System L: Toward System 2-Style Legal Reasoning

Chuanyi Li[1], **Yi Feng**[1*], **Vincent Ng**[2]

[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Human Language Technology Research Institute, University of Texas at Dallas, USA
lcy@nju.edu.cn, fy@nju.edu.cn, vince@hlt.utdallas.edu

Abstract

Dual-system theory distinguishes between fast, intuitive System 1 and slow, deliberative System 2. While this dichotomy describes many forms of reasoning, it oversimplifies the reality of expert legal reasoning. Legal reasoning is not merely a process of slow, logical deliberation. It is intrinsically normative, embedding precedent analysis, statutory interpretation, policy balancing, and social values. This paper envisions a reasoning architecture for legal reasoning, System L (Legal System 2), which extends traditional System 2 by integrating domain-specific normative frameworks in a structured manner. Using the IRAC (Issue–Rule–Application–Conclusion) structure as a backbone model, System L represents a blueprint for the next generation of cognitive and AI systems capable of human-like legal reasoning.

1 Introduction

Reasoning has long been regarded as a central component of Artificial Intelligence (AI), enabling systems to move beyond surface-level pattern recognition toward genuine intelligence (Gupta 2025; Behnke 2024; Liu et al. 2025; Cao et al. 2024). In AI, reasoning broadly refers to the process of deriving conclusions, making inferences, and supporting decision-making based on existing knowledge and evidence, such as mathematical reasoning (Anand et al. 2025; Huang et al. 2025), commonsense reasoning (Xiong et al. 2025; Li et al. 2024), and counterfactual reasoning (Chen et al. 2025b; Bynum et al. 2024). These reasoning tasks can be interpreted within the framework of dual-system theory (Sloman 1996; Evans and Stanovich 2013). This theory distinguishes between System 1, which is fast, intuitive, and pattern-oriented, and System 2, which is slow, more deliberate, and guided by explicit rules and sequential logic.

Contemporary AI systems, particularly Large Language Models (LLMs), exhibit strengths analogous to an enhanced System 1: they excel at rapid pattern recognition, contextual adaptation, and generating plausible inferences when supported by abundant training data. LLMs have been observed to achieve significant high accuracy in number sequence inductive reasoning (Cobbe et al. 2021; Hendrycks et al. 2021), where a representative query is "2, 4, 6, 8, what is the next number?" and the correct answer is "10". System 1 reasoning does not involve explicit reasoning steps: the system only returns an answer to the query. The implicit reasoning steps during model inferences, however, may contain: (1) observing the given numbers, (2) getting an intuitive sense that they follow an increasing pattern, (3) hypothesizing that the rule is "+2", and (4) producing the next number. Note that System 1 reasoning is conclusion-oriented, relying on intuition, pattern recognition, and memory retrieval, which can be characterized as fast but not transparent.

By contrast, current AI systems still face significant challenges in performing robust System 2–style reasoning, which demands sustained multi-step logic, explicit rule-following, and verifiable chains of thought. This gap is especially evident in domains such as mathematical theorem proving (Cao et al. 2025; Welleck et al. 2021), formal logic (Kuzelka 2023), and other structured reasoning tasks. When asked to prove that the square root of 2 is irrational, a model must not only recall the relevant proof technique (i.e., proof by contradiction) but also accurately construct a sequence of logically dependent steps (i.e., assuming $\sqrt{2} = p/q$ in lowest terms, deducing that both p and q must be even, and identifying the resulting contradiction) without omitting or misordering any part of the argument. Unlike simple pattern recognition in number sequences, System 2 reasoning requires deliberate reasoning, error-resistant logic maintenance, and justification at each stage, where current models are far more prone to mistakes.

Within the scope of System 2-style reasoning, *legal reasoning* represents a particularly demanding subset. Legal reasoning underpins a wide range of tasks in the AI & Law domain, including Legal Judgment Prediction (Feng et al. 2022; Wu et al. 2023), Legal Case Retrieval (Feng et al. 2024; Li et al. 2025), Legal Entailment (Custeau and Inkpen 2025), Legal Question Answering (QA) (Louis et al. 2024), and Legal Argument Mining (Habernal et al. 2024). Legal systems depend on rigorous logic, explainability, and transparency to ensure fairness and accountability. Unlike most AI applications, legal decision-making has direct implications for individuals' rights, obligations, and access to justice. Thus, AI-driven legal solutions must transcend surface-level predictions and produce structured, interpretable reasoning chains grounded in legal authority. By examining the chain of reasoning steps, a user can under-

*Corresponding author

Issue: Whether the defendant is excused from performance under the doctrine of frustration.
Rule: A contract may be discharged if an unforeseen event occurs, rendering performance impossible or different, and the event is not self-induced (**Taylor v. Caldwell** [1863]).
Application: Evidence shows the venue was destroyed by a natural disaster before the scheduled event, without fault of either party. According to the precedent, this qualifies as frustration: the venue's destruction made performance impossible, satisfying the criteria in *Taylor v. Caldwell*. Furthermore, public policy supports excusing the party to avoid unfair penalization for uncontrollable events.
Conclusion: The defendant is discharged from the contractual obligation.

Table 1: Example application of IRAC.

stand how a prediction is made by a legal system, thus improving user confidence.

