OBJECTION!
Lawyer Agents Mitigate Guilty Bias in Legal Judgment Prediction

Jaehoon Jeong[1,2] **Jay-Yoon Lee**[1,*]
[1]Seoul National University, Seoul, South Korea
[2]Korean National Police Agency, Seoul, South Korea
{20210042, lee.jayyoon}@snu.ac.kr

Abstract

Legal Judgment Prediction (LJP) models are typically trained on documents that describe facts from a prosecutorial perspective. Existing datasets further exhibit severe label imbalance toward guilty outcomes. Consequently, these models suffer from "Guilty Bias", blindly accepting the prosecution's narrative as objective truth. Previous studies employing three-step reasoning structures or training on synthetically generated innocence data improve overall accuracy, but they still fail to mitigate bias at inference time.

In this paper, we introduce OBJECTION, an inference-time pipeline that integrates an Adversarial Lawyer Agent into each 3-step reasoning of offense, unlawfulness, and culpability. Unlike generic critics, our agent actively challenges the model's presumptions of guilt by injecting legal defense arguments at each reasoning stage. To thoroughly evaluate this, we present a new "Natural Innocent" dataset including 3.4k real-world cases[1], overcoming the limitations of synthetic innocence benchmarks. Test results show that OBJECTION drastically reduces the False Guilty Rate (FGR) from 82.93% (SOTA baseline) to 16.69%, proving its capability to perform substantive legal reasoning. This work denotes a key progress toward aligning Legal AI with the presumption of innocence.

1 Introduction

The maxim "It is better that ten guilty persons escape than that one innocent suffer" epitomizes the presumption of innocence, a basic tenet of criminal law. However, criminal trials are not neutral by design. The very fact that a case has reached the courtroom implies that investigative bodies have established a sufficient suspicion of guilt. Consequently, the 'written judgment' inevitably reflects this bias. These fact descriptions are not neutral event summaries, but prosecutorial narratives written to justify guilt.

The critical issue is that the vast majority of legal texts used to train LLMs consist of precisely this type of content. In Legal Judgment Prediction (LJP), models are typically trained on these prosecutions' stated facts (Luo et al., 2017; Xiao et al., 2018; Zhong et al., 2018; Cui et al., 2023). As a result, models are systematically exposed to guilt-biased narratives even if the actual verdict is not guilty. This guilt-biased nature poses a serious impediment to real-world application. Convicting an innocent individual undermines public trust; thus, a high False Guilty Rate (FGR) is impractical and normatively unacceptable.

Previous research has attempted to mitigate this bias by structured reasoning, such as explicit element schemas (Liu et al., 2025) or multi-step legal reasoning frameworks (Jiang and Yang, 2023; Deng et al., 2023; Zhang et al., 2025b). However, these approaches fundamentally assume that the input fact description is unbiased. As a result, they inherit the guilty bias embedded in the narrative itself. Even recent efforts based on synthetic innocent cases (Zhang et al., 2025a) largely rely on training-time correction. Such synthetic innocence often introduces artificial cues that models can learn as shortcuts, leading to overfitting, which we empirically observe, while inference-time bias remains unaddressed.

We argue that guilty bias should be handled at inference time to neutralize bias effectively. We propose OBJECTION, a lawyer-agent-based critique-and-revise pipeline. In contrast to generic critics (Madaan et al., 2023), the Lawyer Agent does not merely act as a consistency checker. Instead, it operates under the principle of *In Dubio Pro Reo*, actively excavating reasonable doubt hidden within

* Corresponding author.

[1]Available at https://huggingface.co/datasets/Kcsp0042/natural-innocent. The public release (3,200 cases) is a revised edition of the data used here.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 34205–34225
October 24-29, 2026 ©2026 Association for Computational Linguistics

