arXiv:2602.18916v1 [cs.MA] 21 Feb 2026

基于竞技场的论证型大语言模型自适应协作：实现可解释、可质疑的法律推理

Hoang-Loc Cao[†*§]、Phuc Ho[†§]、Truong Thanh Hung Nguyen[‡]、
Phuc Truong Loc Nguyen[◇]、Dinh Thien Loc Nguyen[†]、Hung Cao[‡]
[†]胡志明市自然科学大学，越南
[‡]新不伦瑞克大学，加拿大[◇]埃尔朗根-纽伦堡弗里德里希-亚历山大大学，德国

摘要

法律推理不仅需要高准确率，还需要能够通过可验证、可争议的论证来为决策提供依据。然而，现有的大语言模型（LLM）方法，例如思维链（CoT）和检索增强生成（RAG），往往会生成缺乏正式验证机制或用户干预途径的非结构化解释。为解决这一局限，我们提出论证型大语言模型的自适应协作（*ACAL*），这是一种神经符号框架，将自适应多智能体协作与基于竞技场的定量双极论证框架（A-QBAF）相结合。ACAL 会动态调度专家智能体团队来构建论证，采用冲突解决机制对相互冲突的主张进行裁决，并利用不确定性感知升级机制处理边界情形。尤为重要的是，我们的框架支持人在回路（HITL）可争议工作流，使用户能够直接审查和修改底层推理图，从而影响最终判决。在 LegalBench 基准测试上的实证评估表明，相较于 Gemini-2.5-Flash-Lite 和 Gemini-2.5-Flash 架构上的强基线方法，ACAL 均表现更优，有效兼顾了高效的预测性能、结构化透明度和可争议性。我们的实现见：
https://github.com/loc110504/ACAL。

关键词： 法律推理、大语言模型、多智能体系统、可争辩人工智能、计算论证、神经符号人工智能

1. 引言

法律推理是一个复杂过程，指将法律原则和规则应用于具体事实，从而得出有理有据的结论 [1, 2]。这项专业职责包含多项关键子任务，例如识别相关问题、回忆相关法律，以及解释复杂的制定法或判例 [2, 3]。该领域的从业者必须将既定规则应用于具体事实，得出合乎逻辑的结论，并提出有说服力的论证来支持其主张。近年来，大语言模型（LLM）在这些法律推理任务上取得了显著进展。研究人员发现，为这些模型提供领域知识和特定推理策略，能够大幅提升其性能 [3–5]。近期，研究人员也开始探索多智能体大语言模型框架，通过模拟辩论或司法评议来应对决策问题 [6, 7]。通过让多个智能体相互批评并完善彼此的论证，这些方法能够提升事实准确性和整体稳健性 [8, 9]。

然而，现有的大语言模型法律推理方法以不同方式未能满足这些要求。基于提示的方法（包括少样本提示和思维链（CoT）推理） [3, 4, 10], 往往会生成难以核验或系统性地提出异议的自由形式解释。检索增强生成（RAG）通过引入外部法律文本 [7, 10, 11], 提升事实依据，但最终的决策逻辑仍然隐晦且不透明。多智能体辩论框架（MAD）引入了多样化的观点，但其性能提升在不同任务上的表现并不一致，而且辩论记录本身并不构成可正式提出异议的推理结构。

此外，前沿模型在面对复杂推理需求时，往往会产生无效论证或不相关引文。这一局限引发了对信任、公平性以及

[§]这些作者贡献相同。[*]通讯作者：
chloc22@clc.fitus.edu.vn

高风险领域问责性的关切。与此同时，监管机构日益要求透明度与可争议性，这体现在欧盟《人工智能法案》等全球指令中 [12], 以及加拿大《自动化决策指令》 [13]。因此，既有方法要么在缺乏问责性的情况下优化预测性能，要么所提供的解释缺乏可供审查与争议的原则性机制。

为应对这些挑战，我们提出论证型大语言模型的自适应协作（*ACAL*），这是一种结构化的神经符号多智能体人工智能框架，旨在实现可解释且可争议的法律推理。我们的贡献总结如下：

(1) 我们提出 ACAL，这是一种自适应多智能体法律推理框架，它将论证型大语言模型（ArgLLMs） [14]与基于竞技场的（定量）双极论证（*A-(Q)BAF*）框架中的自适应角色选择相结合，并通过冲突消解（*CR*）和不确定性感知升级（*UAE*）加以增强，从而得出更准确、更明确的法律判决。(2) 我们提出的 ACAL 支持人在回路中的（HITL）争议处理工作流，使用户能够直接质询并修订底层推理过程，并将变更正式传递至最终结果。(3) 我们通过 LegalBench [15]上的实验证明，ACAL 在预测性能方面优于强提示、检索增强生成（RAG）、思维链（CoT）和多智能体辩论（MAD）等基线方法，同时还能提供以往方法不具备的结构化透明度和可争议性。

2. 相关工作

2.1. 基于大语言模型的法律推理方法

人工智能在法律领域的应用已从早期专家系统发展到利用大语言模型的现代数据驱动方法。ChatLaw[16] 和 LegalMind [17] 等专用模型通过整合知识图谱和强化学习，提升了事实准确性和流程优化能力。RAG 等基础技术如今已成为使大语言模型的回答以权威外部来源为依据的标准方法。系统利用 RAG 查询的内容涵盖伦理规则 [11] 、税法 [10], 以及复杂的成文法框架和判例法先例 [18]。与此同时，思维链提示被广泛用于引导多步推理；一些框架还开发了法律专用思维链变体 [4] 或逻辑语义集成模型 [5] ，以构建更稳健的演绎路径。为进一步提高稳健性，研究人员引入了多智能体架构，其中智能体承担不同角色，例如法官、原告和被告，以模拟对抗性法律论辩 [9]。这些系统从模拟法庭辩论 [19] 到协作立法 [20], 不等，并借助迭代式批评来完善论证、达成共识。

