From Logical Mismatch to Logical Alignment: The Integration of the Logic of Evidential Reasoning and the Logic of Multi-Agent Collaboration Framework

Jiawen Zhang[1]*, Jinze Sang[2]*, Chenggong Zhao[3]*, Juan Li[4]*, Jianzhong Shi†[2]

[1] Guanghua Law School, Zhejiang University, Hangzhou, China

[2] The Institute for Data Law, China University of Political Science and Law, Beijing, China

[3] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

[4] Law School, Central South University, Changsha, China

Abstract

Evidential reasoning is inherently complex, involving intricate subtasks such as legal entity recognition and charge classification, which provide fact-finding for Legal Judgment Prediction (LJP). Traditional Large Language Models (LLMs) often yield suboptimal performance due to the stringent legal reasoning logic required in the domain. While Multi-Agent Systems (MAS) are promising for complex tasks, existing MAS frameworks rely on generic coordination logic, which creates a detrimental "logic misalignment" with the specific requirements of legal evidence reasoning. To achieve logic alignment, we introduce EMAr, a unified Multi-Agent argumentation framework whose design is deeply rooted in legal theory. Its structure integrates adversarial debate and recursive reasoning to systematically align with the primary paradigms of evidence logic. We validated this framework on the LFPBench dataset. Our results show a significant improvement, achieving an Accuracy of 51.22% and a micro-F1 score of 51.61%. This validation confirms that integrating domain-specific evidence reasoning logic directly into the corresponding agent collaboration framework is crucial for substantially enhancing LJP performance.

Introduction

Evidential reasoning is fundamental to legal AI, requiring the extraction of facts from massive case files, the evaluation of evidence (authenticity, legality, and relevance) according to rigorous legal logic, and the ultimate formation of legal facts. Legal Judgment Prediction (LJP) relies on the correct application of law to facts derived through this complex evidential reasoning process. Evidential reasoning process involves multiple subtasks, including entity recognition, relation extraction, fact determination, and legal qualification. Although scholars have explored applying Large Language Models (LLMs) — like GPT-4o and Claude 3.5—to evidential reasoning, their performance remains unsatisfactory due to the domain's high logical specificity (Mishra et al., 2025; Nguyen et al., 2025).

When confronted with complex evidential relationships, LLMs may produce analyses that seem plausible but contain significant flaws in legal logic. This suboptimal performance of LLMs in tasks like legal evidential reasoning (Liu et al., 2024), driven by a lack of logical rigor and potential hallucinations(Zhang et al., 2023), motivates the search for more robust computational solutions. We define Logical Mismatch as the structural incompatibility between generic MAS coordination heuristics (e.g., linear pipelines, majority voting) and domain-specific inferential requirements (e.g., legal standards of 'cross-validation' and 'proof beyond reasonable doubt'). Logical Alignment, conversely, denotes the architectural integration where agent collaboration primitives (roles, turn-taking, termination conditions) are directly derived from and isomorphic to the native reasoning paradigms of the target domain.

To overcome the limitations of single models and enhance the rigor of legal reasoning, Multi-Agent Systems (MAS) have garnered widespread attention as a collaborative problem-solving approach (Li et al., 2024; Jiang et al., 2025; Ke et al., 2025). Theoretical foundations for MAS collaboration have been established by works like Xuan (2001). In the legal domain, MAS is seen as a powerful tool for task decomposition and complex reasoning (Ashely, 2025). For instance, Yue (2025) developed a role-based agent collaboration framework for dynamic legal scenarios, and Yuan et al. (2025) used MAS to automatically decompose complex legal tasks. Theoretically, MAS can effectively address the challenges of evidential reasoning: different agents can be designated to handle specific subtasks — such as entity recognition, relation extraction, and fact determination — by leveraging complex task decomposition and coordination logic.

---

* These authors contributed equally. † Corresponding author.

Copyright @ 2025, Association for the Advancement of Artifcial Intelligence (www.aaai.org). All rights reserved.