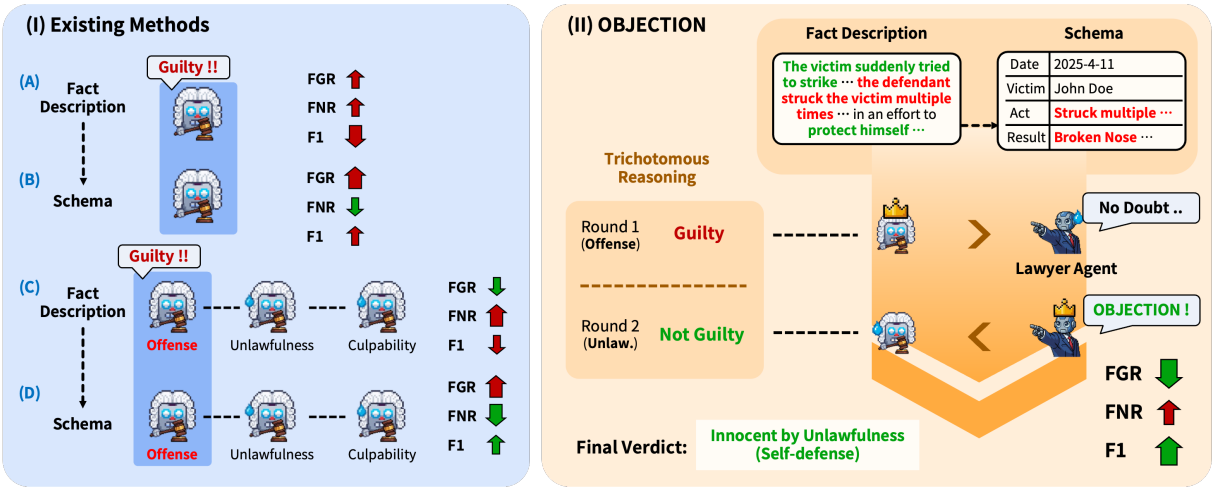

Figure 1: Comparative overview of LJP paradigms corresponding to Table 1 and Sec. 4.2. FGR(↓) and FNR(↓) denote False Guilty Rate and False Not-Guilty Rate, respectively. Green arrow denotes improvements, while the reds denote degradations. Existing methods (A)–(D) suffer from guilty bias from fixed architectures. In contrast, (II) OBJECTION (Ours) employs a Lawyer Agent to actively dispute assumptions, efficiently mitigating bias while maintaining reasoning quality. The further reasoning round (e.g., Round 3) is skipped when a former legal doubt (e.g., Self-defense in Round 2) is established. For a detailed decision flow, see Fig. 2.

the prosecutorial narrative to challenge overconfident guilt assumptions. By integrating Schema-based Information Extraction (IE) and the 3-step reasoning structure into this agent, OBJECTION introduces a training-free pipeline that aligns with LLM inference while incorporating the adversarial nature of criminal law, effectively counteracting guilty bias.

Our contributions are as follows:

- We quantify guilty bias in LJP systems using normative metrics such as False Guilty Rate (FGR), and show that prior LJP models exhibit a substantial inclination toward guilty predictions.
- We propose **OBJECTION**, an automatic judge framework with an adversarial lawyer agent without any extra training, and show that bias in existing frameworks can be neutralized by our proposed adversarial inference.
- We introduce a **Natural Innocent** dataset of real-world acquittal cases and reveal problems of existing approaches.

2 Related Works

2.1 Legal Judgment Prediction and Guilty Bias

Standard LJP settings utilize fact descriptions from written judgments, operating under the assumption that these often guilt-biased narratives represent neutral facts (Luo et al., 2017; Zhong et al., 2018; Xu et al., 2020; Yang et al., 2016; Xiao et al., 2021; Chalkidis et al., 2020). Many studies adopt Information Extraction (IE) or explicitly model Criminal Constitutive Elements to improve accuracy (Feng et al., 2022; Liu et al., 2025). These methods innovate in how they process the input but overlook the inherent bias within it, effectively inheriting the prosecutorial perspective.

Zhang et al. (Zhang et al., 2025a) were the first to explicitly identify this guilty bias and proposed synthetic innocent data and trichotomous training as remedies. While influential, this line of work handles bias primarily at training time, allowing models to rely on artificial cues of innocence. As a result, such models remain vulnerable to guilt-biased narratives at inference time, where no explicit signal of acquittal or adversarial challenge is present.

2.2 Critique & Revise Frameworks with Multi-agent Systems

Recent "critique-and-revise" frameworks utilize multi-agent pipelines, commonly favoring a "divide-and-conquer" strategy for granular feedback (Madaan et al., 2023; Chen et al., 2025b; Xu et al., 2024a; Ferraz et al., 2024; Shinn et al., 2023). However, these systems face a major constraint due to their iterative nature. Previous studies indicate such pipelines may amplify the model's own biases (Xu et al., 2024b): rather than converging on accu-

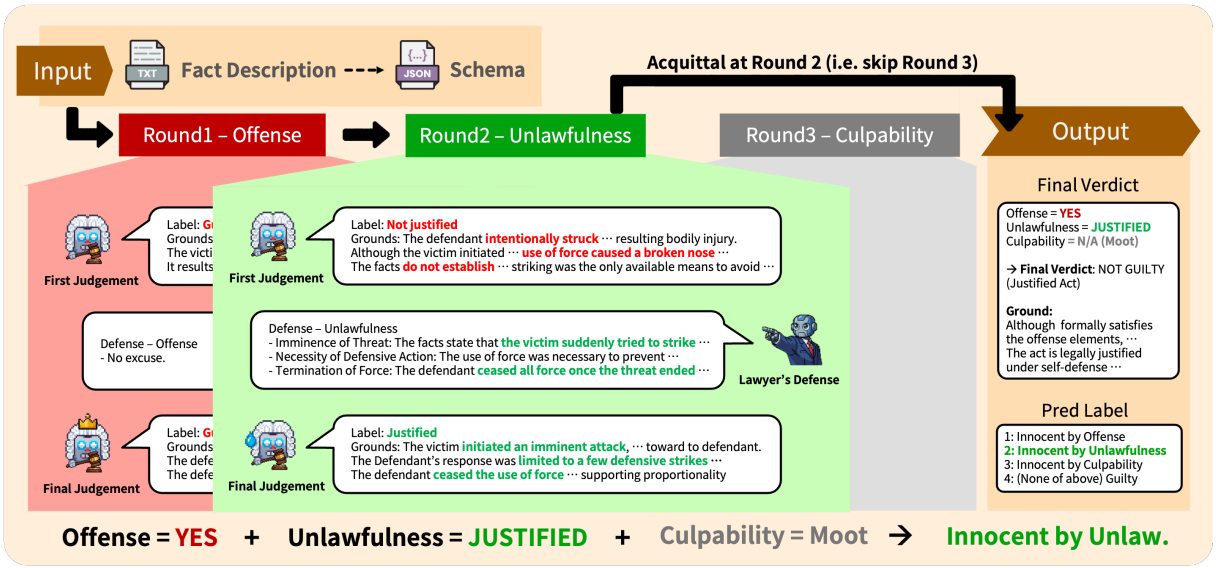

Figure 2: Detailed decision flow of the OBJECTION pipeline shown in Figure 1-(II), illustrated with a self-defense case. The Lawyer Agent overturns the initial biased judgment in Round 2 by highlighting exculpatory details (e.g., imminence), triggering the early-exit mechanism for acquittal. This adversarial intervention effectively neutralizes narrative bias, reducing the False Guilty Rate (FGR) from 82.93% to 16.69% on the Natural Innocent dataset.

racy, iterative refinement can reinforce the initial presumption of guilt.