Broadly, legal reasoning can be defined as the structured reasoning process of interpreting and applying legal norms, principles, and precedents to specific factual circumstances in order to derive conclusions that are logically coherent, normatively justified, and verifiable. The task of Legal Judgment Prediction (LJP), for instance, can be used to illustrate these requirements in practice. LJP aims to predict the outcome of a legal case, such as whether the defendant is guilty or not, as well as the applicable charges, given the textual description of the facts of the case (Feng et al. 2022; Shi et al. 2025; Upadhya and Santosh 2025). While LJP is often cast as a text classification problem that maps case descriptions to outcomes, effective legal reasoning calls for domain-expert IRAC reasoning (Christensen and Kift 2000; Babusiaux 2016), which involves four steps: (1) *Issue*: identifying the legal question(s) or issue(s) that need to be resolved; (2) *Rule*: stating the legal rule(s), including the statutes, regulations, and principles from case law that are applicable to the issue; (3) *Application (or Analysis)*: applying the legal rules to the specific facts of the case, often involving making analogies to precedent and consideration of counterarguments; and finally (4) *Conclusion*: concluding with a well-reasoned answer to the legal issue based on the application of the rules.

As an example of how the IRAC framework can be applied, consider a scenario in which there is a contract dispute over non-performance due to an unforeseen event. More specifically, a music festival organizer entered into a contract to rent a venue for a large public event, but shortly before the scheduled date, a government-mandated COVID-19 lockdown prohibited all mass gatherings, rendering the event impossible to hold. The reasoning structure produced by IRAC when applied to this scenario is shown in Table 1.

As can be seen from the above example (and the resulting reasoning structure), legal reasoning requires more than generic logical analysis, as it must integrate multiple layers of normative assessment. First, the decision-maker must identify and apply the relevant legal norms, such as the doctrine of frustration, which under common law excuses contractual performance when an unforeseeable event—beyond the control of the parties—makes fulfillment impossible or radically difficult. Second, the process demands precedent analysis, comparing the case to earlier decisions like Taylor v. Caldwell (1863), where the destruction of a concert hall prior to a concert excused performance, establishing a principle closely analogous to the current facts. Third, legal reasoning often incorporates policy balancing, evaluating whether excusing performance aligns with broader societal values, such as public health priorities and fairness in the allocation of risk during a pandemic. These steps demonstrate that legal reasoning operates through not only structured logical deduction, but also the coordinated application of codified norms, case-based analogical reasoning, and policy-sensitive judgment.

The examples we have seen so far illustrate that legal reasoning differs markedly from both System 1 and System 2 reasoning. Unlike System 1–style reasoning, which relies on fast, intuitive pattern recognition and yields outputs without interpretations, legal reasoning requires deriving logical conclusions with a step-by-step justification. Unlike System 2, which pursues logical validity, legal reasoning pursues legal acceptability and embeds formal legal norms, precedent reasoning, and policy evaluation, enabling judgments that meet both logical and legal standards. While recent advances in legal reasoning have improved the retrieval and application of legal information, they remain insufficient to meet the demands of legal reasoning, for the following reasons.

First, current systems tend to display shallow integration of legal norms. For example, the doctrine of frustration may be applied as a static checklist: ["Was the event unforeseeable?" $\rightarrow$ Yes, "Is performance impossible?" $\rightarrow$ Yes], leading to the conclusion that the contract is discharged. Such an approach omits the interpretive methods a legal professional would employ, such as analyzing whether the pandemic restrictions truly meet the jurisdiction's doctrinal threshold for "impossibility", or whether obligations could be modified rather than entirely avoided.

Second, many systems exhibit fragmented reasoning with precedents. An AI tool may retrieve Taylor v. Caldwell (1863) as a "similar" case because it contains keywords like "venue" and "impossible performance". Yet a professional lawyer would go further by mapping the facts, noting both the parallels (unforeseeable event, outside parties' control) and differences (physical destruction vs. temporary legal prohibition), and deciding whether the precedent's principle applies in full or requires modification. Without structured analogical reasoning, precedent use remains superficial and potentially misleading.

Third, existing approaches often suffer from a lack of policy-sensitive balancing. A court weighs public health priorities against the financial harm to the venue owner, exploring whether losses should be shared (e.g., partial refunds) rather than placing the full burden on one party. Many current reasoning systems do not integrate this evaluative step, stopping at doctrinal application without considering how broader societal objectives can influence the chosen remedy.

Taken together, these limitations imply that current systems may produce outputs that are internally consistent yet legally incomplete. Specifically, they fail to integrate formal doctrinal rigor, structured precedent evaluation, and nor-

mative policy judgment into a unified, transparent decision-making process—capabilities that are central to achieving the legal reasoning standard.