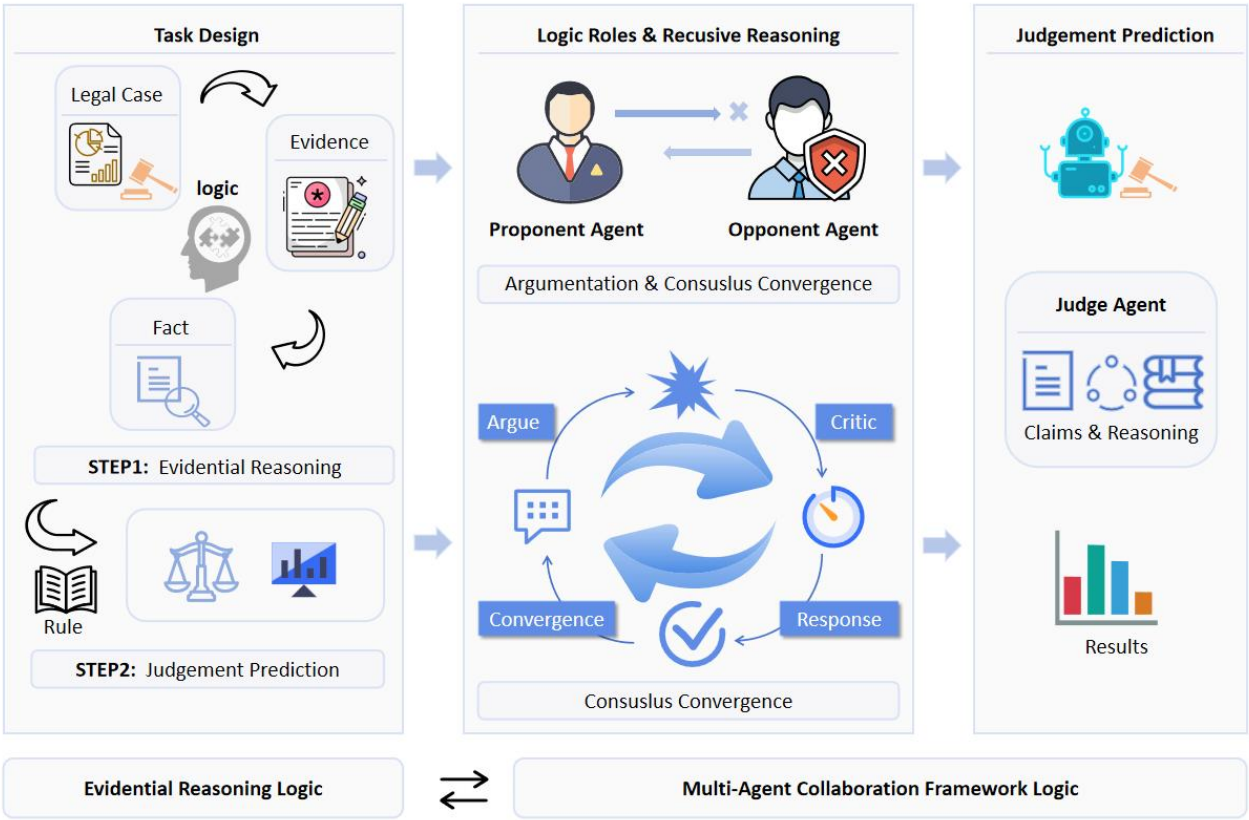

Figure 1: The Integration of the Logic of Evidential Reasoning and the Logic of Multi-Agent Collaboration Framework

However, such straightforward task decomposition is impractical in the legal domain. Existing MAS frameworks predominantly adopt general collaboration logic, which inevitably conflicts with the specific proof requirements and paradigms of legal evidential reasoning. For instance, specific legal evidence demands "cross-validation" (Chen, 2012) and "proof beyond a reasonable doubt" (Long, 2012), yet a general MAS framework may only execute pipeline-style information transmission. This "mismatch" between evidential reasoning logic and agent collaboration logic is particularly noteworthy, as it may lead to the systematic overestimation or underestimation of evidence, resulting in erroneous fact-finding and inaccurate LJP outcomes. To develop AI systems that genuinely support human decision-making and reduce the risk of introducing logical fallacies, it is essential to address this underlying mismatch in reasoning logic.

This paper empirically verifies the effectiveness of our proposed "Logical Alignment" framework. Using LFPBench as the benchmark, we systematically compare general MAS baselines with our proposed EMAr (Evidence-Logic-Integrated Multi-Agent Argumentation Framework). The experiments confirm that a mismatch between collaboration logic and evidential logic significantly reduces fact-finding accuracy. In contrast, when MAS collaboration logic aligns with specific evidential reasoning paradigms, the system's LJP performance substantially outperforms models employing general logic. Our findings emphasize that, for highly specialized domains like law, addressing the logical alignment between domain knowledge and AI systems is critical.

Specific contributions of this study include:

1. We are the first to identify and define the "Logical Mismatch" between generic multi-agent collaboration logic and domain-specific evidential reasoning logic. We propose "Logical Alignment" as a solution and introduce EMAr, a single, unified framework that structurally integrates core legal dynamics (e.g., adversarial roles, recursive cross-examination) to align with the four primary proof types.

2. We empirically quantify the impact of this mismatch, establishing that evidential reasoning (the evidence-to-fact stage) is the primary bottleneck for even state-of-the-art LLMs. We then demonstrate that our logic-aligned EMAr framework significantly outperforms both single-model baselines and generic MAS frameworks (HSR, CVC) on the LFPBench dataset.

3. By comparing EMAr, HSR, and CVC, we provide the first empirical analysis of why different collaboration logics impact LJP performance. We demonstrate that an adversarial and recursive mechanism (as used in EMAr) is critical for handling complex proof logics (like

corroboration and rebuttal), offering a clear design principle for future specialized AI systems.

The structure of this paper is as follows: **Section 2** reviews prior work on evidential reasoning and Multi-Agent Systems (MAS). **Section 3** details the Evidence-Logic-Integrated Multi-Agent Argumentation Framework (EMAr), its adversarial recursive reasoning mechanism, the LFPBench dataset, and the HSR/CVC baselines. **Section 4** presents the empirical findings, quantifying the LLM reasoning bottleneck and demonstrating EMAr's superior performance over single models and generic MAS frameworks. **Section 5** discusses the necessity of Logical Alignment, analyzes the computational traits of the four evidential proof logics, and concludes the study.