This limitation is particularly critical in the legal domain. When the underlying LLM is already biased toward guilt, generic critique-and-revise or multi-agent debate systems (Chen et al., 2025c) tend to reproduce and amplify this bias through iterative agreement, forming an echo chamber that reinforces the presumption of guilt. While recent approaches like AgentCourt (Chen et al., 2025a) have introduced adversarial roles to simulate courtroom dynamics and improve legal question-answering, they focus primarily on civil domains and knowledge accumulation, leaving the fundamental issue of guilty bias in criminal judgments unaddressed.

In contrast, OBJECTION tackles this specific vulnerability by introducing a role-constrained adversarial Lawyer Agent whose sole objective is to construct the strongest possible defense. Unlike generic debaters or QA-oriented agents, our agent does not seek consensus or general knowledge refinement. Instead, it systematically enforces the presumption of innocence by actively injecting grounded, **reasonable doubts** at inference time. This effectively shatters the echo chamber, preventing bias amplification and enabling a genuine adversarial process that structurally corrects guilty bias; Appendix C walks through a full case. Crucially, the Lawyer Agent is injected at each stage rather than after a completed verdict, and our Normal Critic ablation shows the effect stems from role asymmetry, not from adding a second agent (Table 3).

3 Methodology

The OBJECTION framework addresses the inherent guilty bias in Legal LLMs by orchestrating three core components: (1) Schema-based Information Extraction (IE) (Feng et al., 2022; Liu et al., 2025), (2) Trichotomous Reasoning with Issue Isolation (Zhang et al., 2025a), and (3) Adversarial Argumentation. While existing strategies adopt IE or structural reasoning individually, they remain fragile against guilt-biased narratives. Specifically, LLMs tend to treat biased fact descriptions as neutral ground truths (X_{fact}), causing the reasoning process to collapse into a confirmation of guilt.

We propose a formal pipeline in which the Lawyer Agent serves as the orchestrator. Unlike a standard critic (Madaan et al., 2023; Shinn et al., 2023), the Lawyer Agent serves as a binding mechanism that connects individually bias-fragile modules.

We instantiate a single LLM into two distinct roles via domain-specific prompting: the Judge ($\mathcal{M}$) and the Lawyer ($\mathcal{A}$). Instead of the standard prediction $\hat{y} = \mathcal{M}(X_{fact})$, we model the inference as a 3-step adversarial interaction process:

$$\hat{y} = \overrightarrow{\prod_{k \in T}} \Psi_k \left(\mathcal{M}, \mathcal{A}, X_{fact}\right) \quad (1)$$

where $T = \{\text{offense}, \text{unlawfulness}, \text{culpability}\}$ represents each step in the trichotomous reasoning structure, the arrow over $\prod$ denotes an ordered execution, and Ψ_k denotes the sequential decision function refined by the adversarial intervention of $\mathcal{A}$ in step k. This design assures that the final verdict results from a legal debate rather than a unilateral confirmation of the prosecution's narrative.

3.1 Input and Schema-based IE

Legal narratives are highly unstructured. While prior event extraction studies (Feng et al., 2022; Liu et al., 2025) have made meaningful progress in structuring these narratives by decomposing them based on the Four Element Theory (Gao and Ma, 2019), applying these methods at scale often presents practical challenges. Because these approaches typically require defining specific rules for every charge, extending them across diverse crimes and legal systems requires substantial manual effort. Furthermore, while the four-element theory is fundamental in certain jurisdictions, other civil law countries (e.g., Germany, South Korea, Japan) heavily rely on the *Tatbestand* concept, which explicitly separates objective and subjective elements. To overcome the dependency on charge-specific priors and provide a generalized factual abstraction, we propose a SOAM schema that captures the fundamental structure of criminal acts:

- **Subject (S):** The actor or principal offender.
- **Object (O):** The target or victim of the act.
- **Actus Reus (A):** The strictly objective physical conduct itself and its consequences.
- **Mens Rea (M):** The subjective elements inferred from the context, rather than just explicit intent. We use the term broadly here.[2]

By abstracting facts into these four core components, our model performs extraction in a zero-shot manner (Prompt 1), effectively bypassing the need for granular annotation.

However, structure alone is not a panacea. As shown in our experiment (Table 1), relying solely on the schema actually exacerbates the False Guilty Rate (FGR). Forcing the model to explicitly evaluate detailed elements without an adversarial counterbalance often reinforces its inherent prosecutorial bias. Both the schema and the 3-step reasoning are individually "bias-fragile."