To overcome these limitations, we propose System L, a legal reasoning framework that unifies deep integration of legal norms and interpretive methods, structured and explainable precedent analysis, and policy-sensitive value balancing within a transparent IRAC-based process. By embedding doctrinal rigor, analogical reasoning, and societal objective alignment into a single, auditable pipeline, System L not only addresses the technical shortcomings of current AI-based legal reasoning but also advances the broader vision of AI for Social Good—promoting fairness, expanding access to justice, strengthening public trust, and ensuring that legal conclusions are both doctrinally sound and socially beneficial. From an AI for Social Good perspective, enhancing AI's capacity for legal reasoning is pivotal, as legal systems underpin the protection of rights, the resolution of disputes, and the maintenance of social order. By improving the accuracy, transparency, and normative alignment of AI-driven legal analysis, System L can help democratize access to legal knowledge, promote consistency in judicial decision-making, and reduce disparities in the application of law, especially in low-resource contexts where timely and affordable legal assistance is scarce. In doing so, AI can complement human adjudication, strengthen procedural fairness, and reinforce societal trust in legal institutions.

2 Related Work

Broadly, existing Legal AI systems can be divided into two categories: System 1-style and System 2-style systems.

System 1-Style Systems

Early legal AI systems often exhibit no explicit reasoning beyond information retrieval or statistical pattern matching. These System 1-style approaches can learn correlations between text and outcomes, but cannot articulate why a conclusion follows from the given facts. For example, Chalkidis et al. (2019) model LJP as a text classification task and employ a BERT-based model to predict judgment labels directly without outputting any intermediate reasoning steps.

System 2-Style Systems

Early systems. Unlike their System 1 counterparts, System 2–style models incorporate explicit reasoning steps. Early works in this category generate intermediate representations such as rule–fact mappings (Deng et al. 2023; Luo et al. 2017; Xu et al. 2020). For example, in a contract dispute, the model first identifies legal rules from statutory text, e.g., "A contract is void if entered into under coercion", as the major premise. It then detects the facts in the case description indicating coercion (e.g., "Plaintiff was threatened with physical harm unless he signed") as the minor premise, explicitly linking each fact to the corresponding legal rule element before concluding that the contract is void.

There are also works performing precedent similarity assessments (Wu et al. 2023; Shao et al. 2023). Given a negligence case, the system retrieves prior cases involving car accidents (Donoghue v. Stevenson, Caparo Industries plc v. Dickman), then encodes fact patterns using legal-domain embeddings. It compares these representations to assess similarity (e.g., "duty of care existed", "breach caused foreseeable harm") and uses the closest precedent's legal principle as the basis for reasoning in the new dispute.

Other works decompose tasks into sub-questions aligned with legal concepts (Zhong et al. 2020a; Hu et al. 2018). In a criminal law scenario, the system decomposes the problem into sub-questions reflecting statutory elements of the offence - (1) Was there an act causing harm?, (2) Was there intent?, and (3) Do any statutory defences apply? - and answers them sequentially. The answers are then composed into a full legal argument aligned with the statute's structure.

While these strategies make parts of the reasoning chain visible, they often remain incomplete: rule–fact links may be shallow, precedent comparisons may overlook legally significant distinctions, and decomposed reasoning may fail to integrate policy considerations or weigh conflicting values.

Systems producing human-like reasoning. Beyond static reasoning chains, LLMs have been used to perform explicit legal reasoning when addressing LJP problems. Specifically, given the facts of a legal case, LLMs have been instructed via prompting to predict court decisions and at the same time output the reasoning steps (in the form of a natural language paragraph) that eventually led to those predictions. As is known among LJP researchers, these LLM-generated paragraphs are not without their problems. First, legal jargons (e.g., "the scope of intentional infliction of bodily harm", "a second-degree minor injury") are often used without being clearly explained in lay terms, making it difficult for users without a legal background to understand. Second, the reasoning structure is often unclear and has a weak logical flow. Specifically, those outputs often do not match the legal provisions with the case details step by step (e.g., explaining how the defendant's actions meet the requirements for intentional injury under the law, or why they do not qualify as self-defense). Instead, they simply list multiple legal references without integrating them into a clear argument.

In another line of research, LLMs are used to simulate trials. For instance, Chen et al. (2025a) provide a trial debate simulation framework, in which lawyer agents can evolve with adversarial debating. He et al. (2024) propose a judicial decision-making agent with trial debate simulation and legal knowledge augmentation. Zhang et al. (2025) focus on procedural realism and improve the systematic process design and evaluation of simulations.

These procedural simulations mark an important step toward end-to-end reasoning, embedding discourse and evidence evaluation into the process. Yet, much like earlier reasoning systems, they still face several challenges. While they reproduce the form of courtroom dialogue, the substance of doctrinal interpretation and precedent application tends to be shallow. Specifically, precedent use may be unstructured, with retrieved cases mentioned in passing rather than systematically analogised. Moreover, policy balancing is rarely explicit, meaning that the simulated judgments may lack the normative deliberation present in real courts.

3 System L Reasoning Framework

The aforementioned limitations of existing legal AI systems, all of which employ either System 1 or System 2-style reasoning, motivate the design of System L, our proposed framework for legal reasoning. System L is built upon the well-established IRAC backbone (see the Introduction). While IRAC provides a clear and pedagogically sound sequencing of reasoning stages, its conventional form often under-specifies the internal complexity of each stage from the implementation standpoint. To address this, System L extends and refines the IRAC framework by embedding additional, interlinked reasoning layers that more closely reflect expert legal practice, as discussed below.