然而，尽管取得了这些进展，显著的性能差距依然存在。近期研究指出了一种“推理悖论”：在层级任务中，投入更多计算资源的模型往往性能反而下降，难以区分表层事实与更深层的法律区分 [21]。此外，尽管这些先进方法的准确率优于单模型基线 [6], ，但它们生成的辩论记录或自由形式的解释缺乏形式结构。因此，其推理过程难以得到系统验证或质疑，形成了一个关键的问责缺口，而纯粹由性能驱动的架构无法解决这一问题。

2.2. 基于大语言模型的法律推理系统的可解释性与可争议性

法律人工智能中的可解释性，根本上关乎为利益相关方提供合理的论证依据；这与描述模型

内部机制这一技术目标截然不同。当前框架试图通过提高推理过程的透明度来提供这种论证依据。例如，基于检索增强生成的系统通过提供引用机制，将主张与特定源文档相链接，从而实现可解释性 [7,22]。其他方法则利用思维链技术生成结构化、循序渐进的推理链，以模拟司法逻辑 [18, 21]。更先进的系统甚至会借助图尔敏模型等框架，将交互形式化为计算论证图，从而构建可验证的推理轨迹 [9, 23]。然而，存在这些解释性产物并不能保证其可靠性，也无法提供救济机制。近期一项对前沿模型的评估发现，生成的司法分析中超过60%包含无效论证，半数以上还引用了无关引文，尽管其结构连贯[18]。因此，人们越来越希望超越被动的可解释性，转向可争议性这一交互式范式，使用户能够主动质疑并纠正模型的推理过程，而非仅仅旁观 [11, 23]。

可争议人工智能（CAI）通过纳入保障措施，并为用户提供明确途径以质疑和修正系统结论，从而扩展了可解释人工智能的原则 [24, 25]。计算论证的整合已成为可争议性操作化的核心，尤其是在法律人工智能领域。诸如ALEX [23]等框架采用形式化方案，如图尔敏模型和ASPIC+ [23]框架，来组织刑事案件中对立论证之间的交互。这使得用户能够审视推理图中的特定节点。同样，SOLAR [8]等神经符号方法利用结构化本体表示，弥合自然语言与符号推理之间的差距，以进行成文法分析。这提供了能够经受对抗性探测的可验证路径。其他研究则通过面向基于案例的推理的结构化多轮辩证方案 [26] ，或从监管要求中推导政策指导以帮助用户确定应提出的恰当问题 [22]，探索了可争议性。然而，这些现有的法律可争议人工智能方法主要侧重于生成用于事后检查的静态推理结构。它们通常缺乏一种定量机制，用以动态权衡冲突论点的强度。此外，它们不支持完全交互式的工作流程，即人工干预（如修改论证权重或关系）能够以数学方式传播，从而更新最终法律判决。

我们的工作通过将多智能体辩论整合到一个可由人在回路中直接审计的形式化定量论证框架中，弥补了这一差距。这使静态解释转变为动态且可质疑的决策对象。

3. 论证型大语言模型的自适应协作（ACAL）

如图 1 所示，ACAL 包含四个模块：自适应专家团队选择、结合基于竞技场的冲突消解的多智能体论证生成、通过 A-QBAF 进行定量推理，以及结合不确定性感知决策共识的人在回路中的争辩。

3.1. 方面识别与团队选择

法律推理面临着来自不同领域的诸多挑战，需要专业知识。面对需要不同专业人士有效分析的复杂案件，单个专家可能无法涵盖所有观点。为克服这一局限，我们提出一种自适应多智能体架构，可根据特定法律任务动态组建量身定制的专家团队。