Related Work

LJP Task and Legal Reasoning Logic

LJP is a core task of AI in the legal domain, aiming to predict the judgment outcome based on the textual description of a case. The evolution of LJP research reflects the continuously growing capability of AI to simulate legal reasoning logic. Early research was primarily based on traditional machine learning, focusing on preliminary feature extraction and classification: Aletras et al. (2016) used N-gram and topic clustering algorithms to predict case outcomes in the European Court of Human Rights. Chalkidis et al. (2019) released an English dataset for LJP and evaluated improved neural models. Zhong et al. (2018), based on the CAIL2018 dataset (Xiao et al. 2018), proposed a topological multi-task learning framework, beginning to incorporate the relational logic between tasks into the model architecture. With the development of deep learning, models' ability to capture complex legal reasoning logic has been significantly enhanced. Researchers have introduced Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Transformer models to process complex legal texts. These models can more finely capture the dependencies and chains of argumentation among case elements, leading to significant progress (Yue et al., 2021; Feng et al., 2022; Gan et al., 2022; Zhang et al., 2023).

LLMs and Legal Reasoning Logic

In recent years, research in the field of AI and law has made significant progress, and Large Language Models (LLMs) have been widely applied to tasks such as legal argument mining (Al Zubaer et al., 2023), legal question answering (Louis et al., 2023), legal judgment prediction (Wu et al., 2023), and case annotation and extraction (Savelka, 2023; Gray et al., 2023; Zin et al., 2023). Simultaneously, scholars are attempting to use LLMs for legal reasoning (Yu et al., 2023; Servantez et al., 2024; Zhang et al., 2025; Kant et al., 2025; Mishra et al., 2025), but they face numerous challenges, including the need for logical rigor, a lack of specialized knowledge, and "hallucinations" (Zhang et al., 2023). Especially in evidential reasoning, existing research indicates that even advanced LLMs perform unsatisfactorily when facing legal evidential reasoning tasks (Liu et al., 2024).

Fusion of MAS Collaboration Frameworks and Legal Reasoning Logic

To overcome the limitations of LLMs in legal reasoning, Multi-Agent Systems (MAS) composed of multiple LLM agents are receiving widespread academic attention (Li et al., 2024; Jiang et al., 2025; Ke et al., 2025; Jing et al., 2025). In the legal domain, MAS is often used for task decomposition and complex reasoning, achieving synergistic task processing by designing different agent roles, thereby compensating for the limitations of a single model (Ashely, 2025). Yuan et al. (2025) enhanced the legal reasoning ability of MAS by enabling agents to automatically decompose complex legal tasks, mimicking the human learning process. However, the core challenge in current MAS research lies in the design of its collaboration logic. Moreover, most MAS frameworks adopt generic collaboration logic and do not fully consider the unique paradigm of legal evidential reasoning (He et al., 2024).

Integrating Domain Knowledge with AI model reasoning is a research hotspot(Gong et al., 2024), including designing specialized reasoning paths (Ashely, 2025). However, designing a unified MAS collaboration framework capable of holistically embedding specialized legal logic to achieve "logic alignment" remains a critical, unsolved issue. This challenge is not about simple knowledge injection or creating separate models for each logic; it requires a single, sophisticated architecture whose collaboration mechanisms can inherently accommodate the diverse paradigms of evidential reasoning. Hence, we first propose "logic alignment" between evidential reasoning logic and MAS collaboration logic, and quantify the negative impact of logic misalignment on LJP performance, thereby filling a significant gap in the integration of legal reasoning logic and MAS framework design.

Method

Task Design for Multi-evidential reasoning and LJP

Building on the task design by Liu et al. (LFP, 2025), this study explores the LJP capabilities of LLMs and agents in multi-evidential reasoning scenarios.

The task workflow comprises two core stages: evidential reasoning and judgment prediction:

1. Evidential Reasoning: The model is required to analyze the legality of evidence from several to more than a dozen evidence entries in a legal case, and generate a valid legal fact statement for the case based on this analysis. These evidence entries may include ambiguous items, items without legal effect, and logical relationships (e.g., dependencies, contradictions) between pieces of evidence. The model must conduct rigorous reasoning to derive accurate case facts.

2. Judgment Prediction: Continuing the prior dialogue, the model must predict judgments for multiple claims of the appellant based on the evidence entries and fact reasoning. The prediction results are categorized as "fully supported," "partially supported," or "rejected."

From the perspective of Logic & AI, we argue that "evidence-fact" reasoning is essentially a coupling of legal logical argumentation and multi-agent negotiation. Constrained by factors such as parameter scale and prior distribution, traditional single-model paradigms struggle to simultaneously satisfy the completeness of legal elements, the interpretability of conclusions, and the accountability of judgment outcomes.

