pro (Wang et al., 2024b) as our baseline benchmark. MMLU-Pro extends the original MMLU by increasing difficulty through 10 answer choices (vs. 4) and more complex reasoning requirements.

US Bar Examination We construct datasets from two components of the Uniform Bar Examination: (1) **Multistate Bar Exam (MBE)**: multiple-choice questions across Constitutional Law, Contracts, Criminal Law, Evidence, Real Property, Torts, and Civil Procedure; (2) **Multistate Essay Exam (MEE)**: open-ended questions requiring written legal analysis integrating multiple areas of law.

Brazilian Bar Examination (OAB) The national certification for legal practice in Brazil: (1) **First Phase**: multiple-choice questions covering Constitutional, Administrative, Tax, Criminal, Civil, Labor, Business Law, and Legal Ethics; (2) **Second Phase**: subject-specific open-ended questions requiring detailed legal reasoning in Portuguese.

Brazilian National Magistrate Examination The highest-level legal competency assessment for judicial appointments: (1) **First Phase**: highly complex multiple-choice questions testing advanced legal knowledge and judicial reasoning; (2) **Second Phase**: varies by state, typically including essay questions and practical case analysis. All datasets are preprocessed into standardized formats. For open-ended questions, we develop evaluation frameworks based on official scoring guidelines. Detailed dataset statistics and preprocessing procedures are provided in Appendix A.

4.2 Evaluation

We will evaluate the proposed Multi-Agent Debate (MAD) framework on objective, exam-style legal QA tasks. Our evaluation focuses on three complementary dimensions: *answer accuracy*, *citation quality*, and *reasoning consistency*.

Answer Accuracy. Following prior work in multi-agent reasoning and legal QA, we compute the proportion of questions whose final answer chosen by the *Judge* exactly matches the gold answer. For open-ended questions in LEXAM, we adopt the official grading rubric (1 = fully correct, 0.5 = partially correct, 0 = incorrect) and report the normalized score. Accuracy is reported both per-dataset and averaged across domains.

Citation Quality. Because legal reasoning relies on precise doctrinal grounding, we will measure citation quality along two axes: (i) *Correctness*—whether the cited authority genuinely supports the chosen legal principle (binary judged by human annotators on 100 sampled cases); and (ii) *Coverage*—the fraction of gold-labeled authoritative sources that appear in the model's retrieved or cited set. We will also compute the *Citation F1* score combining correctness and coverage.

Reasoning Consistency. To quantify internal coherence of the debate process, we will compare the Judge's rationale and each Debater's argument using semantic similarity (embedding cosine score) and logical entailment detection. A debate will be marked *consistent* if the final rationale logically entails the winning debater's main premise. We will report the mean consistency rate across tasks.

Human Evaluation. For qualitative validation, three legal researchers will rate a subset of outputs on a 5-point Likert scale for *factual accuracy*, *legal relevance*, and *argument clarity*. Inter-rater agreement will be reported via Cohen's κ. Human evaluation will help identify improvements in interpretive reasoning that may not be captured by discrete accuracy metrics.

Statistical Significance. We will test performance differences between MAD variants and baselines using paired bootstrap resampling ($n = 10{,}000$) with $p < 0.05$ significance. All metrics will be reported with mean ± 95% confidence intervals.

Overall, this evaluation plan provides both quantitative and qualitative measures to assess how adversarial debate structures can improve factual accuracy, citation reliability, and reasoning coherence in legal question answering.

References

Odysseas S. Chlapanis, Dimitrios Galanis, and Ion Androutsopoulos. 2024. Lar-echr: A new legal argument reasoning task and dataset for cases of the european court of human rights. *arXiv preprint arXiv:2410.13352*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Andrew Estornell and Yang Liu. 2024. Multi-llm debate: Framework, principles, and interventions. *openreview preprint openreview: sy7eSEXdPC*.

Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. 2025. Lexam: Benchmarking legal reasoning on 340 law exams. *arXiv preprint arXiv:2505.12864*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *arXiv preprint arXiv:2303.17760*.

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2024a. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *arXiv preprint arXiv:2312.04854*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

A Datasets Details

subsectionMMLU-Pro MMLU-Pro (**?**) extends the original MMLU benchmark by increasing question difficulty through additional answer choices (10 options versus 4) and more complex reasoning requirements. The law subset contains questions spanning constitutional law, contracts, torts, criminal law, and civil procedure, providing standardized comparison with prior work on legal reasoning in LLMs. We use the publicly available dataset from HuggingFace.[1]

A.1 US Bar Examination Dataset

We construct a comprehensive dataset from two components of the Uniform Bar Examination:

Multistate Bar Exam (MBE) Multiple-choice questions testing fundamental legal principles across seven subject areas: Constitutional Law, Contracts, Criminal Law and Procedure, Evidence, Real Property, Torts, and Civil Procedure. Each question presents complex legal scenarios requiring application of multiple legal doctrines.

Multistate Essay Exam (MEE) Open-ended essay questions requiring written analysis of legal issues, application of relevant legal principles, and reasoned conclusions. Questions integrate multiple areas of law and demand sophisticated legal reasoning beyond factual recall.