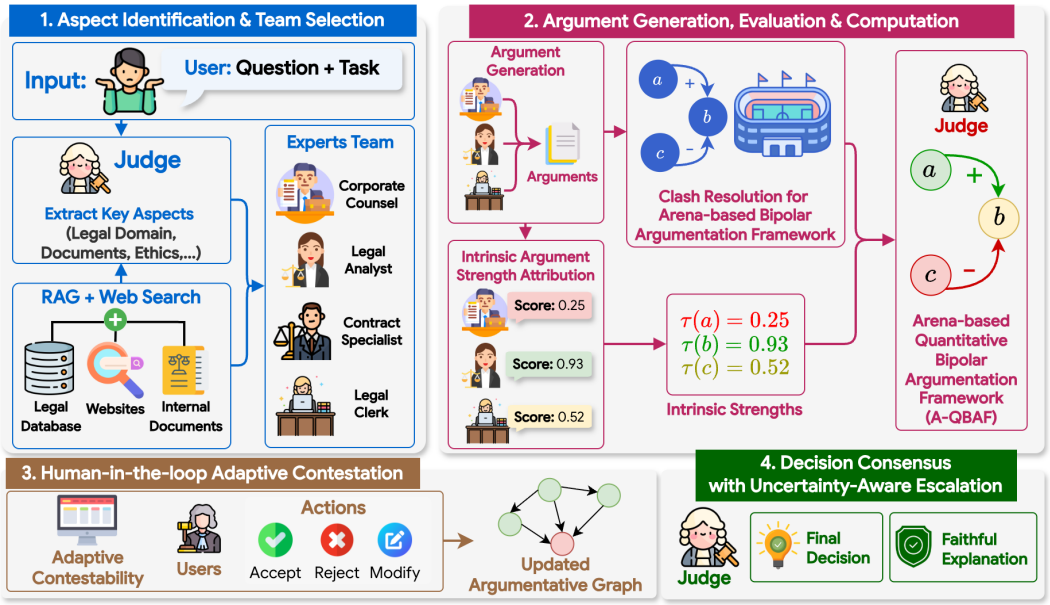

图 *1.* 我们提出的用于法律推理的论证型大语言模型的自适应协作（ACAL）架构。

3.1.1. 法律智能体池定义

我们定义了一个全面的法律智能体集合 $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$，其中每个智能体$a_i$均由一个元组 r_i、 E_i、 P_i、 S_i表示，具体如下：(i) 角色 r_i：专业职称；(ii) 专业领域 E_i：一组领域专长；(iii)重点优先事项 P_i：指导推理的任务导向目标；(iv) 论证风格 S_i：具有代表性的推理方式。如表 1 所示，我们的框架实现了 10 种不同的法律角色，并将其归入不同的职能类别 [27]。

表 *1.* 按功能类别分组的法律角色。

类别	角色
裁判	法官、法律助理 / 司法助理
诉讼与辩护	私人执业律师、检察官、公设辩护人
咨询与交易	公司法律顾问、合规官、知识产权律师
研究与支持	法律分析师、律师助理

3.1.2. 自适应智能体选择

我们并不针对每个案件都使用所有智能体，因为这不仅计算成本高，而且在特定任务中使用非专家时可能引入噪声，因此我们采用了自适应选择机制。对于给定的法律任务 $\mathcal{T}$ 及其上下文 c 和主张ϕ，系统会选择两个子集：

$$\mathcal{A}^+ = \text{Select}(\mathcal{A}, \mathcal{T}, c, \text{support}); \quad \mathcal{A}^- = \text{Select}(\mathcal{A}, \mathcal{T}, c, \text{attack}), \tag{3.1}$$

其中， $\mathcal{A}^+$ 是专业知识与为特定案件情境构建支持性论据相匹配的智能体， $\mathcal{A}^-$ 是适合生成反驳论据的智能体。选择函数利用大语言模型，将智能体专业知识档案与案件特征进行匹配。

3.2. 论证生成、评估与计算

我们框架的核心在于利用 A-QBAF 对论证强度进行结构化生成和计算，而 A-QBAF 是基于传统 QBAF 构建的 [28]。

3.2.1. 基于混合检索增强生成的上下文依据构建

为确保论证以权威法律标准而非泛化知识为依据，我们集成了混合检索模块。在生成论证之前，系统会同时查询用于检索立法文本的向量化法律数据库，以及用于检索最新判例法的外部网络搜索。系统将语义相关度最高的前k 段内容汇总，形成证据上下文 c，以此为所有智能体后续推理提供条件，从而尽量减少幻觉并确保准确率。

3.2.2. 多智能体论证生成

每个选定的智能体都会在上下文 c中直接针对中心主张 ϕ生成论证。对于智能体 $a \in \mathcal{A}^+ \cup \mathcal{A}^-$:

$$\mathrm{Args}(a) = \mathrm{LLM}\left(a, \phi, c, \mathcal{C}, \mathrm{type}(a)\right), \tag{3.2}$$

其中，type(a) $\in \{support, attack\}$ 决定该智能体构建的证据是支持主张为 True($support$) 还是 False($attack$)。每个智能体通常会生成 2-5 条论证，具体数量根据可用证据和案件复杂程度自行确定。

3.2.3. 内在强度赋值

每个论证 α_i 都会从大语言模型获得一个内在强度分数 $\tau(\alpha_i)$。我们设计了一套评分标准（见表 2），通过严格区分分数来防止分数饱和，要求评估者对笼统陈述予以扣分，并对案件特定的法律精确性予以加分。

表 2. 基于大语言模型的论证强度评分细则。

分数范围	解释
$0.1 - 0.2$	法律分析错误或关键要素识别错误
$0.3 - 0.4$	部分正确，但遗漏了关键要素，或表述过于笼统
$0.5 - 0.6$	分析合理，但存在细微疏漏，或未充分结合具体案例进行分析
$0.7 - 0.8$	分析扎实，结合了具体事实，法律推理正确
$0.9 - 1.0$	精准度极高，引文权威，逻辑无懈可击

3.2.4. 基于大语言模型的论证间关系识别

在构建 A-QBAF 图之前，我们必须确定论证之间的成对关系 $\mathcal{R}^-$和 $\mathcal{R}^+$。在启发式模式下，系统遵循一条简单规则：持相同立场的论证（均表示支持或均表示反对）被视为相互支持，而持相反立场的论证则被视为相互攻击。然而，这种做法过度简化了法律推理，因为立场相反的两个论证可能涉及完全不同的法律层面，因此在逻辑上彼此独立，而非相互冲突。

为构建更完善的逻辑结构，我们实现了基于大语言模型的语义关系识别模式。对于每一对 (α_i, α_j)，若满足 $i < j$，我们会提示大语言模型将关系分为反对、支持或中性三类，并给出置信度分数。为确保稳健性，置信度分数为 < 0.6 的任何支持或反对关系都会降级为中性。所有已建立的关系均设为双向，以确保在后续 QBAF 传播中进行对称

处理。中性论据对不引入边，从而保持图的稀疏性和可解释性。为提高效率，每次大语言模型调用会批量分析 b 对论据对（默认值为 $b = 10$），将 API 调用次数从 $\binom{n}{2}$ 降至 $\lceil \frac{\binom{n}{2}}{b} \rceil$

