Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence
Canadian AI 2026
25–29 May 2026, Vancouver, BC

Neuro-Symbolic Adaptive Collaboration of Arena-Based Argumentative LLMs for Contestable Legal Reasoning

Hoang-Loc Cao[†*§], Phuc Ho[†§], Truong Thanh Hung Nguyen[‡],
Phuc Truong Loc Nguyen[◇], Dinh Thien Loc Nguyen[†], Hung Cao[‡]
[†]University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[‡]Analytics Everywhere Lab, University of New Brunswick, Canada
[◇]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Abstract

Legal reasoning requires not only accuracy but also verifiable, contestable justification. Existing LLM methods, such as Chain-of-Thought (CoT) and Retrieval-Augmented Generation (RAG), often produce unstructured explanations that lack formal verification or user intervention. We propose *Adaptive Collaboration of Argumentative LLMs* (ACAL), a neuro-symbolic framework that combines multi-agent collaboration with an Arena-based Quantitative Bipolar Argumentation Framework (A-QBAF). ACAL constructs and resolves arguments adaptively, escalates uncertain cases, and supports Human-in-the-Loop (HITL) contestability by allowing users to inspect and modify the reasoning graph. Experiments on LegalBench show that ACAL outperforms strong baselines while improving transparency and contestability.[1].

Keywords: Legal Reasoning, Contestable AI, Neuro-Symbolic AI

1. Introduction and Related Work

Legal reasoning applies legal rules and principles to case facts to reach justified outcomes, requiring issue identification, legal retrieval, statutory or precedential interpretation, and persuasive argumentation [1–3]. Recent LLM-based methods have shown strong potential in these tasks, especially when augmented with domain knowledge, legal-specific prompting, or structured reasoning strategies [4, 5]. Retrieval-augmented generation (RAG) improves factual grounding by linking model outputs to legal sources such as statutes, case law, and regulatory texts [6, 7], while chain-of-thought (CoT) prompting elicits stepwise reasoning and has inspired law-specific variants for more robust legal analysis [4, 5]. Multi-agent frameworks further extend this direction by simulating adversarial or collaborative legal discourse, where agents critique and refine one another's arguments to improve robustness and decision quality [8–11].

Despite these advances, current approaches remain limited in transparency and contestability. Prompting-based methods, including CoT and few-shot prompting, often generate free-form explanations that are difficult to verify or systematically challenge [4, 12]. RAG strengthens factual support, but its final decision logic often remains implicit [6]. Multi-agent debate (MAD) introduces diverse perspectives, yet debate transcripts do not provide a formally contestable reasoning structure, and performance gains can be inconsistent across tasks [8, 13]. Prior studies also report that frontier models may produce invalid arguments or irrelevant citations in complex legal settings, raising concerns about trust, fairness, and accountability [7, 14]. At the same time, growing regulatory emphasis on transparency and recourse increases the need for systems whose reasoning can be inspected and disputed [15].

Related work on explainable and contestable legal AI has begun to address this gap. Some systems provide citations or structured reasoning traces for post-hoc inspection [16],

[1]Our implementation is available at: https://github.com/loc110504/ACAL

[§]These authors contributed equally.
[*]Corresponding author: chloc22@clc.fitus.edu.vn