Step I: Issue Analysis

This step aims to identify the legal *issues*, which are issues that require resolution from a case description. It is important because missing or misidentifying an issue can cause the reasoning chain to apply an irrelevant law or overlook critical arguments.

The output of this step is a set of legal issues and the statements in the input from which they are inferred. Consider a case in which a collision occurred when driver A, traveling 65 km/h in a 50 km/h zone and turning left on a green light, struck person B as he entered a marked crosswalk on a green pedestrian signal while pushing an e-bike. The manufacturer plate showed the e-bike's rated maximum speed as 28 km/h. A system should identify the issue "Whether B is a pedestrian" and the relevant fact statement "he entered a marked crosswalk on a green pedestrian signal while pushing an e-bike. The manufacturer plate showed the e-bike's rated maximum speed as 28 km/h". This issue is legally significant because the classification of B's status directly determines the allocation of liability in traffic accident law as non-pedestrians bear more responsibility than pedestrians.

Challenge 1: Extracting implicit issues. Legal issues often require making inferences from the background fact statements. For instance, in the traffic collision example, the fact statements do not reveal the legal issue directly. While recent claim detection approaches have improved the extraction of surface-level claims, they remain limited when confronted with normative reasoning that depends on a domain-specific schema (Ni et al. 2024; Stammbach et al. 2023). Identifying implicit issues demands not only recognizing textual signals, but also mapping fact patterns to legally codified categories within a jurisdiction's normative framework.

To address this challenge, we suggest a two-phase pipeline, where we (1) extract the *relevant* fact statement (i.e., the sentences from which a legal issue can be derived) from the input, and then (2) identify the legal issue from each fact statement extracted in the first phase. To train a model for fact extraction and issue generation, we can build a training set where each input case is annotated with a set of (fact statement, issue) pairs. We can adapt existing resources such as COLIEE Task 4 (Case Entailment) (Tang et al. 2025), JEC-QA (Legal QA) (Zhong et al. 2020b), and manually add fact statement and issue annotations.

Challenge 2: Handling dynamic reframing of issues. Legal issues are dynamic as (1) real-world litigation is iterative and (2) the legal issues themselves evolve as additional facts are unveiled over time. As an example, suppose that the initial fact is "Defendant accused of stealing a company laptop" and the initial issue is Theft of Property. If additional evidence reveals that the laptop contained trade secrets, the issue should be updated to Misappropriation of Confidential Information. Existing legal NLP systems simply treat inputs as static, ignoring the temporal evolution of disputes. There is virtually no pipeline for legal issue state tracking in dynamic, multi-turn contexts, a gap that exists even in the broader state-of-the-art dialogue and reasoning literature.

To train a model that can handle dynamic reframing of issues, we propose adapting dialogue state tracking frameworks (Pyun et al. 2025) where "slots" correspond to legal issues, with a change detection module to flag new, dropped, or modified issues, and event-time reasoning to capture fact sequences that trigger legal reclassification. Specifically, to create data for training this model, we can curate multi-turn case timelines by segmenting judgments, court hearing transcripts, and legal reports into chronological fact snapshots, annotating the active issue set at each stage. A good starting point would be to adapt the MultiWOZ dialogue dataset format (Quan et al. 2020) to store evolving fact–issue states, and initially populate it using publicly available hearings (e.g., U.S. federal/state archives) and manually annotated case timelines.

Step II: Legal Rule Identification

The next step is Legal Rule Identification (LRI). Specifically, for each (fact statement, issue) pair identified in Step I, this step aims to identify all the norm(s) that are applicable to the issue, such as statutes, regulations, and case law principles. Failure to do so will cause the system to apply irrelevant norms, yielding incorrect predictions. To exemplify, reconsider the traffic collision example. Given the issue and the relevant fact statement, a model should output the regulation GB17761-2018, in which an e-bike qualifies as a non-motor vehicle only if it satisfies the national standard, including maximum design speed $\leq$ 25 km/h. Below we discuss two key challenges associated with this step.

Challenge 3: Reducing the risk of shallow norm integration. Shallow norm integration occurs when a system identifies a rule's minimal triggering conditions or relies on surface-level fact matching, but neglects the rule's underlying multi-factor and multi-condition doctrinal structure.

Consider the traffic collision example again. If a system applies only surface-level matches, e.g., [Is person B in a marked crosswalk? $\rightarrow$ Yes] and [Is person B on foot? $\rightarrow$ Yes], it will incorrectly match the fact and the issue directly to Traffic Law, Article 119 (defining a "pedestrian" as a person walking on the road). It will then lead to the absolute pedestrian-priority, assigning primary liability to the driver.

In contrast, a model that performs deep doctrinal integration would interpret the regulation through the lens of underlying legal doctrines and assess whether B's device satisfies the legal criteria for a motor vehicle—for example, by comparing the manufacturer-rated maximum speed with the 25

km/h statutory threshold. If the device exceeds this limit, B would not be classified as a legal "pedestrian", even when walking alongside it. Consequently, the conclusion would shift from granting absolute pedestrian priority to applying a mixed-duty framework, which necessitates a comparative fault assessment between the driver and the crosser. Hence, a system must capture not only surface-level factual cues but also the doctrinal standards embedded within the rule that determine legal conclusions, such as the speed threshold.