3.2.5. 通过竞技场辩论轮次进行冲突消解

当对立论证的基础分数相近（差值 $< \delta$，默认值为 $\delta = 0.2$）时，我们提出一种冲突消解（CR）机制来裁定二者，如图 2 所示。对于每一对冲突论据（α_s，α_a），其中 α_s 支持该主张，而 α_a 反对该主张：

（1）将两项论证提交给一个充当法律推理专家的大语言模型。（2）结合案件特定事实和法律标准，评估哪项论证更有力。（3）根据所有冲突中的胜负结果调整分数。

调整采用基于胜率 w的对称公式：

$$\Delta\tau(\alpha) = \beta \cdot (2w - 1), \tag{3.3}$$

其中，β 是基础调整幅度，$w \in [0, 1]$ 是赢得冲突的比例。这样可以确保获胜者获得奖励分，落败者受到按比例计算的惩罚分，从而保持分数差异。

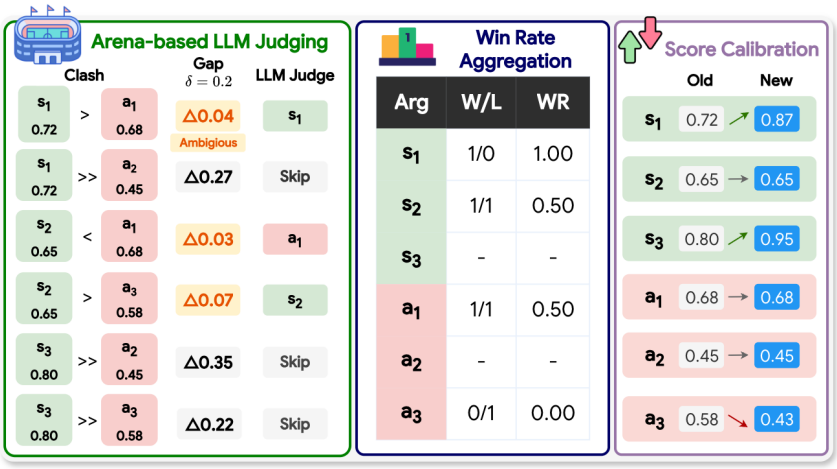

图 2. 通过冲突消解（CR）进行竞技场辩论轮次的示意图。

3.2.6. 基于竞技场的定量双极论证框架（A-QBAF）

我们将论证结构形式化为 QBAF $\langle \mathcal{X}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ 其中：

- $\mathcal{X} = \{\phi\} \cup \{\alpha_1, \ldots, \alpha_n\}$：论证集合，包括中心主张 ϕ，
- $\mathcal{R}^- \subseteq \mathcal{X} \times \mathcal{X}$：攻击关系，
- $\mathcal{R}^+ \subseteq \mathcal{X} \times \mathcal{X}$：支持关系，
- τ： $\mathcal{X} \to [0, 1]$：基础强度函数。

图构造：中心主张节点 ϕ 以基础分数 $\tau(\phi)$ =0.5（中性）初始化。每个论证 α_i 通过以下方式连接到 ϕ：

- $(\alpha_i, \phi) \in \mathcal{R}^+$ 若 α_i 为支持论证，
- $(\alpha_i, \phi) \in \mathcal{R}^-$ 若 α_i 为攻击论证。

我们使用二次能量（*QE*）语义 [29],计算最终论证强度，该语义将分数传播建模为连续动力系统，以确保收敛性和公理稳定性，具体如下：

能量计算：对于论证 $j \in \mathcal{X}$，总能量 E_j 是其支持者当前强度之和减去其攻击者当前强度之和

$$E_j = \sum_{i \in \mathrm{Sup}_j} \sigma_i - \sum_{i \in \mathrm{Att}_j} \sigma_i, \tag{3.4}$$

其中，σ_i 表示论证 i在时间 t的强度。

影响函数：能量的影响由二次函数 $h(x)$决定，该函数连续可微，且在正值范围内严格单调：

$$h(x) = \frac{\max\{x, 0\}^2}{1 + \max\{x, 0\}^2}, \tag{3.5}$$

平衡强度：最终传播强度 σ_j^* （后文也记作 $\sigma(j)$）即平衡状态 $\sigma^* = \lim_{t\to\infty} \sigma(t)$。在平衡状态下，该强度满足：

$$\sigma_j^* = \tau(j) + (1 - \tau(j)) \cdot h(E_j) - \tau(j) \cdot h(-E_j), \tag{3.6}$$

其中，$\tau(j)$是论证的初始基础权重。这确保了反对和支持对论证强度产生对称（对偶）的影响。

3.3. 人在回路中的（HITL）争议

结构化论证虽然提高了透明度，但法律决策还需要具备可争议性：受影响的当事人和法律专业人士必须能够质疑系统的推理、纠正错误，并切实影响结果。因此，我们引入了基于角色的HITL争议模块，将我们的 A-QBAF 图从静态解释转变为可编辑、可修订的决策对象。

3.3.1. 争议工件

根据论证结构 $\langle \mathcal{X}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ 以及计算得出的度数 $\sigma(\cdot)$，系统会生成参与摘要和论证仪表板。每张论证卡片都会显示：(i) 论证文本和立场（支持/反对）；(ii) 发起代理角色 r_i；(iii) 指向所用具体证据段落（制定法、判例、案件事实）的链接；(iv) 内在得分 $\tau(\alpha)$和传播得分 $\sigma(\alpha)$；以及 (v) 该论证在 A-QBAF 中的局部邻域（传入的支持者/攻击者），并附上一段简短的“重要性说明”轨迹，描述其对 $\sigma(\phi)$ 的影响。