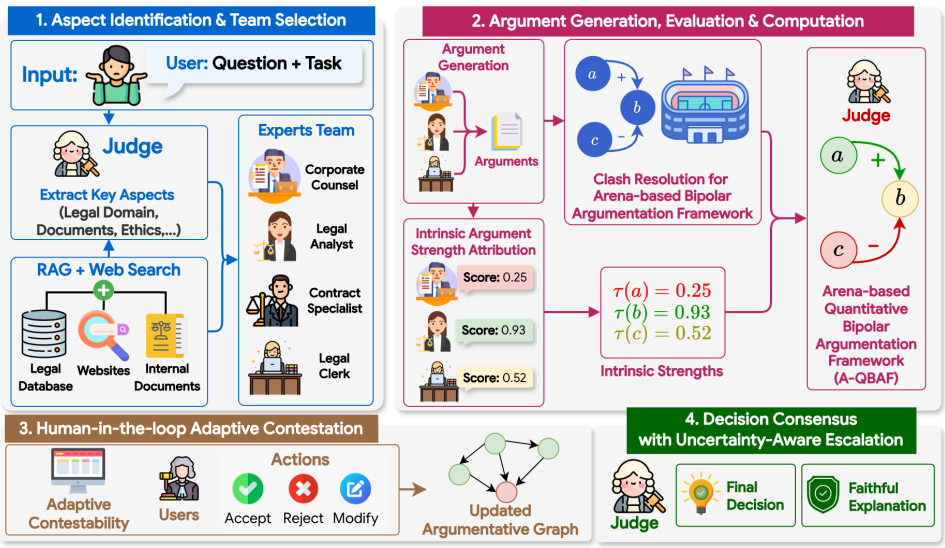

Figure 1. Our proposed Adaptive Collaboration of Argumentative LLMs.

while others use formal argumentation schemes to represent competing legal claims more explicitly [16]. Neuro-symbolic approaches also show promise for connecting natural language reasoning with symbolic legal structures [13]. However, these methods typically produce static artifacts for inspection rather than a dynamic framework that quantitatively weighs conflicting arguments and formally propagates user interventions to the final judgment.

To address these limitations, we propose *Adaptive Collaboration of Argumentative LLMs (ACAL)*, a structured neuro-symbolic framework for explainable and contestable legal reasoning. ACAL integrates argumentative LLMs with an arena-based quantitative bipolar argumentation framework, adaptive role selection, clash resolution, and uncertainty-aware escalation. Unlike prior approaches, it also supports a human-in-the-loop (HITL) workflow, allowing users to directly inspect and revise the reasoning graph, with those changes formally affecting the final outcome. Experiments on LegalBench show that ACAL outperforms strong prompting, RAG, CoT, and MAD baselines while providing structured transparency and contestability unavailable in existing methods.

2. Proposed Method

As shown in Figure 1, ACAL consists of four components: adaptive expert selection, multi-agent argument generation, quantitative reasoning via A-QBAF, and human contestation with uncertainty-aware decision escalation.

2.1. Aspect Identification and Team Selection

Legal cases often involve multiple legal dimensions, so relying on a single expert can miss relevant perspectives. ACAL therefore dynamically assembles a task-specific team of legal agents. We define a legal agent pool $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$, where each agent a_i is characterized by its professional role, expertise, reasoning priorities, and argument style. Our framework instantiates 10 legal roles across four functional groups: *Adjudication*, *Litigation & Advocacy*, *Advisory & Transactional*, and *Research & Support*. Rather than using all agents for every case, we adaptively select task-specific subsets for support and attack.

Given a legal task $\mathcal{T}$ with context c and claim ϕ, the system selects:

$$\mathcal{A}^+ = \text{Select}(\mathcal{A}, \mathcal{T}, c, \text{support}); \quad \mathcal{A}^- = \text{Select}(\mathcal{A}, \mathcal{T}, c, \text{attack}), \tag{2.1}$$

where $\mathcal{A}^+$ contains agents suited to construct supporting arguments and $\mathcal{A}^-$ contains agents suited to generate counter-arguments. The selection function uses an LLM to match agent profiles to the characteristics of the case.

2.2. **Argument Generation, Evaluation and Computation**

Our reasoning module generates grounded legal arguments and computes their final strengths under A-QBAF, an extension of QBAF [17].

2.2.1. **Contextual Grounding via Hybrid RAG**

To reduce hallucination and anchor reasoning in legal authority, we retrieve evidence from both a vectorized legal corpus and external web search for recent case law. The top-k relevant passages are aggregated into the evidentiary context c used by all agents.

2.2.2. **Multi-Agent Argument Generation**

Each selected agent generates arguments addressing claim ϕ under context c:

$$\text{Args}(a) = \text{LLM}\left(a, \phi, c, \mathcal{C}, \text{type}(a)\right), \tag{2.2}$$

where $\text{type}(a) \in \{\text{support}, \text{attack}\}$ specifies whether the agent argues that the claim is true or false. Each agent typically produces 2–5 arguments depending on case complexity and evidence sufficiency.