Therefore, we utilize this extraction exclusively in the *Offense* stage. While the schema alone may amplify bias, it provides a crucial *structured factual anchor* for the Lawyer Agent. Without this S-O-A-M breakdown, the agent struggles to pinpoint specific legal weaknesses in unstructured narratives. By synergizing these components with an adversarial counterbalance, OBJECTION converts them into a robust bias-mitigation pipeline, while the subsequent *Unlawfulness* and *Culpability* stages access the raw description to capture subtle contextual nuances.

3.2 Trichotomous Reasoning Framework

We implement the dogmatic three-stage structure of criminal law prevalent in civil law systems (Dubber, 2005). The reasoning process follows a sequential chain $T = \{\text{offense}, \text{unlawfulness}, \text{culpability}\}$ adopting the logical flow proposed in recent LJP studies (Zhang et al., 2025a):

(1) Offense: Does the act satisfy the constitutive elements of a crime?
(2) Unlawfulness: Are there justification grounds (e.g., self-defense) that negate unlawfulness?
(3) Culpability: Can the subject be held personally responsible? (e.g., insanity)

The core of this pipeline is contextual isolation. Each step is separated to ensure logical integrity. For instance, when evaluating unlawfulness, the model must proceed on the basis that the offense is established, independent of culpability factors.

3.3 Adversarial Lawyer Agent

The main contribution of this work is the introduction of the Adversarial Lawyer Agent, which transcends the role of a standard self-critique agent (Madaan et al., 2023; Xu et al., 2024b).

Difference from Standard Critics While a standard critic passively verifies logical consistency (asking *"Is the reasoning sound?"*), our Lawyer Agent actively shifts the burden of proof (similar to asking *"Is there any room for reasonable doubt?"*). Even if the prosecution's logic appears flawless on the surface, the Lawyer Agent is explicitly prompted to hypothesize exculpatory contexts (e.g., self-defense, intent negation), enforcing a rigorous adversarial review that generic critics cannot perform. In other words, while generic critics assess the validity of the model's logic, Lawyer

[2] We retain the term *Mens Rea* to maintain alignment with the traditional *Actus Reus* and *Mens Rea* dichotomy. Here, it serves as a broad umbrella term encompassing various subjective states required for different offenses, consistent with broader criminal law jurisprudence.

Agent actively searches for a loophole that could be a reasonable doubt of guilt.

Adversarial Interaction Design Formally, we define the internal mechanism of Ψ_k at each step as a dialectical interaction between $\mathcal{M}$ and $\mathcal{A}$:

$$j_k = \mathcal{M}(X_{fact}), \quad d_k = \mathcal{A}(j_k, X_{fact}) \quad (2)$$
$$\hat{y}_k = \mathcal{M}(j_k, d_k, X_{fact}) \quad (3)$$

First, $\mathcal{M}$ produces an initial judgment j_k. Second, $\mathcal{A}$ inspects j_k to generate a defense argument d_k (Eq. 2). Finally, this argument is injected back into the context, forcing $\mathcal{M}$ to re-evaluate the case to derive the final stage verdict $\hat{y}_k$ (Eq. 3).

This formulation highlights that the final decision is explicitly conditioned on the defense narrative, structurally blocking the model's inherent guilty bias. Our design of interaction systematically implements the *Exercise of Defense Rights* in criminal procedure.

3.4 Sequential Verdict Determination

The final verdict is determined by synthesizing the results of three stages. In accordance with the hierarchy of criminal law, the pipeline employs an Early Exit Mechanism: if a higher-level stage is negated (e.g., Offense is not established), the process terminates immediately with an acquittal, without evaluating subsequent stages. This ensures both computational efficiency and logical consistency by preventing error propagation.

Unlike simple binary classification, our system produces Granular Acquittal Labels (e.g., *Innocent by Offense*, *Unlawfulness*, *Culpability*). This granularity significantly improves the explainability of the judgment, clarifying the precise legal grounds for the verdict.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate our framework on three different datasets to verify both in-domain performance and cross-domain generalizability.

- **LJPIV-CAIL:** A benchmark dataset (Zhang et al., 2025a) containing criminal cases with synthetic innocent samples generated via counterfactual rewriting from the CAIL dataset (Xiao et al., 2018).
- **LJPIV-ELAM:** An out-of-domain dataset (Zhang et al., 2025a) based on ELAM (Yu et al., 2022), constructed via the same generation pipeline to evaluate cross-domain generalization.
- **Natural Innocent (Ours):** To resolve the limitations of synthetic innocence, we introduce a comprehensive dataset consisting of **3,412 real criminal judgments** from South Korean criminal courts. Detailed information can be found in Appendix B.

Baselines We categorize our baselines into two groups to identify the individual effects of prompting strategies and architectural choices. All inference experiments utilize `Qwen-2.5-7B-Instruct` (Yang et al., 2024), `Llama-3.1-8B-Instruct` (Dubey et al., 2024), and `Gemma-2-9B-it` (Gemma Team, 2024) as the backbones.

Prompting Baselines To investigate whether simply adding Schema or Structure mitigates bias, we designed four incremental prompting strategies.

- **(A) Plain LLM:** Performs direct zero-shot prediction using standard prompts.
- **(B) Schema-based:** Utilizes the extracted SOAM elements to focus on factual components.
- **(C) Naive Trichotomy:** Compels the model to follow the 3-step reasoning structure.
- **(D) Schema + Trichotomy:** Naively combines both the schema and the 3-step structure.