3.3.2. 统一争议工作流

为在单一的面向用户的工作流中支持可争议性，界面提供了一个引导式质疑流程，围绕系统推理中受到质疑的内容展开：*(a)* 事实问题（事实或证据不正确或缺失）、*(b)* 法律规则问题（错误适用法规/标准或采用错误的检验标准）、*(c)* 判例问题（依据无关、已被推翻或遭到误读）、*(d)* 缺失的例外/抗辩事由（例如，同意、紧急避险、诉讼时效期间），以及*(e)* 程序公平方面的顾虑（缺少通知、答辩机会或减轻责任）。选择争议类型后，系统会触发有针对性的提示，(1) 请求用户提供更多支持材料，或 (2) 指示相关智能体根据用户提出的质疑主张重新生成、完善或反驳论证。所有由此产生的更新都会反映在可编辑的论证记录（论证、关系和得分）中，并附有可审计变更日志，从而确保质疑能够实质性地影响传播的主张得分，而非仅仅停留在事后反馈层面。

3.3.3. 允许的人工干预

质疑过程中，用户可以通过以下方式对论证记录提供反馈：(i) 接受或驳回论证；(ii) 编辑论证以更正事实或完善法律理据；以及 (iii) 补充缺失的论证（如有引文，则一并提供）。如果论证结构明显设定不当，用户还可以建议调整论证的强度或其支持/攻击关系。

更新与重新计算。应用所有获准的编辑，生成更新后的 A-QBAF（$\mathcal{X}^H$，$\mathcal{R}^{+H}$，$\mathcal{R}^{-H}$，τ^H），随后重新运行相同的传播语义，得到更新后的得分 $\sigma^H(\cdot)$ 和修订后的主张得分 $\sigma^H(\phi)$。因此，质疑被视为一种模型输入，可以改变最终决策，而非事后标注。

监督与日志记录。对于高不确定性或高影响案件，系统可以触发额外的审核步骤；所有质疑操作都会记录在审计日志中（谁更改了什么，以及这如何影响 $\sigma(\phi)$），以支持可追溯性。

3.4. 不确定性感知升级下的决策共识

最终答案根据主张得分得出：

$$\text{Answer} = \begin{cases} \text{Yes}, & \text{if } \sigma(\phi) \geq \theta, \\ \text{No}, & \text{otherwise}, \end{cases} \tag{3.7}$$

其中，θ 是决策阈值（默认值为 $\theta = 0.5$）。然而在实践中，初始的基于大语言模型的基础评分可能会使支持与攻击论证的得分十分接近（例如，二者看起来同样可信），导致传播后的主张得分趋近于中性区间。因此，依据阈值作出的决策若接近边界，可能会变得武断，结果受边际评分噪声而非有意义的法律区分左右。因此，既有研究 [30] 建议在高不确定性情形下推迟预测或弃权，通常的做法是升级至更强大的模型或交由人工审核者处理。

因此，我们针对 $\sigma(\phi) \in [0.49, 0.51]$ ($\sigma(\phi) \approx 0.5$) 的边界情形，提出不确定性感知升级（*UAE*）方法：该方法绕过受基础分数限制的决策规则，并启用一名最终法官智能体。法官会对案件进行独立的法律分析（重新评估证据和法律标准，并解决关键冲突），然后给出具有约束力的裁决。该机制专门解决反对论证与支持论证之间的同分饱和问题，并确保在高度不确定的情况下得出明确结论。

4. 实验设置

4.1. 数据集

我们在**LegalBench** [15],这一协作构建的基准测试上评估我们的方法。该基准测试包含多种多样的法律推理任务，旨在用于大语言模型的上下文评估。在本研究中，我们聚焦于两项分类任务：

- **传闻证据**（hearsay）：给定一段简短陈述，预测其是否属于《联邦证据规则》第801(c)条规定的传闻证据（传闻证据与非_传闻证据）。
- **法院**（learned_hands_courts）：给定一段真实世界叙述（例如论坛帖子），预测其是否属于 Learned Hands 分类体系中的法院类别（是 / 否）。

4.2. 基准测试

所有实验均使用 Google 的两种模型：Gemini-2.5-Flash-Lite 和 Gemini- 2.5-Flash。对于检索增强基线，我们实现了一个标准的检索增强生成流程，其中

文档使用 OpenAI text-embedding-3-large 进行嵌入，并提取最相关的 top-k ($k = 5$) 个文本块，在进行预测前为生成器提供额外的上下文。对于我们提出的框架，我们为冲突解决机制设置了基础调整幅度 $\beta = 0.15$，该数值经实证确定，以实现最佳性能。

我们将其与法律推理和一般推理中常用的提示方法及多智能体基线方法进行比较：**SP** [31]（少样本标准提示）使用固定的指令模板，并提供少量上下文示例；**思维链** [32] 鼓励模型在生成最终标签之前进行中间推理；**检索增强生成** [33] 通过检索法律证据（例如相关的法条或法理片段）补充输入，并基于这些外部上下文进行预测；而**多智能体辩论** [34] 则采用多智能体辩论设置，由三个智能体经过两轮批评与修订，最终得出决策。

对于这两项任务，我们报告准确率（Acc）、精确率（Prec）、召回率（Rec）和宏平均F1（F1）。宏平均F1通过对各类别F1分数取平均计算得出，确保各类别权重相等，从而使评估对类别不平衡更具稳健性。