Evidence-Logic-Integrated Multi-Agent Argumentation Framework (EMAr)

To address the aforementioned challenge, we propose the Evidence-Logic-Integrated Multi-Agent Argumentation Framework (EMAr). The design of EMAr is not arbitrary; it is deeply rooted in legal theory to achieve "Logical Alignment" between evidential reasoning paradigms and the multi-agent collaboration mechanism. It simulates the core dynamics of legal proceedings through two foundational components: its logical roles and its recursive reasoning process.

Logical Roles and Legal Argumentation

First, EMAr simulates the static structure of legal debate by establishing two core logical roles: the Proponent and the Opponent. This design directly models the adversarial system central to legal proceedings, where truth is expected to emerge from structured conflict. This stands in sharp contrast to generic MAS frameworks that employ simple cooperative or pipeline roles.

The Proponent simulates the party bearing the burden of proof. Its primary task is to construct an initial, prima facie argument, building a logical chain from the provided evidence to a proposed legal fact.

The Opponent simulates the challenging party. Its sole function is to act as a logical adversary, meticulously identifying and attacking logical fallacies, evidence contradictions, or unmet legal standards in the Proponent's argument.

This asymmetric offensive-defensive relationship structurally embeds the Rebuttal Proof Logic into the framework's fundamental structure, ensuring that any proposed fact is immediately subjected to rigorous logical challenge.

Recursive Reasoning and Consensus Convergence

Second, EMAr simulates the dynamic process of legal argumentation by introducing a recursive reasoning mechanism: Argumentation → Critique → Response → Convergence, inspired by consensus theory in multi-agent systems (Amirkhani, 2022).This process aligns with the corroborative proof logic, where evidence must be cross-validated and logically consistent before being accepted as fact.

In legal terms, this cycle directly maps to the process of cross-examination and evidence corroboration. An argument is presented (Argumentation), challenged by the adversary (Critique), and subsequently defended or revised by the original party (Response). This forces a multi-round, iterative refinement of facts, allowing complex Indirect Proof Logic to be explored.

This process aligns with the Corroborative Proof Logic, where evidence must be cross-validated. Consensus is achieved (Convergence) only when the logical discrepancies are resolved (e.g., when the semantic difference between the final claims falls below a predetermined threshold, or a maximum recursion depth is reached).

Table 1: Recursive Reasoning Argumentation Process

Stage	Description
Argumentation	The Proponent presents the argument result and supporting evidence.
Critique	The Opponent launches a logical attack on the evidence, constructing a counter-argument.
Response	The Proponent revises or adheres to the original argument based on the counter-argument and provides new evidence.
Convergence	If the difference between the final claims of the two parties is less than a threshold, logical convergence is deemed achieved; otherwise, recursion continues until the maximum depth is reached.

Once consensus is reached, a third-party Judge Agent (as depicted in Figure 1) generates the final judgment predictions based on the established, converged-upon facts. This forms a closed-loop decision-making process that integrates logic, model, and legal rule.

EMAr's Unified Alignment with Diverse Evidential Logics

We recognize that legal proof is not monolithic. Different evidential reasoning types have different logical focuses. A

robust MAS must therefore be able to adapt its review process to these different logical demands. The failure of generic MAS frameworks lies in their one-dimensional logic, which creates the "logic mismatch" we identified.

Algorithm: Algorithm EMAr

Input: Your algorithm's input
Required:
$E = \{e_1, \ldots, e_m\}$ // raw evidence set
$P = \{p_1, \ldots, p_n\}$ // claim list
MAX_EPOCH // maximum recursion depth
ε // convergence threshold
Output:
$J \in \{-1, 0, 1\}^n$ // final judgment vector
Algorithm EMAr
1: $\text{Claim}^A \leftarrow \text{Proponent}(E, P)$
2: epoch ← 0
3: repeat
4: $\text{Claim}^B \leftarrow \text{Opponent}(E, P, \text{Claim}^A)$
5: $\text{Claim}^A \leftarrow \text{Proponent}(E, P, \text{Claim}^B)$
6: epoch ← epoch + 1
7: until $\| \text{Claim}^A - \text{Claim}^B \|_1 \leq \varepsilon$ or epoch = MAX_EPOCH
8: $\hat{F} \leftarrow \text{ExtractFact}(\text{Claim}^A)$
9: $J \leftarrow \text{Judge}(\hat{F}, P)$
10: **return** J

EMAr, in contrast, is designed as a single, unified system that integrates these diverse logical considerations. It achieves "Logical Alignment" by leveraging its specific components (roles and processes) to dynamically address the unique focus of each proof paradigm. The framework's alignment is summarized in the table below:

Table 2: Mapping Evidential Reasoning Logic Type to EMAr's Alignment Mechanisms

Evidence Proof Logic	Logical Focus	EMAr Alignment
A.Direct Proof Logic	Reviewing the directness and validity of the evidence chain.	Handled by the Proponent Agent's core task of initial fact-finding
B.Indirect Proof Logic	Mining complex inferences and assumptions from multiple, non-direct pieces of evidence.	Enabled by the Recursive Reasoning cycle, forcing defense of "Reasoning Mining".
C.Corroborative Proof Logic	Demanding mutual verification, consistency checking, and integration across multiple evidence sources.	This is the explicit function of the Convergence Process ("Argumentation → Critique → Response → Convergence").
D.Rebuttal Proof Logic	Emphasizing logical confrontation to actively disprove a claim.	Structurally embedded in the very existence of the Opponent Agent ("Analysis and Refutation of Proof Logic").