2.2.3. **Intrinsic Strength Attribution**

Each argument α_i is assigned an intrinsic score $\tau(\alpha_i) \in [0, 1]$ by the LLM under a task-specific evaluation rubric. The score reflects the legal correctness, relevance to the case facts, specificity of reasoning, and overall logical soundness of the argument.

2.2.4. **LLM-based Inter-Argument Relation Identification**

To construct $\mathcal{R}^-$ and $\mathcal{R}^+$, we identify pairwise relations between arguments. Simple stance-based heuristics are often too coarse, as arguments from opposite sides may address different legal aspects and thus be logically independent. We therefore use an LLM to classify each pair (α_i, α_j), $i < j$, as *support*, *attack*, or *neutral*. Support and attack relations are instantiated as bidirectional edges, while neutral pairs introduce no edge.

2.2.5. **Clash Resolution via Arena Debating Round**

Since intrinsic scoring is performed independently for each argument, it may not fully capture the comparative strength of closely competing support and attack arguments. When a support argument and an attack argument have similar intrinsic scores (difference $< \delta$, default $\delta = 0.2$), score comparison alone may be unreliable. We therefore introduce a clash resolution (CR) step. For each conflicting pair (α_s, α_a), an LLM acting as a legal evaluator compares the two arguments under the case facts and legal standards and determines the stronger one. Scores are then adjusted according to win rate w:

$$\Delta\tau(\alpha) = \beta \cdot (2w - 1), \tag{2.3}$$

where β is the adjustment magnitude. This gives winners a bonus and losers a proportional penalty, preserving score differentiation in close conflicts.

2.2.6. **Arena-based Quantitative Bipolar Argumentation Framework (A-QBAF)**

We formalize the argument set as a QBAF $\langle \mathcal{X}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$, where $\mathcal{X} = \{\phi\} \cup \{\alpha_1, \ldots, \alpha_n\}$, $\mathcal{R}^-$ and $\mathcal{R}^+$ are attack and support relations, and τ is the base strength function. The claim node is initialized with $\tau(\phi) = 0.5$. Each support argument connects to ϕ through $\mathcal{R}^+$ and each attack argument through $\mathcal{R}^-$.

We compute final strengths using Quadratic Energy (QE) semantics [18], which propagates support and attack signals through a convergent continuous dynamical system. For each argument $j \in \mathcal{X}$, the net energy is:

$$E_j = \sum_{i \in \mathrm{Sup}_j} \sigma_i - \sum_{i \in \mathrm{Att}_j} \sigma_i, \tag{2.4}$$

where σ_i is the current strength of argument i. The impact of this energy is defined by:

$$h(x) = \frac{\max\{x, 0\}^2}{1 + \max\{x, 0\}^2}, \tag{2.5}$$

and the equilibrium strength is:

$$\sigma_j^* = \tau(j) + (1 - \tau(j)) \cdot h(E_j) - \tau(j) \cdot h(-E_j), \tag{2.6}$$

where $\sigma^* = \lim_{t \to \infty} \sigma(t)$. This formulation yields bounded, symmetric propagation of support and attack effects around the initial weight $\tau(j)$.

2.3. **Human-in-the-loop (HITL) Contestation**

To support contestability, we treat the A-QBAF graph as an editable decision object that allows users to challenge the system's reasoning. The interface presents the generated arguments together with their supporting evidence, roles, and scores, so that users can inspect how they influence the final claim score. During contestation, users may reject, revise, or add arguments, and suggest corrections to their strengths or support/attack relations. Accepted interventions update the argument graph and trigger recomputation of the propagated scores, allowing human feedback to materially affect the final decision rather than remain as post-hoc explanation. All interventions are logged for traceability, and high-uncertainty or high-impact cases may trigger additional review.