5. 结果

表3. LegalBench 中 Learned Hands Courts 和 Hearsay 数据集上的对比结果。所有指标均以百分比表示。最高分以**粗体**标示。

模型	方法	Learned Hands Courts				传闻证据			
		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（本文方法）	**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（本文方法）	**75.5**	76.4	**75.5**	**75.3**	76.7	77.1	**77.6**	**76.7**

5.1. 定量分析

表3展示了我们提出的框架 ACAL 与四种基线方法（SP、思维链、检索增强生成和多智能体辩论）在 LegalBench 的*Learned Hands Courts* 和传闻证据任务上的比较结果。我们使用两种骨干模型评估性能：Gemini-2.5-Flash-Lite 和 Gemini-2.5-Flash。如表所示，ACAL 在两种模型架构下均表现出更优或极具竞争力的性能。

Gemini-2.5-Flash-Lite: 在资源受限环境中，ACAL 整体表现最佳。在*Learned Hands Courts* 数据集上，我们的方法优于所有基线方法，取得最高准确率（70.8%）和 F1分数（70.7%）。同样，在传闻证据 数据集上，ACAL 在精确率（77.1%）、召回率（76.3%）和 F1分数（74.4%）方面均优于最强基线方法（多智能体辩论），同时准确率（74.5%）达到最高水平。

Gemini-2.5-Flash： 扩展至更大模型时，ACAL仍保持其稳健性。在*Learned Hands Courts*上，ACAL的准确率与多智能体辩论（MAD）的最高准确率持平

表4. 对我们提出的两个模块——冲突消解（CR）和不确定性感知升级（UAE）进行消融研究，测试模型为 Gemini-2.5-Flash-Lite（传闻证据）。

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**

表5. 对参数 β （基础调整幅度）进行消融研究，测试模型为 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

基线方法（75.5%），并取得了优于检索增强生成（74.3%）的F1分数（75.3%）。值得注意的是，在传闻证据数据集上，ACAL在所有对比方法中取得了最高召回率（77.6%）和F1分数（76.7%）。这些结果表明，ACAL持续优于标准提示方法，并与复杂的检索与辩论式基线方法保持竞争力。

5.2. 可解释性与可争辩性分析

除了预测性能之外，ACAL 还解决了传统法律人工智能的不透明问题。与思维链或多智能体辩论等基线方法不同，后者生成的是非结构化文本或辩论记录，而 ACAL 会生成 A-QBAF，即一种结构化图，其中的决策由显式论证和内在分数通过数学方式推导得出。此外，ACAL 将被动可解释性提升为主动可争辩性。检索增强生成系统提供的是静态引文，而我们的人在回路争辩工作流则赋予用户直接审查和修改推理图的能力。这些干预会通过数学方式逐层传递，从而更新最终判决，确保决策过程透明且可验证。

5.3. 消融研究

冲突消解 (CR) 和不确定性感知升级 (UAE) 机制。为验证所提出模块的有效性，我们使用 Gemini-2.5-Flash-Lite 在传闻证据数据集上进行了消融研究，从而有效分离 CR 机制和 UAE 策略的影响。这项消融研究还与原始论证大语言模型（不使用 CR 和 UAE）进行了比较评估 [14], ，同时采用 QE 语义计算最终的论证强度。

如表4所示，结果表明 CR 是推动性能提升的主要因素，单独应用即可使准确率提高 7.9%（64.9% $\rightarrow$ 72.8%），这证实了消解大语言模型生成的论证中的分数饱和现象的必要性。值得注意的是，单独使用 UAE 会降低性能（62.8%），这表明若没有 CR 提供的校准，不确定性估计并不可靠。然而，完整的 ACAL框架取得了最高的准确率（74.5%）和 F1分数（74.4%），体现了二者的互补作用：CR 建立了论证结构，使 UAE 能够在边界情形中有效运行。

基础调整幅度 β。我们研究了超参数β （基础调整幅度）的敏感性，该参数决定冲突消解过程中分数更新的强度。如表5所示，模型在$Learned\ Hands\ Courts$ 数据集上的表现呈现明显的钟形趋势：随着 β 从0.05逐渐增大，性能逐步提升，并在 $\beta = 0.15$ 时达到峰值，准确率（70.8%）和F1分数（70.7%）均为最高。然而，当 β 超过这一阈值后，性能便会下降。这一发现表明，适度调整对于有效区分冲突论点至关重要，但幅度过大（$\beta \geq 0.20$）会导致波动，使

最终传播得分过度受单次冲突结果影响，而无法体现整体论证结构。

6. 结论

本文提出了 ACAL，这是一种神经符号框架，旨在满足自动化法律推理对高性能与可争议性的关键需求。通过将自适应多智能体协作与 A-QBAF 相结合，ACAL 成功地将非结构化的大语言模型输出转化为形式化、可验证的推理图。我们在 LegalBench 基准测试上进行的实证评估表明，ACAL 表现出具有竞争力的性能，在 Gemini-2.5-Flash-Lite 和 Gemini-2.5-Flash 模型上均优于包括思维链和检索增强生成在内的强基线方法。尤为重要的是，ACAL 不仅在标准性能指标上表现出色，还通过支持人在回路可争议工作流弥合了问责缺口。该工作流使利益相关方能够直接审查并介入推理过程，确保法律判决不仅准确，而且有据可依，并符合针对高风险人工智能日益完善的监管标准。未来的工作将着重优化多智能体架构的计算效率，在不牺牲推理深度的前提下降低推理成本。此外，我们还计划将自适应智能体池扩展到更广泛的复杂法律任务及其他需要可争议决策的高风险领域，以验证 ACAL 框架的泛化能力。