To address this challenge, we propose a two-phase approach. First, we build an annotated doctrinal corpus in which each rule is explicitly encoded in its doctrinal structure, including (1) *Citation*, for traceability; (2) *Text*, the exact statutory or precedential wording, for semantic precision; (3) *Legal elements*, the doctrinal criteria (e.g., factors, conditions, threshold) for operational applicability; and (4) *Applicability*, metadata such as jurisdiction, effective dates, for correct contextual usage. The purpose of having this *structured* representation of a rule is to make it explicit the doctrinal criteria associated with each rule, so as to reduce the risk of the shallow norm integration problem mentioned above. In the second phase of our approach, we train a model that outputs all the rules that are applicable to each (fact statement, issue) pair extracted from Step I. Specifically, this model takes as input not only a (fact statement, issue) pair but also the entire set of rules obtained from Phase 1, and retrieves all and only those rules that are applicable. To train this model, we can develop a fact/issue–rule mapping dataset linking facts/issues to the cited rule(s).

Step III: Rule Application

In this step, we apply a rule to an issue to derive a conclusion. In our collision example, rule application involves applying the rule GB17761-2018 to the issue "Whether B is a pedestrian" to derive the conclusion "B should be classified as a motor vehicle operator and the mixed-duty rule should be applied". The conclusion could be the *final* conclusion (e.g., the predictions made by an LJP system such as the charges and term of penalty for a case), or, in our collision example, an *intermediate* conclusion in the reasoning process that needs to be combined with other intermediate conclusions using another rule to derive the final conclusion.

At first glance, rule application seems trivial: since the rules applicable to a (fact statement, issue) pair have already been identified in Step II, all we need to is to apply the rule to obtain the conclusion. Nevertheless, there are two key challenges associated with rule application.

Challenge 4: Precedent-grounded doctrinal mapping. Even with a doctrinally structured rule in hand, a system needs to align each operative legal/doctrinal element with the relevant facts/issues, a process that often requires analogical reasoning with precedents. Many current systems still fail at this task. Consider again the contract dispute example in the Introduction: an AI tool might apply *Taylor v. Caldwell* merely because it shares surface terms like "venue" and "impossible performance" with the facts/issues. An expert lawyer, by contrast, would ground the analysis in doctrinal elements. This involves first identifying the applicable elements (e.g., the occurrence of an unforeseeable event beyond the parties' control) and then systematically comparing the facts with the precedent. The comparison would highlight the parallel, such as an external impediment to performance, to show how the facts align with the doctrinal elements in the precedent, and the distinction, such as physical destruction of the venue versus a temporary legal prohibition, to assess whether the difference is significant enough to affect the application of the law. Without a structured, element-by-element mapping from precedent to current facts, even a doctrinally represented rule remains vulnerable, potentially misleading application.

To address this challenge, an initial step could be to construct a hand-annotated dataset that decomposes a precedent's rule into its individual doctrinal elements and maps each element to the corresponding facts. For each element, the mapping would specify whether the precedent is applicable (parallel) or not (distinction). A model for constructing this mapping can then be trained on this dataset.

Challenge 5: Policy-sensitive balancing. Existing approaches often suffer from a lack of policy-sensitive balancing. Courts frequently weigh doctrinal outcomes against broader societal objectives. For example, in the pandemic-related frustration claim we saw in the music festival example in the introduction, the court may need to balance public health priorities against the financial harm to the venue owner, and consider remedial options like partial loss-sharing instead of placing the full burden on one party. Current systems stop at doctrinal matching, failing to incorporate this evaluative step where societal and remedial policy considerations may influence the outcome.

To address this challenge, we propose augmenting the reasoning process with an explicit policy-sensitive balancing layer following doctrinal analysis. Specifically, the system would undergo (1) a *doctrinal* stage, in which an element-by-element doctrinal mapping is constructed; (2) a *policy evaluation* phase, in which relevant societal objectives (e.g., public health, economic stability, environmental protection) are identified and assigned indicative weights; and (3) a *calibration* stage, where the final conclusion is assessed against both doctrinal outcomes and policy priorities.

Step IV: Argument Tree Generation

By the end of Step III, what we obtain is an *argument*, which is composed of a *conclusion* (Step III's output) as well as the *facts and issues* extracted in Step I that are used to derive the conclusion by applying the *rule* identified in Step II. In Computational Argumentation, an argument is commonly represented as a *tree*. In this tree, each leaf node corresponds to a fact/issue that is relevant to the derivation of a conclusion, and all these leaves have the same parent node, which corresponds to an intermediate or final conclusion. For our collision example, the argument tree would be composed of a parent node that corresponds to the intermediate conclusion "B should be classified as a motor vehicle operator and the mixed-duty rule should be applied".