2.4. **Decision Consensus with Uncertainty-Aware Escalation (UAE)**

The final answer is Yes if $\sigma(\phi) \geq \theta$ and No otherwise, where $\theta = 0.5$ by default. However, near the decision boundary, small differences in base scores can cause support and attack arguments to offset each other, driving $\sigma(\phi)$ toward neutrality and making threshold-based decisions unstable. Prior work, therefore, recommends deferring or escalating highly uncertain cases [19]. Accordingly, when $\sigma(\phi) \in [0.49, 0.51]$, we invoke a *Final Judge* agent instead of relying on the threshold rule alone. This agent performs an independent legal analysis of the evidence, legal standards, and key conflicts, and outputs a binding decision. The mechanism specifically addresses near-tie saturation between support and attack arguments and ensures a decisive outcome under high uncertainty.

3. **Experiment and Results**

We evaluate on **LegalBench** [3] using two classification tasks: **Hearsay**, which predicts whether a statement is hearsay under FRE 801(c), and **Courts**, which predicts whether a narrative belongs to the `courts` category in the Learned Hands taxonomy. Experiments use Gemini-2.5-Flash-Lite and Gemini-2.5-Flash. For the RAG baseline, we retrieve the top-5 chunks using OpenAI `text-embedding-3-large`. In ACAL, the clash resolution parameter

Table 1. Comparative results on Learned Hands Courts and Hearsay from LegalBench. All metrics are reported as percentages. The highest scores are in **bold**.

Model	Method	Learned Hands Courts				Hearsay			
		Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Gemini-2.5-Flash-Lite	SP	57.8	63.1	57.8	53.1	69.2	73.6	71.5	68.9
	CoT	69.3	69.3	69.2	69.2	69.2	74.7	71.8	68.7
	RAG	70.3	70.5	70.3	70.3	71.3	70.8	70.9	70.9
	MAD	69.8	69.8	69.8	69.8	**74.5**	75.8	72.4	72.7
	ACAL (Ours)	**70.8**	**71.2**	**70.8**	**70.7**	**74.5**	**77.1**	**76.3**	**74.4**
Gemini-2.5-Flash	SP	65.1	72.5	65.1	62.0	75.5	75.7	74.2	74.5
	CoT	72.3	74.1	72.6	71.3	77.2	77.4	75.6	75.9
	RAG	75.0	**78.1**	75.0	74.3	75.5	**79.5**	72.8	73.0
	MAD	**75.5**	76.8	**75.5**	75.2	**77.6**	78.5	76.1	76.5
	ACAL (Ours)	**75.5**	76.4	**75.5**	**75.3**	76.7	77.1	**77.6**	**76.7**

is set to $\beta = 0.15$. We compare against **SP**, **CoT**, **RAG**, and **MAD**. Performance is measured with Accuracy, Precision, Recall, and Macro-F1.

Quantitative Analysis. Table 1 presents the comparative results of our proposed framework, ACAL, against four baselines (SP, CoT, RAG, and MAD) on the *Learned Hands Courts* and *Hearsay* tasks from LegalBench. We evaluate performance across two backbone models: Gemini-2.5-Flash-Lite and Gemini-2.5-Flash. As shown in the table, ACAL demonstrates superior or highly competitive performance across both model architectures.

Gemini-2.5-Flash-Lite: In the resource-constrained setting, ACAL achieves the best overall performance. On the *Learned Hands Courts* dataset, our method surpasses all baselines, achieving the highest Accuracy (70.8%) and F1-score (70.7%). Similarly, on the *Hearsay* dataset, ACAL outperforms the strongest baseline (MAD) in Precision (77.1%), Recall (76.3%), and F1-score (74.4%), while matching the highest Accuracy (74.5%).

Gemini-2.5-Flash: Scaling to the larger model, ACAL maintains its robustness. On *Learned Hands Courts*, ACAL matches the top accuracy of the MAD baseline (75.5%) and achieves a superior F1-score (75.3%) compared to RAG (74.3%). Notably, on the *Hearsay* dataset, ACAL achieves the highest Recall (77.6%) and F1-score (76.7%) among all compared methods. These results indicate ACAL consistently outperforms standard prompting methods and remains competitive with complex retrieval and debate-based baselines.