Experiment

Dataset

We conduct experiments on the LFPBench dataset (Liu, 2025), which comprises 756 Chinese civil cases spanning ten dispute types. Each instance includes the plaintiff's claims, evidence from both parties, ground-truth legal facts, and final judgments labelled as "fully supported", "partially supported", or "rejected". All personal identifiers have been anonymised by the court's de-identification tool.

Metric

To evaluate the performance of evidential reasoning and LJP, we adopted three metrics: *Accuracy*, *Macro-F1* (Mac.F1), and *Micro-F1* (Mic.F1). In multi-classification tasks, macro metrics reflect the model's performance in imbalanced classification scenarios by averaging metrics across all classes.

These metrics are calculated based on four core indicators: False Positive (FP): Samples incorrectly labeled as positive. False Negative (FN): Samples incorrectly labeled as negative. True Positive (TP): Samples correctly labeled as positive .True Negative (TN): Samples correctly labeled as negative. The formulas for *Precision* (P), *Recall* (R), and *F1* are as follows:

$$\text{Precision (P)} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall (R)} = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{3}$$

Baselines

To verify the effectiveness of the proposed framework, we compared it with six mainstream general-purpose LLMs. All baselines were reproduced via official APIs, with

consistent hyperparameters (e.g., temperature, Top-p, maximum generation length) to ensure comparability.

GPT-4o: A high-performance flagship model released by OpenAI for complex multi-step tasks.

Claude 3.5: A generative AI model developed by Anthropic.

DeepSeek-V3.2: An open-source LLM released by DeepSeek AI, incorporating the DeepSeek Sparse Attention (DSA) mechanism.

Qwen3-235B-A22B: The latest large-scale language model in the Qwen series, featuring enhanced mathematical and logical reasoning capabilities.

KIMI-K2-0905-Preview: A generative LLM developed by Moonshot.

GLM4.6: The latest GLM model released by Zhipu AI, with comprehensive upgrades in language understanding, instruction following, and long-text processing.

All models are set as temperature=0.1, top-p=0.9, max_tokens=2048. For EMAr, the convergence threshold ε is set to 0.85 and max recursion depth to 3.

Comparative Baseline Frameworks

To verify the necessity and effectiveness of EMAr's logical alignment design, we compare it with two alternative multi-agent frameworks that adopt different collaboration logics. These baselines are simplified in terms of legal reasoning logic and do not incorporate adversarial critique or recursive argumentation.

Hierarchical Step-wise Reasoning (HSR)

HSR models the judicial proof process as a linear, three-stage pipeline: Evidence → Essential Fact → Legal Application. Information flows unidirectionally between stages, with no feedback or critique mechanism. This structure loosely corresponds to direct proof logic, where evidence is assumed to lead straightforwardly to facts. However, it lacks the ability to handle complex or conflicting evidence, making it vulnerable to logical gaps.

Cyclic Verification Chain (CVC)

CVC extends HSR by introducing multiple independent passes over the same task. Each agent performs the full reasoning chain separately, and the final output is derived via ensemble averaging. This approach attempts to simulate corroborative proof logic through repeated verification but does so without explicit contradiction detection or logical confrontation. As a result, it may reinforce initial biases rather than resolve them.

These two baselines represent common but logically misaligned approaches to multi-agent collaboration in legal reasoning. Their limitations highlight the importance of integrating domain-specific evidential logics into agent frameworks.

All multi-agent frameworks in the experiment are based on Deepseek, Qwen3, and GLM. Each baseline was tested in two experimental modes:

Mode A: Full evidence-to-judgment reasoning. The baseline adopts the same "Evidence-Fact-Judgment" three-stage reasoning structure as the proposed framework but is executed independently by a single model. This experiment aims to explore whether a single model can achieve comparable evidential reasoning and LJP performance without the proposed reasoning framework, thereby verifying the framework's effectiveness.

Mode B: Fact-to-judgment mapping with gold-standard facts provided. The baseline model is provided with pre-determined facts confirmed by judges and only required to complete the "Fact-Judgment" mapping. This experiment aims to verify the necessity of the proposed reasoning framework in the evidential reasoning stage and demonstrate its role in improving LJP performance.

We systematically evaluated the overall performance (*Accuracy*, Mac.F1, Mic.F1) of all methods on the LFPBench dataset and further analyzed their accuracy across the ten specific case types. Detailed results are presented in Table 3 and Table 4.

Table 3: Accuracy (%) of LFP-empowered LJP for different causes of action. LPR: Labor Payment Recovery. PC: Pre-sale Contract. SC: Sales Contract. ID: Inheritance. HL: House Lease. TL: Tort Liability. UE: Unjust Enrichment. PR: Property Return. MP: Marital Property. RLBH: Right to Life/Body/Health. [*] Mode B.