参考文献

[1] E. Linna 和 T. Linna。《生成式人工智能在法律推理中面临的挑战》。载于：*Discover Artificial Intelligence*（2026）。ISSN：2731-0809。[2] N. Guha 等。《LegalBench：协作构建的基准测试，用于衡量大语言模型的法律推理能力》。载于：神经信息处理系统进展。第36卷。2023年，第44123–44279页。[3] Y.-C. Yu 等。《大语言模型法律推理的结构化评估：用于检索鲁棒性的思维链提示与人工评分》。载于：*NII*机构知识库（2025）。[4] S. Yue 等。《LawLLM：具备法律推理与可验证检索能力的智能法律系统》。载于：高级应用数据库系统国际会议。施普林格。2024年，第304–321页。[5] R. Yao 等。《提升法律大语言模型的回答质量：利用可训练逻辑结构与语义知识进行法律推理》。载于：*arXiv*预印本 *arXiv:2502.07912*（2025）。[6] G. Chen 等。《AgentCourt：使用对抗性可进化律师智能体模拟法庭》。载于：计算语言学协会年会论文集：*ACL 2025*。2025年。[7] P. Devadiga、O. J. Shetty 和 P. Agarwal。《SAMVAD：用于模拟印度司法审议动态的多智能体系统》。载于：*arXiv*预印本 *arXiv:2509.03793*（2025）。[8] A. Sadowski 和 J. A. Chudziak。《论可验证法律推理：采用形式化知识表示的多智能体框架》。载于：第*34*届*ACM*信息与知识管理国际会议论文集。2025年，第2535–2545页。[9] S. Jung 和 J. Jung。《Courtroom-LLM：受法律启发的多大语言模型框架，用于解决歧义文本分类问题》。载于：第*31*届国际计算语言学会议论文集。2025年，第7367–7385页。[10] J. J. Nay 等。《大语言模型作为税务律师：法律能力涌现案例研究》。载于：英国皇家学会哲学汇刊*A*辑：数学、物理与工程科学 382.2270（2024年2月），第20230159页。ISSN：1364-503X。[11] C. G. O'Grady 和 C. O'Grady。《法律实践中的智能体工作流——作为伦理顾问的人工智能智能体》。载于：亚利桑那州法律研究讨论论文（2024），第25–03页。[12] 欧盟。《欧盟〈人工智能法案〉》。载于：欧盟《人工智能法案》（2024）。[13] 加拿大政府。自动化决策指令-*Canada.ca*。2019年。

[14] G. Freedman、A. Dejl、D. Gorur、X. Yin、A. Rago 和 F. Toni。《用于可解释且可争辩的主张核验的论证型大语言模型》。载于：人工智能促进协会人工智能会议论文集。第39卷。第14期。2025年，第14930–14939页。[15] N. Guha 等。《LegalBench：协作构建的基准测试，用于衡量大语言模型的法律推理能力》。载于：神经信息处理系统进展36（2023）。[16] J. Cui 等。《Chatlaw：由知识图谱增强的混合专家大语言模型驱动的多智能体协作法律助手》。载于：arXiv:2306.16092（2023）。[17] N. Raju 等。《LegalMind：利用 DeepSeek 优化流程并降低法律服务成本的智能体人工智能》。载于：IEEE Access（2025）。[18] S. S. Han 等。《COURTREASONER：大语言模型智能体能像法官一样推理吗？》。载于：2025年自然语言处理实证方法会议论文集。2025年。[19] Z. He 等。《AgentsCourt：借助法庭辩论模拟与法律知识增强构建司法决策智能体》。载于：计算语言学协会年会论文集：EMNLP 2024。2024年，第9399–9416页。[20] A. Hota 和 J. P. Jokinen。《NomicLaw：协作立法过程中大语言模型中涌现的信任与策略性论证》。载于：AAAI/ACM人工智能、伦理与社会会议论文集8.2（2025），第1278–1289页。[21] L. Zhang、M. Grabmair、M. Gray 和 K. Ashley。《想得更久，却不总是更聪明：评估大语言模型的层级法律推理能力》。载于：arXiv预印本（2025）。[22] I. Cheong 等。《我不是律师，但……：邀请法律专家参与制定负责任的法律建议大语言模型政策》。载于：2024年ACM公平、问责与透明度会议论文集。FAccT '24。美国纽约州纽约市，2024年。[23] S. Park 等。《异议，法官阁下！：一种由大语言模型驱动的韩国刑事案件反驳论点生成方法》。载于：人工智能与法律（2025）。[24] H. Nguyen、A. Rahimi、V. Whitford、H. Fournier、I. Kondratova、R. Richard 和 H. Cao。《Heart2Mind：利用可穿戴心电监测仪实现以人为本且可争辩的精神障碍预测系统》。载于：ACM医疗保健计算汇刊（2026）。[25] L. P. T. Nguyen 等。《Motion2Meaning：面向帕金森病步态解读的以临床医生为中心的可争辩大语言模型框架》。载于：第9届聊天机器人与以人为本人工智能国际研讨会论文集（CONVERSATIONS）2025。2025年。[26] M. Gray、L. Zhang 和 K. D. Ashley。《使用大语言模型生成基于案例的法律论证》。载于：2025年计算机科学与法律研讨会论文集。CSLAW '25。德国慕尼黑：计算机协会，2025年，第160–168页。ISBN：9798400714214。[27] S. Liu 等。《法律领域的大语言模型智能体：分类体系、应用与挑战》。载于：arXiv预印本 arXiv:2601.06216（2026）。[28] X. Yin、N. Potyka 和 F. Toni。《定量双极论证框架中的论证归因解释》。载于：（2023）。[29] N. Potyka。《加权双极论证的连续动力系统》。载于：KR 2018（2018），第148–57页。[30] M. M. Hasan 等。《关于利用不确定性估计构建可信深度神经网络的综述：拒绝选项与训练后处理案例》。载于：ACM计算调查57.9（2025），第1–35页。[31] T. Brown 等。《语言模型是少样本学习者》。载于：神经信息处理系统进展33（2020），第1877–1901页。[32] J. Wei 等。《思维链提示能够激发大语言模型的推理能力》。载于：神经信息处理系统进展35（2022），第24824–24837页。[33] P. Lewis 等。《面向知识密集型自然语言处理任务的检索增强生成》。载于：神经信息处理系统进展33（2020），第9459–9474页。[34] Y. Du、S. Li、A. Torralba、J. B. Tenenbaum 和 I. Mordatch。《通过多智能体辩论提升语言模型的事实准确性与推理能力》。载于：第41届国际机器学习会议论文集。ICML'24。JMLR.org，2024年。