Explainability and Contestability Analysis. Beyond predictive performance, ACAL addresses the opacity of traditional legal AI. Unlike baselines such as CoT or MAD, which produce unstructured text or debate transcripts, ACAL generates an A-QBAF, a structured graph where decisions are mathematically derived from explicit arguments and intrinsic scores. Moreover, ACAL advances from passive explainability to active contestability. While RAG systems offer static citations, our HITL contestation workflow empowers users to directly audit and modify the reasoning graph. These interventions are mathematically propagated to update the final judgment, ensuring transparent, verifiable decision-making.

4. Conclusion

We propose ACAL, a neuro-symbolic framework that combines adaptive multi-agent collaboration with A-QBAF for accurate and contestable legal reasoning. On LegalBench, ACAL outperforms strong baselines, including CoT and RAG, across Gemini-2.5-Flash-Lite and Gemini-2.5-Flash. Beyond accuracy, it enables human-in-the-loop contestability by allowing users to inspect and revise the reasoning graph. Future work will improve efficiency and extend ACAL to broader legal and other high-stakes domains.

Acknowledgement

This work was supported by NSERC Discovery Grant No RGPIN-2025-04478 and NSERC Discovery Supplement Award No DGECR-2025-00129.

References

[1] E. Linna and T. Linna. "Challenges for generative AI in legal reasoning". In: *Discover Artificial Intelligence* (2026). ISSN: 2731-0809.

[2] H. T. Nguyen et al. "LLMs for legal reasoning: A unified framework and future perspectives". In: *Computer Law & Security Review* 58 (2025), p. 106165.

[3] N. Guha et al. "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 44123–44279.

[4] S. Yue et al. "LawLLM: Intelligent Legal System with Legal Reasoning and Verifiable Retrieval". In: *International Conference on Database Systems for Advanced Applications*. Springer. 2024, pp. 304–321.

[5] R. Yao et al. "Elevating Legal LLM Responses: Harnessing Trainable Logical Structures and Semantic Knowledge with Legal Reasoning". In: *arXiv preprint arXiv:2502.07912* (2025).

[6] C. G. O'Grady and C. O'Grady. "Agentic Workflows in the Practice of Law—AI Agents as Ethics Counsel". In: *Arizona Legal Studies Discussion Paper* (2024), pp. 25–03.

[7] S. S. Han et al. "COURTREASONER: Can LLM Agents Reason Like Judges?" In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025.

[8] G. Chen et al. "AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents". In: *Findings of the Association for Computational Linguistics: ACL 2025*. 2025.

[9] Z. He et al. "AgentsCourt: Building Judicial Decision-Making Agents with Court Debate Simulation and Legal Knowledge Augmentation". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. 2024, pp. 9399–9416.

[10] A. Hota and J. P. Jokinen. "NomicLaw: Emergent Trust and Strategic Argumentation in LLMs During Collaborative Law-Making". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8.2 (2025), pp. 1278–1289.

[11] S. Jung and J. Jung. "Courtroom-LLM: A Legal-Inspired Multi-LLM Framework for Resolving Ambiguous Text Classifications". In: *Proceedings of the 31st International Conference on Computational Linguistics*. 2025, pp. 7367–7385.

[12] Y.-C. Yu et al. "Structured Evaluation of Legal Reasoning in LLMs: Chain-of-Thought Prompting and Human Scoring for Retrieval Robustness". In: *NII Institutional Repository* (2025).

[13] A. Sadowski and J. A. Chudziak. "On Verifiable Legal Reasoning: A Multi-Agent Framework with Formalized Knowledge Representations". In: *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 2025, pp. 2535–2545.

[14] L. Zhang, M. Grabmair, M. Gray, and K. Ashley. "Thinking Longer, Not Always Smarter: Evaluating LLM Capabilities in Hierarchical Legal Reasoning". In: *arXiv preprint* (2025).

[15] H. Nguyen et al. "Heart2Mind: Human-Centered Contestable Psychiatric Disorder Prediction System Using Wearable ECG Monitors". In: *ACM Trans. Comput. Healthcare* (2026).

[16] S. Park et al. "Objection, your honor!: an LLM-driven approach for generating Korean criminal case counterarguments". In: *Artificial Intelligence and Law* (2025).

[17] X. Yin, N. Potyka, and F. Toni. "Argument attribution explanations in quantitative bipolar argumentation frameworks". In: (2023).