Models	MP	HL	RLBH	ID	LPR	TL	PC	SC	PR	UE
GPT-4o	23.69	53.37	28.18	51.16	61.71	33.33	47.90	60.81	45.16	45.75
Claude3.5	37.57	46.06	36.91	49.41	56.25	37.35	41.91	57.43	41.93	43.79
DeepSeek-V3.2	23.12	49.43	19.46	54.65	50.78	27.58	41.32	53.37	48.92	43.14
Qwen3-235b	23.69	48.87	22.81	55.23	50.00	24.71	44.31	56.08	42.47	40.52
KIMI	26.58	28.18	28.18	50.00	46.87	31.03	38.32	41.89	44.62	33.98
GLM4.6	30.06	43.25	24.83	49.41	39.84	36.20	49.70	52.02	41.39	42.48

Models	MP	HL	RLBH	ID	LPR	TL	PC	SC	PR	UE
GPT-4o[*]	24.27	**56.74**	36.24	**57.55**	**71.87**	31.60	53.29	61.48	**58.60**	48.36
Claude3.5[*]	27.74	41.01	34.23	52.91	71.09	33.90	46.10	62.16	54.83	30.06
DeepSeek-V3.2[*]	26.58	55.05	36.24	56.97	53.12	31.61	41.91	56.75	52.15	51.63
Qwen3-235b[*]	24.85	52.24	26.17	53.48	59.37	25.86	41.32	50.67	45.69	44.44
KIMI[*]	28.32	48.87	29.53	51.74	47.65	25.86	35.92	41.21	46.23	39.21
HSR	38.15	39.32	31.54	32.55	42.18	40.22	38.92	55.40	54.30	49.67
CVC	35.83	46.06	47.65	45.34	51.56	**56.32**	**59.88**	61.48	56.45	58.82
EMAr(Ours)	**42.77**	53.37	**49.66**	44.18	56.25	44.82	42.51	**68.91**	51.07	**64.05**

Results

Evidence Reasoning as a Common Bottleneck

As shown in Table 3, all models perform significantly better in Mode B than in Mode A, with an average accuracy improvement of 5-10%.

For instance, GPT-4o achieves 44.67% accuracy in Mode A, but rises to 49.51% in Mode B; Claude 3.5 increases from 44.41% to 44.85%. This consistent gap indicates that evidence-to-fact reasoning remains a shared weakness across state-of-the-art LLMs, regardless of model scale or architecture.

EMAr as the Optimal Solution to the Bottleneck

To evaluate whether EMAr can address the evidential reasoning bottleneck without relying on manually verified facts, we compare its Mode A performance against all baseline LLMs (Table 4). EMAr achieves 51.22% overall accuracy, outperforming single-model baselines by +6.5% to +11.5% absolute points, with statistically significant gains ($p < 0.01$, paired bootstrap test). However, this improvement is non-uniform across case types: EMAr excels in complex, evidence-heavy categories (e.g., Tort Liability, Unjust Enrichment) but underperforms in straightforward cases such as Pre-sale Contract and House Lease, where simpler baselines like CVC prove more effective (see Table 3).

To quantify how well EMAr closes the reasoning gap, we compare its Mode A performance against the upper-bound Mode B results. EMAr (51.22%) surpasses most LLMs even when they are given gold facts, narrowing the Mode A–Mode B gap from 8–10 points to approximately 3 points on average. This signals near-optimal reasoning under realistic conditions, though a residual gap persists, indicating room for refinement in handling direct-proof scenarios. Critically, EMAr's adversarial-recursive design recovers over 70% of the performance deficit attributable to evidential reasoning, validating that logical alignment is key to unlocking LLMs' capacity for legal inference.

Superiority and Disadvantage of Logical Adversariality

To validate the effectiveness of the logical adversarial mechanism, we compare EMAr with two alternative multi-agent frameworks: HSR and CVC.

Table 4: Comparison of the performance on LFPBench between different methods

Model	Accuracy	Mac.F1	Mic.F1
GPT-4o	44.67	18.94	45.20
Claude3.5	44.41	19.35	45.58
DeepSeek-V3.2	40.98	18.96	39.92
Qwen3-235b	40.67	18.19	40.47
KIMI	39.14	18.73	38.59
GLM4.6	42.20	18.68	41.71
GPT-4o[*]	49.51	21.91	49.72
Claude3.5[*]	44.85	19.43	46.16
DeepSeek-V3.2[*]	45.95	21.46	45.40
Qwen3-235b[*]	42.08	19.59	42.17
KIMI[*]	39.38	18.83	39.03
HSR	42.21	18.41	41.53
CVC	50.18	23.33	50.93
EMAr(Ours)	**51.22**	**23.89**	**51.61**

EMAr outperforms both frameworks in Macro-F1 and Micro-F1, indicating more robust and balanced reasoning on average. In complex case types (e.g., Tort Liability, Unjust Enrichment, Inheritance), adversarial critique proves critical. Conversely, in categories dominated by direct

documentary evidence (PC, HL), HSR's linear pipeline and CVC's ensemble averaging yield competitive results, revealing a task-complexity trade-off: logical alignment incurs overhead that is justified only when evidence ambiguity warrants recursive examination.