Main Baselines We compare OBJECTION against **LJPIV** (Zhang et al., 2025a) (fine-tuned SOTA, in-domain upper bound), **Self-Refine** (Madaan et al., 2023) (generic critic without legal persona), and **Debate-Feedback** (Chen et al., 2025c) (round-table debate via stochastic sampling).

Metrics We report **G-F1** and **NG-F1**, which represent the F1 scores for the Guilty and Not-Guilty classes, respectively. We also measure the macro Precision/Recall/F1 of the 3-step reasoning process. Crucially, beyond standard performance metrics, we introduce normative error metrics that directly reflect legal risk:

- **False Guilty Rate (FGR):** The ratio of innocent cases predicted as guilty ($FP/(TN + FP)$). A high FGR indicates a violation of the presumption of innocence.
- **False Not-Guilty Rate (FNR):** The ratio of guilty cases predicted as innocent ($FN/(TP + FN)$).

Table 1: Impact of adversarial intervention. (A)–(D) correspond to the existing methods in Fig. 1. The Naive combination **(D)** exhibits severe guilty bias despite high accuracy. Integrating the **Lawyer Agent** reduces FGR (53.21% → 10.71%) while improving reasoning.

Method	G-F1	NG-F1	Macro F1	FGR↓	FNR↓
(A) Plain LLM	0.37	0.58	0.48	29.64	70.36
(B) Schema	0.67	0.53	0.60	56.07	21.79
(C) Naive Tri.	0.31	0.63	0.47	17.86	78.21
(D) Schema+Tri.	**0.77**	0.62	0.70	53.21	**3.93**
(D) + Lawyer	0.75	**0.80**	**0.78**	**10.71**	33.93

4.2 Experiment 1: Limitations of Naive Prompting

Table 1 highlights the limitations of naive prompting strategies and the necessity of adversarial intervention. Approaches such as Schema-based extraction (B) and Naive Trichotomy (C) do not achieve a balance between accuracy and fairness. Even when combined (D), these methods produce logically structured yet significantly biased predictions, as indicated by an FGR of 53.21%.

In contrast, incorporating the Lawyer Agent into this pipeline ((D)+Lawyer) reduces the FGR to 10.71%. These results indicate that a structured reasoning framework alone is insufficient and active adversarial reasoning is essential to mitigate bias.

4.3 Experiment 2: Main Comparative Results

Table 2 compares OBJECTION with main baselines. We examine the results, focusing on mitigating guilty bias and cross-domain performance.

Baselines vs. Generalization Existing methods encounter problems with stability and generalization across different LLM architectures. LJPIV attains high in-domain accuracy but suffers from severe guilty bias on out-of-domain datasets such as ELAM (FGR 47.60%). Crucially, Debate-Feedback shows substantial volatility depending on the backbone model. For instance, it fails to effectively mitigate bias on Qwen-2.5 (e.g., CAIL FGR 22.14% vs. 10.71%), experiences uncontrolled over-acquittal on Llama-3.1 (e.g., 56.00% FNR on ELAM), and exhibits inconsistent bias control on Gemma-2. These fluctuations indicate that debate through stochastic sampling lacks a reliable steering mechanism. By comparison, OBJECTION demonstrates a consistent directional effect, reducing guilty bias on most backbones and achieving the highest overall accuracy (Macro-F1) across backbones; significance tests are reported in Appendix G.

Structural Integrity and Reasoning The superior generalization of OBJECTION is attributed to its rigorous adherence to legal procedure. As indicated in the "3-Step Reasoning" columns, this method maintains high element-wise detection scores (F1: 0.82 on CAIL, 0.81 on ELAM) across domains. In contrast, Baseline D fails on the ELAM dataset (F1 0.37), demonstrating that, without the Lawyer Agent's guidance, the model loses logical coherence.

Synthetic vs. Natural Innocence The most distinct performance disparity is observed in the Natural Innocent dataset. LJPIV performs poorly (FGR 82.93%) as a result of overfitting to synthetic innocent cues. Debate-Feedback also underperforms in this context; when using the Qwen backbone, it produces a high FGR of 42.95%, indicating a failure to protect innocent defendants. In contrast, OBJECTION exhibits greater robustness, achieving an FGR of 16.69% (Qwen). These results indicate that the Lawyer Agent effectively identifies reasonable doubt concealed within the narrative, enabling the detection of innocence without dependence on explicit shortcuts; a breakdown by acquittal ground is given in Appendix H.1.

4.4 Ablation Studies and Analysis

Ablation Study To validate the contribution of each component, we conducted an ablation study (Table 3).

- **Role of Lawyer Agent:** Removing the Lawyer Agent causes FGR to spike from 23.20% to 38.80%, while NG-F1 drops from 0.77 to 0.75. Notably, G-F1 rises to 0.83 in this setting, above the full pipeline. Given the guilty bias established in Section 4.2, the Lawyer Agent acts as a counterweight rather than a reasoning enhancer: it shifts where the decision boundary sits, making verdicts fairer rather than sharper.
- **Role of Structural Anchors:** Removing either the SOAM schema (factual anchor) or the 3-step structure (logical anchor) results in high FNRs (62.40% and 68.00%, respectively). Without these constraints to ground the debate, the Lawyer Agent becomes overly dominant and blindly argues for innocence, leading to excessive acquittals.
- **Ours w/ Normal Critic:** Replacing the Lawyer Agent with a generic critic yields a high guilty bias (FGR 56.40%). Without a spe-

Table 2: Main Comparative Results. **Bold** indicates the best performance within each group. The results show that OBJECTION achieves a consistent directional effect across diverse backbones. While **Debate-Feedback** exhibits severe volatility depending on the backbone architecture (e.g., uncontrolled over-acquittal on Llama), OBJECTION consistently mitigates guilty bias (low FGR) while preserving a solid conviction capability (Macro-F1). * denotes datasets whose innocent cases are synthetic, while † denotes datasets whose innocent cases are real-world acquittals.