[18] N. Potyka. "Continuous Dynamical Systems for Weighted Bipolar Argumentation." In: *KR* 2018 (2018), pp. 148–57.

[19] M. M. Hasan et al. "Survey on leveraging uncertainty estimation towards trustworthy deep neural networks: The case of reject option and post-training processing". In: *ACM Computing Surveys* 57.9 (2025), pp. 1–35.

[20] G. Freedman, A. Dejl, D. Gorur, X. Yin, A. Rago, and F. Toni. "Argumentative Large Language Models for Explainable and Contestable Claim Verification". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 14. 2025, pp. 14930–14939.

Appendix A. Ablation Study

A.1. Clash Resolution (CR) and Uncertainty-Aware Escalation (UAE) Mechanism

To validate the effectiveness of our proposed modules, we conducted an ablation study on the *Hearsay* dataset using Gemini-2.5-Flash-Lite, effectively isolating the impact of the CR mechanism and the UAE strategy. This ablation study also serves as a comparative evaluation against vanilla ArgLLMs (without CR and UAE) [20], while adopting QE semantics for final argument strengths computation.

As shown in Table 2, the results identify CR as the primary driver of performance, yielding a substantial 7.9% increase in accuracy (64.9% → 72.8%) when applied independently, which confirms the necessity of resolving score saturation in LLM-generated arguments. Interestingly, deploying UAE in isolation negatively impacts performance (62.8%), suggesting that uncertainty estimation is unreliable without the calibration provided by CR. However, the full ACAL framework achieves the highest accuracy (74.5%) and F1-score (74.4%), demonstrating a complementary effect in which CR establishes the argument structure required for UAE to operate effectively in borderline cases.

A.2. Base Adjustment Magnitude β

We investigated the sensitivity of the hyperparameter β (base adjustment magnitude), which governs the intensity of score updates during Clash Resolution. As detailed in Table 3, performance on the *Learned Hands Courts* dataset exhibits a distinct bell-shaped trend, gradually improving as β increases from 0.05 and peaking at $\beta = 0.15$ with the highest Accuracy (70.8%) and F1-score (70.7%). However, increasing β beyond this threshold results in performance degradation. This finding suggests that while a moderate adjustment is essential to effectively differentiate between conflicting arguments, an excessively aggressive

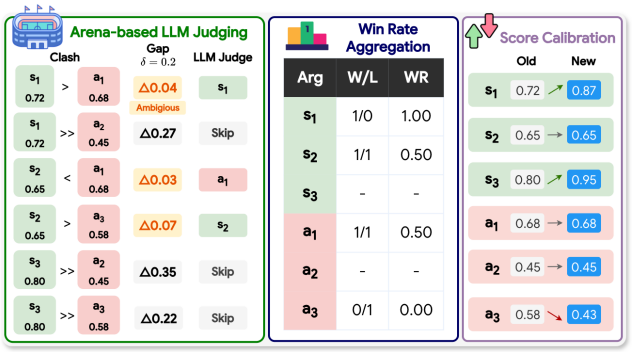

Figure 2. Illustration of an Arena Debating Round through Clash Resolution (CR).

Table 2. Ablation study of our two proposed modules, Clash Resolution (CR) and Uncertainty-Aware Escalation (UAE), on Gemini-2.5-Flash-Lite (Hearsay).

CR	UAE	Acc	Prec	Rec	F1
✗	✗	64.9	69.8	67.5	64.4
✗	✓	62.8	72.5	66.4	61.2
✓	✗	72.8	75.1	74.5	72.8
✓	✓	**74.5**	**77.1**	**76.3**	**74.4**

Table 3. Ablation study of parameter β (base adjustment magnitude) on Gemini-2.5-Flash-Lite (Learned Hands Courts).

β	Acc	Prec	Rec	F1
0.05	69.2	69.2	69.1	69.1
0.10	70.0	70.4	69.9	69.8
0.15	**70.8**	**71.2**	**70.8**	**70.7**
0.20	70.2	70.5	70.2	70.0
0.25	69.1	69.6	68.1	68.0

magnitude ($\beta \geq 0.20$) introduces volatility, leading the final propagated scores to become overly sensitive to individual clash outcomes rather than reflecting the holistic argument structure.