As mentioned before, an intermediate conclusion will be combined with other facts, issues, and/or intermediate conclusions to derive high-level conclusions. This process will

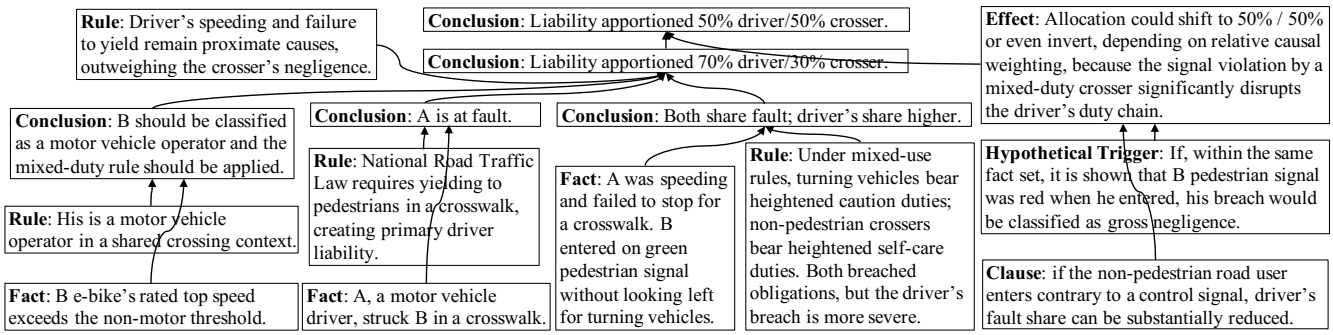

Figure 1: Argument tree for the collision example.

continue until the final conclusion is reached. In other words, Steps II and III need to be repeated in order to produce additional arguments, and the argument tree will be incrementally augmented with these additional arguments.

Returning to the collision example, in addition to the argument that comprises the intermediate conclusion "B should be classified as a motor vehicle operator and the mixed-duty rule should be applied" (henceforth Argument 1), there are three arguments involved, as shown in the top three blocks of Table 2. Specifically, Arguments 1, 2, and 3 can all be generated independently of each other: Arguments 2 and 3 are generated using the case facts as input, whereas Argument 1 is generated using the implicit issue as input. Argument 4, on the other hand, is generated by applying a rule to the intermediate conclusions from the first three arguments to derive the final conclusion for this example. Hence, the resulting argument tree will be composed of all four arguments, with Argument 4's conclusion being the root node, the remaining three arguments' conclusions being the children of the root node, and the facts/issues supporting each intermediate conclusion being the leaves.

While argument trees are a widely known concept in Computational Argumentation, to our knowledge we are the first to propose representing legal reasoning as an argument tree. The key advantage of this argument tree representation is that it makes the reasoning process completely unambiguous, as it is clear which fact(s), issue(s), and/or intermediate conclusion(s) are involved in each step of the process.

Challenge 6: Handling exceptions. The argument tree construction task is compounded by the need to integrate *defeasible reasoning*, where legal conclusions can be overridden by *exceptions*, and *deontic reasoning*, which encodes obligations, permissions, and prohibitions. Without such machinery, AI-generated arguments risk being brittle, overly absolute, and non-reflective of real-world legal reasoning where rules are not universally binding and context can generate justified departures.

To handle exceptions, one can consider leveraging Defeasible Deontic Logic (DDL) (Governatori 2005) to formalize obligations and rights in a way that allows exceptions and priority rules to be integrated with an argument tree. Returning to our collision example, an exception clause (see the last block of Table 2) is applicable. The argument tree with this exception integrated is shown in Figure 1.

Argument 2:
Fact: A, a motor vehicle driver, struck B in a crosswalk.
Rule: National Road Traffic Law requires yielding to pedestrians in a crosswalk, creating primary driver liability.
Conclusion: A is at fault.
Argument 3:
Facts: A was speeding and failed to stop for a crosswalk. B entered on green pedestrian signal without looking left for turning vehicles.
Rule: Under mixed-use rules, turning vehicles bear heightened caution duties; non-pedestrian crossers bear heightened self-care duties. Both breached obligations, but the driver's breach is more severe.
Conclusion: Both share fault; driver's share higher.
Argument 4
Intermediate conclusions: Those from Arguments 1, 2, 3.
Rule: Driver's speeding and failure to yield remain proximate causes, outweighing the crosser's negligence.
Conclusion: Liability apportioned 70% driver/30% crosser.
Exception
Exception Clause: if the non-pedestrian road user enters contrary to a control signal, driver's fault share can be substantially reduced.
Hypothetical Trigger: If, within the same fact set, it is shown that B pedestrian signal was red when he entered, his breach would be classified as gross negligence.
Effect: Allocation could shift to 50% / 50% or even invert, depending on relative causal weighting, because the signal violation by a mixed-duty crosser significantly disrupts the driver's duty chain.

Table 2: Other arguments involved in the collision example. While the rules here are expressed in natural language, in reality a structured rule representation is used (see Step II).

4 Conclusion

We discussed our vision for System L, a framework we proposed for AI-assisted legal reasoning that unites doctrinal depth, structured factual analysis, and transparent argumentation. We believe that System L offers a pathway toward AI systems that can operate as trustworthy, explainable, and normatively sensitive partners in legal decision-support, setting a foundation for transformative impact in both academic research and practical legal workflows.

Acknowledgments

We thank the reviewers for their valuable comments on an earlier draft of this paper. This work was supported by National Natural Science Foundation of China (No. 62406139), State Key Laboratory for Novel Software Technology at Nanjing University (KFKT2025A15, ZZKT2025B14, KFKT2024A07, ZZKT2024B02).