Dataset	Backbone & Method	G-F1	NG-F1	Macro-F1	3-Step Reasoning			FGR (%)↓	FNR (%)↓
					Prec.	Recall	F1		
LJPIV-CAIL*	LJPIV (Fine-tuned)	0.80	0.83	0.82	0.62	0.71	0.61	9.29	27.14
	Qwen-2.5-7B								
	Baseline D	**0.77**	0.62	0.70	0.54	0.40	0.39	53.21	**3.93**
	Self-Refine	**0.77**	0.64	0.71	0.66	0.56	0.57	49.29	6.79
	Debate-Feedback	0.74	0.76	0.75	0.70	0.61	0.62	22.14	27.50
	OBJECTION (Ours)	0.75	**0.80**	**0.78**	**0.87**	**0.81**	**0.82**	**10.71**	33.93
	Llama-3.1-8B								
	Baseline D	**0.76**	0.56	0.66	**0.83**	0.57	0.58	60.36	**2.50**
	Self-Refine	0.69	0.44	0.57	0.74	0.59	0.60	68.93	10.36
	Debate-Feedback	0.58	**0.72**	0.65	0.60	0.61	0.60	**20.36**	52.86
	OBJECTION (Ours)	0.73	0.67	**0.70**	0.73	**0.69**	**0.69**	40.36	18.57
	Gemma-2-9B								
	Baseline D	0.77	0.59	0.68	0.76	0.58	0.60	57.50	1.07
	Self-Refine	0.77	0.57	0.67	**0.87**	0.62	**0.65**	60.00	**0.36**
	Debate-Feedback	0.81	0.80	0.81	0.71	0.62	0.63	**22.50**	16.43
	OBJECTION (Ours)	**0.84**	**0.80**	**0.82**	0.67	**0.64**	**0.65**	26.43	9.29
LJPIV-ELAM*	LJPIV (Fine-tuned)	0.80	0.68	0.74	0.87	0.66	0.71	47.60	2.40
	Qwen-2.5-7B								
	Baseline D	0.75	0.58	0.67	0.51	0.38	0.37	56.40	5.60
	Self-Refine	0.76	0.58	0.67	0.75	0.55	0.59	57.60	**3.20**
	Debate-Feedback	**0.80**	0.74	0.77	**0.87**	0.65	0.67	27.00	15.00
	OBJECTION (Ours)	0.78	**0.77**	**0.78**	0.82	**0.80**	**0.81**	**23.20**	21.60
	Llama-3.1-8B								
	Baseline D	**0.76**	0.53	0.65	0.52	0.54	0.52	64.00	**0.40**
	Self-Refine	0.66	0.33	0.50	0.67	0.55	0.54	77.60	12.00
	Debate-Feedback	0.61	**0.74**	**0.68**	0.56	0.54	0.54	**23.20**	56.00
	OBJECTION (Ours)	0.71	0.64	**0.68**	**0.73**	**0.69**	**0.67**	41.60	22.00
	Gemma-2-9B								
	Baseline D	0.75	0.51	0.63	0.81	0.56	0.58	66.00	**0.00**
	Self-Refine	0.73	0.41	0.57	**0.84**	0.54	0.56	74.40	**0.00**
	Debate-Feedback	0.80	0.72	0.76	0.73	0.63	0.67	38.40	9.60
	OBJECTION (Ours)	**0.83**	**0.77**	**0.80**	0.67	**0.68**	**0.68**	**36.60**	2.00
Natural Innocent†	LJPIV (Fine-tuned)	0.35	0.29	0.32	0.47	0.06	0.11	82.93	50.12
	Qwen-2.5-7B								
	Baseline D	0.55	0.27	0.41	0.48	0.34	0.21	84.49	9.99
	Self-Refine	0.57	0.23	0.40	0.54	0.36	0.24	86.28	**2.87**
	Debate-Feedback	**0.71**	0.73	0.72	0.47	0.48	0.44	42.95	5.34
	OBJECTION (Ours)	0.70	**0.82**	**0.76**	**0.57**	**0.48**	**0.47**	**16.69**	31.29
	Llama-3.1-8B								
	Baseline D	0.51	0.66	0.58	0.46	0.42	0.38	36.73	45.08
	Self-Refine	0.53	0.33	0.43	0.43	0.41	0.34	78.55	**16.89**
	Debate-Feedback	0.41	**0.78**	0.59	0.46	0.44	0.41	**7.69**	70.72
	OBJECTION (Ours)	**0.63**	0.65	**0.64**	**0.51**	**0.50**	**0.49**	46.39	19.05
	Gemma-2-9B								
	Baseline D	0.57	0.18	0.38	0.53	0.41	0.32	90.19	1.16
	Self-Refine	**0.60**	0.34	0.47	**0.62**	0.46	0.39	79.73	**0.15**
	Debate-Feedback	0.49	0.69	0.59	0.50	0.48	0.47	37.01	50.74
	OBJECTION (Ours)	**0.60**	**0.73**	**0.67**	0.52	**0.51**	**0.50**	**29.89**	36.10

cific mandate to establish reasonable doubt, LLMs act as compliant approvers of the input narrative rather than adversarial challengers.