A.2 Brazilian Bar Examination Dataset (OAB)

The Brazilian Bar Examination (Ordem dos Advogados do Brasil) serves as the national certification requirement for legal practice in Brazil, offering Portuguese-language evaluation of legal reasoning:

First Phase Multiple-choice questions covering Constitutional Law, Administrative Law, Tax Law, Criminal Law, Civil Law, Labor Law, Business Law, and Legal Ethics. Questions reflect Brazilian legal doctrine and civil law tradition, providing cross-jurisdictional comparison with common law systems.

Second Phase Subject-specific open-ended questions requiring written legal analysis and argumentation. Candidates must demonstrate application of Brazilian law to complex factual scenarios with detailed legal reasoning and citation of applicable statutes and precedents.

A.3 Brazilian National Magistrate Examination Dataset

The National Magistrate Examination represents the highest level of legal competency assessment in Brazil, used for judicial appointments:

[1] https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro

First Phase Highly complex multiple-choice questions testing advanced legal knowledge and judicial reasoning across all major legal domains. Questions require sophisticated interpretation of statutes, constitutional principles, and established jurisprudence.

Second Phase Varies by state jurisdiction, typically including essay questions and practical case analysis requiring demonstration of judicial reasoning capabilities at the highest professional level.

A.4 Dataset Preprocessing and Evaluation

All datasets are preprocessed into standardized formats with consistent structure for questions, answer options (where applicable), and ground truth answers. For multiple-choice questions, evaluation is straightforward accuracy measurement. For essay and open-ended questions, we develop evaluation frameworks based on official scoring guidelines and rubrics provided by the respective bar authorities, adapted for automated assessment while maintaining fidelity to legal evaluation standards.

B Theoretical Foundations: When Does Debate Improve Reasoning?

Not all reasoning tasks benefit from multi-agent debate. Drawing on prior work in multi-agent systems (**??**) and legal reasoning (**?**), we identify task characteristics that predict when adversarial architectures provide advantages over single-agent approaches.

Interpretive Ambiguity Tasks where multiple interpretations are initially plausible but differ in ultimate correctness particularly benefit from debate. Legal QA—both multiple-choice and open-ended—exhibits high interpretive ambiguity: exam questions deliberately include answer choices that invoke legitimate legal doctrines but may misapply them, or require constructing arguments where several framings are defensible. Single agents tend toward the first plausible interpretation they generate, while debate forces systematic consideration of alternatives.

Evidence-Based Reasoning Tasks requiring synthesis and application of multiple pieces of evidence to reach conclusions benefit from adversarial testing of which evidence is truly relevant. Legal reasoning fundamentally depends on citing relevant authorities and demonstrating their application to specific facts—a process that improves when agents must defend citations against critique. When Agent A cites Precedent P, Agent B can challenge whether P is actually "on point" or can be distinguished, forcing more rigorous citation practices.

Comparative Evaluation Structure Tasks where correctness is relative ("X is better than Y") rather than absolute ("X is true") align naturally with debate formats. Legal exam questions, particularly multiple-choice, are inherently comparative—selecting the "best" answer among options, not determining binary truth. Open-ended questions similarly require arguing why one legal framework or interpretation is superior to alternatives that a thorough analysis must address.

Mechanisms of Improvement We identify three mechanisms by which debate improves legal reasoning, corresponding to our hypotheses:

Error Detection via Adversarial Probing (H2): When agents defend opposing positions, they are incentivized to identify weaknesses in alternatives. This adversarial probing surfaces errors that self-review misses, particularly for high-confidence mistakes where single agents commit strongly to incorrect interpretations. Different agents may catch different error types (factual vs. logical vs. doctrinal), and explicit counterarguments force articulation of why alternatives fail.

Citation Grounding via Cross-Examination (H1): Legal reasoning requires not just citing authorities but demonstrating they apply to the case at hand. Adversarial debate subjects each citation to scrutiny—opposing agents challenge whether cited precedents are "on point," correctly interpreted, or distinguishable from the current case. This reduces both hallucinated citations (authorities that don't exist or don't say what is claimed) and irrelevant citations (authorities that exist but don't support the argument).

Interpretive Convergence and Diminishing Returns (H3): Over multiple rounds, agents refine interpretations through critique and response. Initial rounds typically generate the most new information as fundamentally different perspectives clash. However, as agents incorporate critiques and reasoning converges toward shared understanding of the strongest interpretation, additional rounds provide diminishing marginal value. This predicts performance plateaus after 2-3 rounds.

Application to MCQ vs. Open-Ended Questions These mechanisms apply to both question

formats but manifest differently. In multiple-choice questions, debate focuses on *comparative evaluation*—agents systematically explore why each given alternative is or is not legally superior. Error detection involves identifying misapplications of doctrine to specific facts, citation testing verifies that cited authorities distinguish between answer choices, and convergence occurs as agents agree on which alternative is most defensible.

In open-ended questions, debate emphasizes *constructive synthesis with adversarial refinement*. Error detection involves probing for missing issues or unstated assumptions in proposed analyses. Citation testing ensures comprehensive coverage of relevant authorities and proper application to facts. Convergence integrates multiple agents' insights into a coherent, thoroughly-reasoned response that addresses potential counterarguments.

This theoretical framework guides our experimental design: we systematically vary debate rounds to test H3, measure citation quality to test H1, and stratify analysis by single-agent confidence to test H2. Our evaluation assesses whether legal QA's high interpretive ambiguity, evidence-based structure, and comparative nature make it particularly suitable for multi-agent debate approaches.