Appendix B. Case Study of ACAL in Legal Reasoning

This appendix presents an illustrative case study of the ACAL framework applied to a complex legal-reasoning scenario for a sample from the Hearsay task on LegalBench. Figure 3 demonstrates the end-to-end neuro-symbolic workflow, including: (1-2) Case Aspect Identification and Adaptive Expert Team recruitment; (3) Multi-Agent Argument Generation; (4) HITL contestation where a user explicitly modifies the reasoning graph; (5-6) A-QBAF Graph Construction and Clash Resolution for score calibration; and (7) Final Answer Generation with faithful explanation.

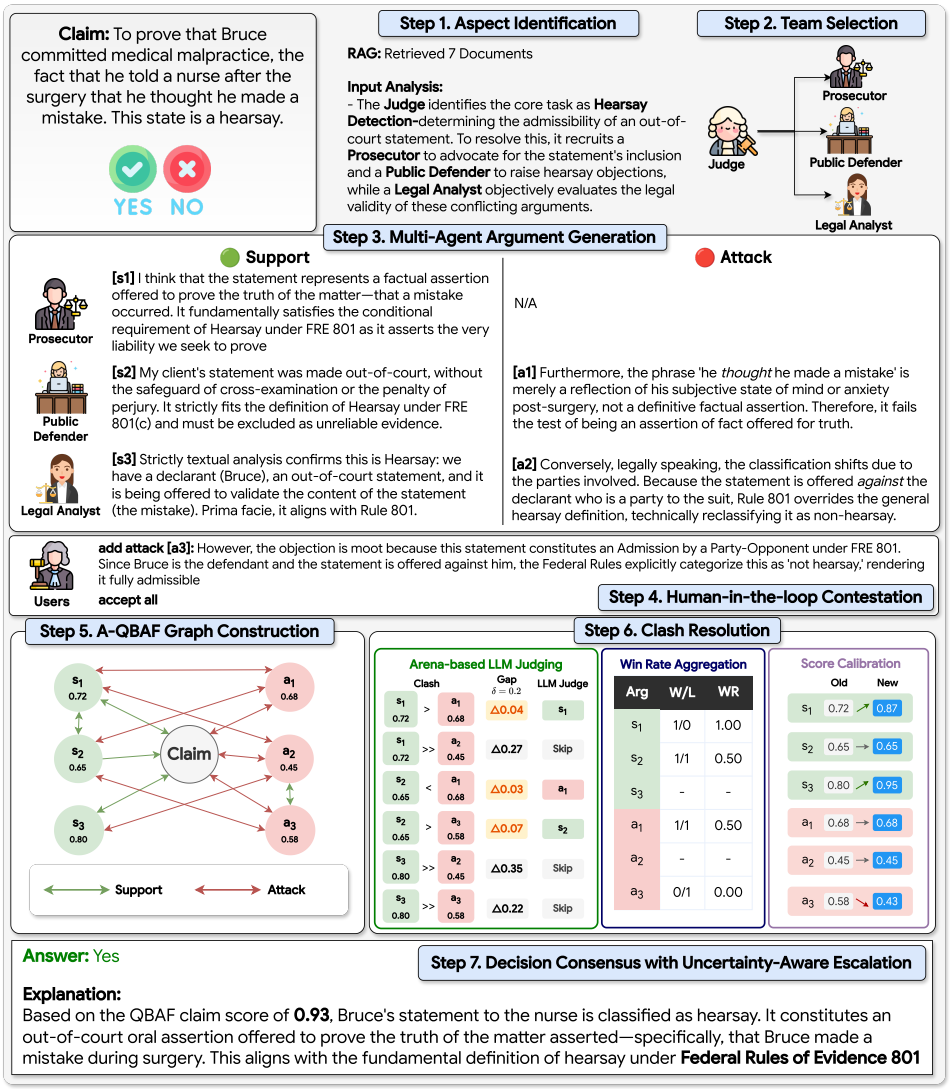

Figure 3. Illustrative Case Study of ACAL on the LegalBench Hearsay Task.