- **Fine-tuning + Lawyer:** Adding our Lawyer Agent to the fine-tuned LJPIV leads to failure (FNR 99.60%). Since LJPIV has already learned to detect synthetic innocence, the additional adversarial intervention amplifies doubt excessively, creating an "echo chamber" (Xu et al., 2024b) that leads to indiscriminate acquittals.

Distribution Analysis We analyzed the prediction distribution (Figure 3). Baselines like Baseline D and Self-Refine skew heavily towards the "Guilty" class (Gray bars), diverging significantly from the GOLD standard. In contrast, OBJECTION aligns most closely with the GOLD distribution across all datasets. This indicates that our model does not merely "guess" innocence to lower

Table 3: Ablation Results (w/ `Qwen 2.5`) on ELAM dataset. The Lawyer Agent is effective in a low False Guilty Rate (FGR). Similar ablation trends are observed on the CAIL dataset where LJPIV was originally trained; see Appendix F for details.

Configuration	G-F1	NG-F1	FGR↓	FNR↓
OBJECTION (Full)	0.78	**0.77**	23.20	21.60
(-) Lawyer Agent	**0.83**	0.75	38.80	**2.00**
(-) Schema	0.52	0.72	**8.00**	62.40
(-) 3-Step Structure	0.38	0.56	34.80	68.00
Ours w/ Normal Critic	0.77	0.60	56.40	**2.00**
LJPIV	0.80	0.68	47.60	2.40
LJPIV + Lawyer Agent	0.01	0.67	0.40	99.60

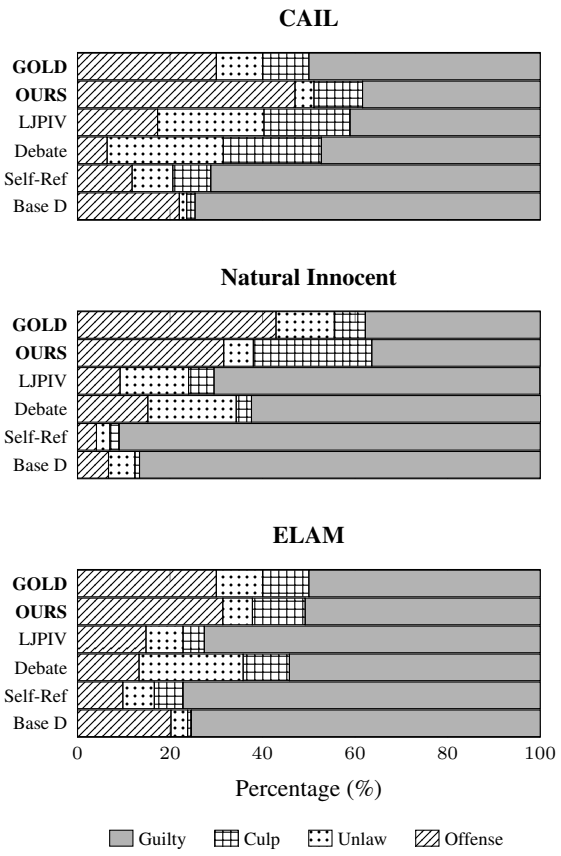

Figure 3: Prediction distributions across datasets. Baselines over-predict the "Guilty" class (gray bars), whereas OBJECTION closely matches the GOLD standard by accurately identifying specific grounds for acquittal.

FGR. Instead, it accurately maintains necessary guilty convictions while correctly classifying the specific legal grounds for acquittal (Offense vs. Unlawfulness vs. Culpability), demonstrating a highly calibrated understanding of legal liability.

Generalization to Commercial APIs To assess scalability, OBJECTION was applied to commercial APIs (`GPT-4.1-mini`, `Gemini-2.5-Flash`). As presented in Table 4, the framework consistently

Table 4: Robustness on Commercial APIs compared to baselines. While Debate-Feedback offers moderate bias reduction, OBJECTION consistently attains fairness (low FGR) and reasoning quality (high 3-Step F1) across both GPT-4.1 and Gemini architectures.

Model	Method	CAIL		Natural	
		FGR↓	3-S F1↑	FGR↓	3-S F1↑
GPT-4.1 mini	Baseline D	40.36	0.54	59.71	0.08
	Debate.	30.84	0.66	**54.92**	0.22
	OBJECTION	**21.43**	**0.75**	58.46	**0.34**
Gemini 2.5-Flash	Baseline D	51.07	0.56	76.75	0.03
	Debate.	42.63	0.63	68.41	0.17
	OBJECTION	**36.07**	**0.65**	**61.95**	**0.33**

reduces guilty bias on the CAIL dataset, lowering the FGR to 21.43%. Furthermore, on the Natural Innocent benchmark, OBJECTION achieves a substantial improvement in 3-Step Reasoning F1 (0.34), significantly surpassing the baseline (0.08). These findings show that the pipeline effectively integrates the three-stage judgment logic of criminal law into general-purpose large language models, ensuring that acquittals are grounded in valid legal reasoning.

Table 5: Controllability via Persona Design. Unlike uncontrolled baselines, OBJECTION allows explicit steering through distinct lawyer variants. Our experiments reveal a trade-off between fairness and accuracy, demonstrating that OBJECTION can actively steer this trade-off by modulating defense intensity.