附录A. ACAL在法律推理中的案例研究

本附录展示了一项说明性案例研究，将ACAL框架应用于LegalBench传闻证据任务中某一样本的复杂法律推理场景。图3展示了端到端的神经符号工作流，包括：（1-2）案例方面识别与自适应专家团队招募；（3）多智能体论证生成；（4）人在回路中的争辩，用户在其中显式修改推理图；（5-6）用于分数校准的A-QBAF图构建与冲突消解；以及（7）带有忠实解释的最终答案生成。

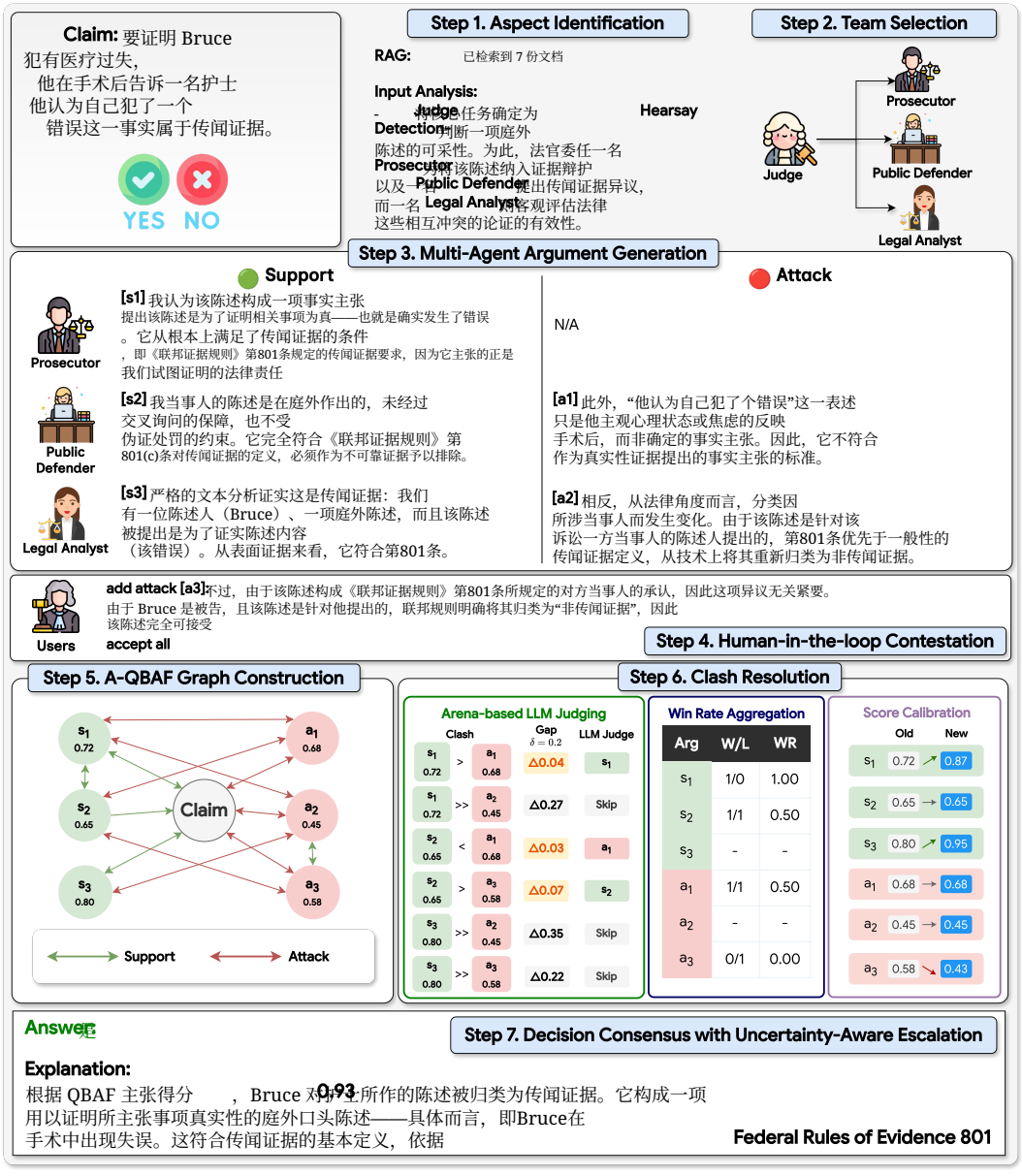

图 *3.* ACAL 在 LegalBench 传闻证据任务中的示例案例研究。