References

Anand, A.; Prasad, K.; Kirtani, C.; Nair, A. R.; Nema, M. K.; Jaiswal, R.; and Shah, R. R. 2025. Multilingual Mathematical Reasoning: Advancing Open-Source LLMs in Hindi and English. In *Proceedings of the 39th AAAI Conference on Artifical Intelligence*, 23415–23423. Philadelphia, PA, USA: AAAI Press.

Babusiaux, U. 2016. Legal writing and legal reasoning. *The Oxford Handbook of Roman Law and Society, Oxford*, 176–187.

Behnke, G. 2024. Symbolic Reasoning Methods for AI Planning. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 22661. Vancouver, Canada: AAAI Press.

Bynum, L. E. J.; Loftus, J. R.; and Stoyanovich, J. 2024. A New Paradigm for Counterfactual Reasoning in Fairness and Recourse. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, 7092–7100. Jeju, South Korea.

Cao, C.; Fu, Y.; Xu, S.; Zhang, R.; and Li, S. 2024. Enhancing Human-AI Collaboration Through Logic-Guided Reasoning. In *Proceedings of the 12th International Conference on Learning Representations*. Vienna, Austria.

Cao, C.; Li, M.; Dai, J.; Yang, J.; Zhao, Z.; Zhang, S.; Shi, W.; Liu, C.; Han, S.; and Guo, Y. 2025. Towards Advanced Mathematical Reasoning for LLMs via First-Order Logic Theorem Proving. *CoRR*, abs/2506.17104.

Chalkidis, I.; Androutsopoulos, I.; and Aletras, N. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4317–4323. Florence, Italy: Association for Computational Linguistics.

Chen, G.; Fan, L.; Gong, Z.; Xie, N.; Li, Z.; Liu, Z.; Li, C.; Qu, Q.; Alinejad-Rokny, H.; Ni, S.; and Yang, M. 2025a. AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, 5850–5865. Vienna, Austria: Association for Computational Linguistics.

Chen, J.; Wang, Y.; Wang, J.; Xie, X.; Hu, J.; Wang, Q.; and Xu, F. 2025b. Understanding Individual Agent Importance in Multi-Agent System via Counterfactual Reasoning. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 15785–15794. Philadelphia, PA, USA: AAAI Press.

Christensen, S.; and Kift, S. 2000. Graduate attributes and legal skills: Integration or disintegration. *Legal Education Review*, 11: 207.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *CoRR*, abs/2110.14168.

Custeau, M.; and Inkpen, D. 2025. Enhancing Legal Text Entailment: Evaluating Model Architectures, Training Approaches, and Interpretability. In *Proceedings of the 38th Canadian Conference on Artificial Intelligence*. Calgary, AB, Canada.

Deng, W.; Pei, J.; Kong, K.; Chen, Z.; Wei, F.; Li, Y.; Ren, Z.; Chen, Z.; and Ren, P. 2023. Syllogistic Reasoning for Legal Judgment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13997–14009. Singapore: Association for Computational Linguistics.

Evans, J. S. B.; and Stanovich, K. E. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3): 223–241.

Feng, Y.; Li, C.; and Ng, V. 2022. Legal Judgment Prediction via Event Extraction with Constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 648–664. Dublin, Ireland: Association for Computational Linguistics.

Feng, Y.; Li, C.; and Ng, V. 2024. Legal Case Retrieval: A Survey of the State of the Art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6472–6485. Bangkok, Thailand: Association for Computational Linguistics.

Governatori, G. 2005. Representing Business Contracts in *RuleML*. *International Journal of Cooperative Information Systems*, 14(2-3): 181–216.

Gupta, V. 2025. Advancements in AI for Reasoning with Complex Data. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 28711. Philadelphia, PA, USA: AAAI Press.

Habernal, I.; Faber, D.; Recchia, N.; Bretthauer, S.; Gurevych, I.; genannt Döhmann, I. S.; and Burchard, C. 2024. Mining Legal Arguments in Court Decisions. *Artificial Intelligence and Law*, 32(3): 1–38.

He, Z.; Cao, P.; Wang, C.; Jin, Z.; Chen, Y.; Xu, J.; Li, H.; Jiang, X.; Liu, K.; and Zhao, J. 2024. AgentsCourt: Building Judicial Decision-Making Agents with Court Debate Simulation and Legal Knowledge Augmentation. *arXiv preprint arXiv:2403.02959*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*. Virtual.

Hu, Z.; Li, X.; Tu, C.; Liu, Z.; and Sun, M. 2018. Few-Shot Charge Prediction with Discriminative Legal Attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, 487–498. Santa Fe, New Mexico, USA.

Huang, Y.; Liu, X.; Gong, Y.; Gou, Z.; Shen, Y.; Duan, N.; and Chen, W. 2025. Key-Point-Driven Data Synthesis with

Its Enhancement on Mathematical Reasoning. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 24176–24184. Philadelphia, PA, USA: AAAI Press.

Kuzelka, O. 2023. Counting and Sampling Models in First-Order Logic. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 7020–7025. Macao, SAR, China.

Li, H.; Ai, Q.; Han, X.; Chen, J.; Dong, Q.; and Liu, Y. 2025. DELTA: Pre-Train a Discriminative Encoder for Legal Case Retrieval via Structural Word Alignment. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 27072–27080. Philadelphia, PA, USA: AAAI Press.