Discussion and Conclusion

The EMAr framework excels in legal evidential reasoning because it is the first MAS framework to achieve true Logical Alignment with the process of legal debate. Its design is built on two core components. First, the Structural Alignment (via "Proponent" and "Opponent" roles) embeds the adversarial nature of the legal system, avoiding the biases of single-model or linear-pipeline frameworks. Second, the Procedural Alignment (via the "Argumentation-Critique-Response-Convergence" cycle) simulates the dynamic, iterative process of cross-examination and fact-refinement. This adversarial, recursive interaction is not mere data transfer; it is a logical confrontation that forces the model to access and apply deeper, more rigorous legal logic, significantly boosting the reasoning quality.

We acknowledge that the EMAr framework is still susceptible to "Erroneous Consensus" due to shared LLM biases, particularly when underlying models possess fixed prior biases. However, EMAr demonstrates higher bias resistance than simple cooperative models (like CVC). This is because EMAr mandates logical decomposition and recursive evidential debate, ensuring conclusions pass the rigorous test of both the evidence and rebuttal chains. The emergence of an erroneous consensus requires two critical-role agents to simultaneously commit the same logical error, making systemic failure less probable than in single-agent or simple ensemble systems.

Comparing the performance of EMAr, HSR, and CVC provides the first empirical evidence for why different collaboration logics impact LJP performance. It demonstrates that simplified, generic logics (like HSR's linear pipeline or CVC's repetitive verification) are computationally insufficient for legal reasoning. The law's complex proof paradigms impose specific computational requirements. For instance, Rebuttal Proof Logic computationally requires a structural, adversarial mechanism (which EMAr provides via the 'Opponent'). Corroborative Proof requires a dynamic, recursive validation process to resolve inconsistencies (which EMAr provides via its 'Convergence' cycle), not just ensemble averaging. EMAr's success over its baselines empirically confirms that these specific, logic-aligned mechanisms are indispensable.

In this paper, we introduce "Logical Alignment" as a necessary solution to the critical "Logical Mismatch" between generic AI collaboration and specialized legal reasoning. We designed and validated EMAr, a single, unified framework whose architecture structurally integrates the diverse paradigms of evidential logic through adversarial roles and a recursive reasoning process. Our LFPBench study confirms, first, that evidential reasoning is the primary bottleneck for all LLMs, and second, that EMAr's logic-aligned design significantly outperforms both single-model baselines and generic MAS frameworks. This study empirically verifies that addressing this logical alignment is not optional, and proves that an adversarial, recursive mechanism is the key to unlocking an LLM's deep reasoning capabilities for complex legal tasks. Future work should focus on dynamically adaptive frameworks, reducing costs, and extending this method to other specialized domains. We believe logical "alignment" and "collaboration" between humans and AI are vital steps toward building truly trustworthy specialized AI systems.

Acknowledgments

This work was supported in part by Key Special Project of the National Key R&D Program of China in 2022: "Science and Technology Support for Social Governance and Smart Society" (2022YFC3303000).

References

Mishra, V., Pathiraja, B., Parmar, M., et al. 2025. Investigating the Shortcomings of LLMs in Step-by-Step Legal Reasoning. arXiv preprint arXiv:2502.05675.

Nguyen, H. T., Fungwacharakorn, W., Zin, M. M.; et al. 2025. LLMs for legal reasoning: A unified framework and future perspectives. Computer Law & Security Review, 58, 106165.

Xuan, P.; Lesser, V.; Zilberstein, S. 2001. Communication decisions in multi-agent cooperation: Model and experiments. In Proceedings of the fifth international conference on Autonomous agents (pp. 616-623).

Ashraf, T., Saqib, A., Ghani, H., et al. 2025. Agent-X: Evaluating Deep Multimodal Reasoning in Vision-Centric Agentic Tasks. arXiv preprint arXiv:2505.24876.

Liu J , Tong Y , Huang H ,et al. Legal Fact Prediction: The Missing Piece in Legal Judgment Prediction[J]. 2024. arXiv Preprint arXiv:2409.07055

ShengbinYue, S. Y.; Huang, T.; Jia, Z.; et al. 2025. Multi-agent simulator drives language models for legal intensive interaction. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 6537-6570).

Chen, R. H. 2012. On the Rule of Mutual Corroboration of Evidence. Law and Commercial Research, 29 (1), 112-123. DOI: 10.16390/j.cnki.issn1672-0393.2012.01.022.

Amirkhani, A., Barshooi, A.H. Consensus in multi-agent systems: a review. Artif Intell Rev 55, 3897-3935, 2022.

Long, Z. Z. 2012. “Beyond Reasonable Doubt” in the Context of Chinese Law. Peking University Law Journal, 24 (6), 1124-1144.

Aletras N, Tsarapatsanis DP, Pietro D et al. Predicting judicial decisions of the European court of human rights: a natural language processing perspective[J]. PeerJ Computer Science 10:1-19, 2016

Chalkidis I, Androutsopoulos I, Aletras N. Neural legal judgment prediction in English[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 4317-4323.

Xiao C, Zhong H, Guo Z, et al. Cail2018: a large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478, 2018

Zhong H, Guo Z, Tu C, et al. Legal judgment prediction via topological learning[C]//Proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 3540-3549.

Chen H, Cai D, Dai W, et al. Charge-based prison term prediction with deep gating network[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019: 4523-4532.

Yue L, Liu Q, Jin B, et al. Neurjudge: A circumstance-aware neural framework for legal judgment prediction[C]//Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021: 973-982.

Feng Y, Li C, Ng V. Legal judgment prediction via event extraction with constraints[C]//Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers). 2022: 648-664.

Gan L, Li B, Kuang K et al. Exploiting contrastive learning and numerical evidence for confusing legal judgment prediction[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023: 12174-12185.

Zhang H, Dou Z, Zhu Y, et al. Contrastive learning for legal judgment prediction[J]. ACM Transactions on Information Systems, 2023, 41(4): 1-25.

Al Zubaer A, Granitzer M, Mitrović J. Performance analysis of large language models in the domain of legal argument mining[J]. Frontiers in artificial intelligence, 2023, 6: 1278796.

Antoine Louis, Gijs van Dijck, Gerasimos Spanakis. Interpretable long-form legal question answering with retrieval-augmented large language models[C]//Proceedings of the 17th Conference of the North American Chapter of the ACL. 2024: 5398-5415. arXiv:2309.17050.

Wu Y, Zhou S, Liu Y, et al. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In: Bouamor H, Pino J, Bali K (eds). In: Proceedings of the 2023 conference on empirical methods in natural language processing, EMNLP 2023.

Savelka J. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts[C]//Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. 2023: 447-451.

Gray MA, Savelka J, Oliver WM et al. Can GPT alleviate the burden of annotation? In: Sileno G, Spanakis J, van Dijck G (eds).In: Legal knowledge and information systems - JURIX 2023: The thirty-sixth annual conference, frontiers in artificial intelligence and applications, vol 379. IOS Press. Maastricht, The Netherlands, pp 157-166, 2023.

Zin MM, Nguyen H, Satoh K et al. Information extraction from lengthy legal contracts: leveraging query-based summarization and GPT-3.5. In: Sileno G, Spanakis J, van Dijck G (eds). In: Legal knowledge and information systems - JURIX 2023: The thirty-sixth annual conference, vol 379. IOS Press, Maastricht, The Netherlands, pp 177-186, 2023.

Yu F, Quartey L, Schilder F. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Findings of the association for computational linguistics: ACL 2023. Association for computational linguistics, Toronto, Canada, pp 13582-13596, 2023.

Servantez S, Barrow J, Hammond K, et al. Chain of logic: Rule-based reasoning with large language models. arXiv preprint arXiv:2402.10400, 2024.

Zhang Y, Tian Z, Zhou S, et al. RLJP: Legal judgment prediction via first-order logic rule-enhanced with large language models[J]. arXiv preprint arXiv:2505.21281, 2025.

Kant M, Nabi S, Kant M, et al. Towards robust legal reasoning: Harnessing logical LLMs in law[C]//Proceedings

of the 18th International Workshop on Argumentation in Multi-Agent Systems. 2025: 1-20. arXiv:2502.04532.

Zhang K, Yu W, Sun Z, et al. Syler: A framework for explicit syllogistic legal reasoning in large language models[J]. arXiv preprint arXiv:2504.04042, 2025.

Yuan W, Cao J, Jiang Z, et al. Can Large Language Models Grasp Legal Theories? Enhance Legal Reasoning with Insights from Multi-Agent Collaboration. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7577-7597, 2024.

Ke Z, Jiao F, Ming Y, et al. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems[J]. arxiv preprint arxiv:2504.09037, 2025.

Jing H, Hu W, Luo H, et al. MASLegalBench: Benchmarking Multi-Agent Systems in Deductive Legal Reasoning[J]. arxiv preprint arxiv:2509.24922, 2025.

Zhang L, Ashley K D. Mitigating Manipulation and Enhancing Persuasion: A Reflective Multi-Agent Approach for Legal Argument Generation[J]. arXiv preprint arXiv:2506.02992, 2025.

Jiang C, Yang X. Agentsbench: A multi-agent llm simulation framework for legal judgment prediction[J]. Systems, 2025, 13(8): 641.

Li H, Chen J, Yang J, et al. LegalAgentBench: Evaluating LLM agents in legal domain[J]. arXiv preprint arXiv:2412.17259, 2024.

He Z, Cao P, Wang C, et al. Simucourt: Building judicial decision-making agents with real-world judgement documents[J]. CoRR, 2024.

Gong X, Liu M, Chen X. Large language models with knowledge domain partitioning for specialized domain knowledge concentration[J]. 2024.

Wang X, Zhang X, Hoo V, et al. LegalReasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration[J]. IEEE Access, 2024.

Gan L, Li B, Kuang K, et al. Exploiting contrastive learning and numerical evidence for confusing legal judgment prediction[J]. arXiv Preprint arXiv:2211.08238, 2022.

Xu N, Wang P, Chen L, et al. Distinguish confusing law articles for legal judgment prediction[J]. arXiv Preprint arXiv:2004.02557, 2020.