Li, J.; Cao, P.; Wang, C.; Jin, Z.; Chen, Y.; Zeng, D.; Liu, K.; and Zhao, J. 2024. Focus on Your Question! Interpreting and Mitigating Toxic CoT Problems in Commonsense Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9206–9230. Bangkok, Thailand: Association for Computational Linguistics.

Liu, Y.; Du, Y.; Ji, T.; Wang, J.; Liu, Y.; Wu, Y.; Zhou, A.; Zhang, M.; and Cai, X. 2025. The Role of Visual Modality in Multimodal Mathematical Reasoning: Challenges and Insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 22596–22611. Vienna, Austria: Association for Computational Linguistics.

Louis, A.; van Dijck, G.; and Spanakis, G. 2024. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 22266–22275. Vancouver, Canada: AAAI Press.

Luo, B.; Feng, Y.; Xu, J.; Zhang, X.; and Zhao, D. 2017. Learning to Predict Charges for Criminal Cases with Legal Basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2727–2736. Copenhagen, Denmark: Association for Computational Linguistics.

Ni, J.; Shi, M.; Stammbach, D.; Sachan, M.; Ash, E.; and Leippold, M. 2024. AFaCTA: Assisting the Annotation of Factual Claim Detection with Reliable LLM Annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1890–1912. Bangkok, Thailand: Association for Computational Linguistics.

Pyun, H.; Park, Y.; and Jo, Y. 2025. Improving Dialogue State Tracking through Combinatorial Search for In-Context Examples. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 28694–28714. Association for Computational Linguistics.

Quan, J.; Zhang, S.; Cao, Q.; Li, Z.; and Xiong, D. 2020. RiSAWOZ: A Large-Scale Multi-Domain Wizard-of-Oz Dataset with Rich Semantic Annotations for Task-Oriented Dialogue Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 930–940. Online: Association for Computational Linguistics.

Shao, Y.; Wu, Y.; Liu, Y.; Mao, J.; and Ma, S. 2023. Understanding Relevance Judgments in Legal Case Retrieval. *ACM Transactions on Information Systems*, 41(3): 76:1–76:32.

Shi, W.; Zhu, H.; Ji, J.; Li, M.; Zhang, J.; Zhang, R.; Zhu, J.; Xu, J.; Han, S.; and Guo, Y. 2025. LegalReasoner: Stepwised Verification-Correction for Legal Judgment Reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7297–7313. Vienna, Austria: Association for Computational Linguistics.

Sloman, S. A. 1996. The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, 119(1): 3.

Stammbach, D.; Webersinke, N.; Bingler, J. A.; Kraus, M.; and Leippold, M. 2023. Environmental Claim Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1051–1066. Toronto, Canada: Association for Computational Linguistics.

Tang, Y.; Qiu, R.; and Huang, Z. 2025. UQLegalAI@COLIEE2025: Advancing Legal Case Retrieval with Large Language Models and Graph Neural Networks. *CoRR*, abs/2505.20743.

Upadhya, R.; and Santosh, T. Y. S. S. 2025. LexCLiPR: Cross-Lingual Paragraph Retrieval from Legal Judgments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13971–13993. Vienna, Austria: Association for Computational Linguistics.

Welleck, S.; Liu, J.; Bras, R. L.; Hajishirzi, H.; Choi, Y.; and Cho, K. 2021. NaturalProofs: Mathematical Theorem Proving in Natural Language. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*. Virtual.

Wu, Y.; Zhou, S.; Liu, Y.; Lu, W.; Liu, X.; Zhang, Y.; Sun, C.; Wu, F.; and Kuang, K. 2023. Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12060–12075. Singapore: Association for Computational Linguistics.

Xiong, K.; Ding, X.; Cao, Y.; Yan, Y.; Du, L.; Zhang, Y.; Gao, J.; Liu, J.; Qin, B.; and Liu, T. 2025. Com^2 : A Causal-Guided Benchmark for Exploring Complex Commonsense Reasoning in Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16119–16140. Vienna, Austria: Association for Computational Linguistics.

Xu, N.; Wang, P.; Chen, L.; Pan, L.; Wang, X.; and Zhao, J. 2020. Distinguish Confusing Law Articles for Legal Judgment Prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3086–3095. Online: Association for Computational Linguistics.

Zhang, K.; Li, J.; Wu, Y.; Li, H.; Luo, C.; Zou, S.; Zhou, Y.; Su, W.; Ai, Q.; and Liu, Y. 2025. Chinese Court Simulation with LLM-Based Agent System. *arXiv preprint arXiv:2508.17322*.

Zhong, H.; Wang, Y.; Tu, C.; Zhang, T.; Liu, Z.; and Sun, M. 2020a. Iteratively Questioning and Answering for Interpretable Legal Judgment Prediction. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 1250–1257. New York, NY, USA: AAAI Press.

Zhong, H.; Xiao, C.; Tu, C.; Zhang, T.; Liu, Z.; and Sun, M. 2020b. JEC-QA: A Legal-Domain Question Answering Dataset. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 9701–9708. New York, NY, USA: AAAI Press.