Method	CAIL		ELAM		Natural	
	G-F1	FGR	G-F1	FGR	NG-F1	FGR
LJPIV	0.80	9.29	0.80	47.60	0.29	82.93
Baseline D	0.77	53.21	0.75	56.40	0.27	84.49
Debate.	0.74	22.14	0.80	27.00	0.73	42.95
OBJECTION	0.75	**10.71**	0.78	**23.20**	**0.82**	**16.69**
OBJECTION-Soft	**0.83**	31.07	**0.84**	31.60	0.58	58.84

4.5 Controllability: The "Soft" Variant

A key advantage of our framework is the ability to explicitly control the model's behavior by modulating the Lawyer Agent's intensity. This stands in sharp contrast to debate-based methods (Chen et al., 2025c), where response diversity relies on stochastic sampling with identical prompts, lacking a mechanism to systematically control the "degree of doubt". By designing a softened persona (OBJECTION-Soft), we provide a deterministic "steering knob" that allows users to control the trade-off between conviction accuracy and the presumption of innocence.

Table 5 provides empirical validation of this steerability. OBJECTION-Soft attains higher conviction accuracy (G-F1 0.83 on CAIL) compared to the standard version (0.75), but this improvement is accompanied by an increased False Guilty Rate (31.07%). Additionally, the Soft variant's decline in performance on the Natural Innocent benchmark (FGR 58.84%) indicates that real-world acquittals involving subtle implicit cues require stronger adversarial intervention, emphasizing the need for a tunable framework for domain-adaptive deployment. This FGR–FNR asymmetry is deliberate. A false conviction is irreversible whereas a false acquittal is correctable (*in dubio pro reo*), and across our settings the reduction in FGR is substantially larger than the accompanying rise in FNR.

4.6 Human Expert Evaluation

To rigorously assess the practical applicability and legal validity of the generated defense arguments, we conducted a qualitative evaluation with human experts. The evaluation panel consisted of active law enforcement investigators with extensive experience in criminal investigations.

Evaluation Setup We randomly sampled 50 cases from the Natural Innocent dataset. The experts blindly evaluated the defense arguments generated by our Lawyer Agent and the debater agents in the Debate-Feedback pipeline. Each argument was scored on a 5-point Likert scale (1: Poor, 5: Excellent) across four criteria:

- **Legal Relevance (Rel.):** Legal validity and appropriateness.
- **Argument Strength (Str.):** Persuasiveness and logical coherence.
- **Factual Grounding (Grd.):** Strict anchoring to the provided input facts.
- **Human Similarity (Sim.):** Resemblance to a professional legal opinion.

Analysis and Insights As shown in Table 6 (A), the generic Debate-Feedback method fails to construct legally sound arguments, averaging a score of 2.87. It particularly struggles to mimic professional legal reasoning (Human Similarity: 2.32). In contrast, OBJECTION achieves an expert-level average of 4.42.

Furthermore, the step-wise analysis in (B) highlights that our Lawyer Agent excels particularly in the Unlawfulness stage (Avg 4.57, Grd 5.00), demonstrating exceptional capability in verifying strict legal requirements and establishing rigorous factual grounding without hallucination. Factual Grounding is also the highest-scoring criterion overall (4.86), indicating that the arguments stay anchored to the given facts rather than inventing exculpatory detail; the panel penalized any claim not traceable to the input. Grounding errors are nonetheless possible, and retrieval-augmented grounding over statutes and precedent is a natural mitigation we leave to future work.

Table 6: **Human Expert Evaluation Results** ($N = 50$)**.** (A) Our Lawyer Agent outperforms the baseline across all metrics. (B) Step-wise analysis reveals that the agent performs best in the Unlawfulness stage (Avg 4.57), verifying its strong logical reasoning capability.

(A) Comparative Analysis (Baseline vs. Ours)					
Model	**Rel.**	**Str.**	**Grd.**	**Sim.**	**Avg**
Debate-Feedback	2.64	2.88	3.64	2.32	2.87
OBJECTION	**4.16**	**4.52**	**4.86**	**4.12**	**4.42**
(B) Step-wise Analysis (Ours: Lawyer Agent)					
Step	**Rel.**	**Str.**	**Grd.**	**Sim.**	**Avg**
Offense	**4.20**	4.45	4.85	**4.40**	4.48
Unlawfulness	**4.20**	**4.87**	**5.00**	4.20	**4.57**
Culpability	4.07	4.27	4.73	3.67	4.19

5 Conclusion

We identified that the Guilty Bias stemming from prosecutorial narratives is a critical point in Legal NLP. To address this, we proposed OBJECTION, an inference-time adversarial pipeline. By combining a Lawyer Agent into the 3-step reasoning structure, our framework actively challenges the model's presumption of guilt.

Empirically, OBJECTION reduced the False Guilty Rate (FGR) on the Natural Innocent dataset, where fine-tuned SOTA models failed, while being robust to out-of-domain datasets without additional training.

Ultimately, this work implements the Presumption of Innocence at the system architecture level. Our pipeline favors preventing false convictions over raw accuracy. Nevertheless, OBJECTION attains the highest or tied-highest Macro-F1 in every backbone–dataset setting, maintaining a robust balance between guilty and innocent predictions. We present OBJECTION not as a simple technical improvement, but as a step toward normative alignment in Legal AI, assuring that digital judgments adhere to the basic principles of criminal justice.

6 Limitations

First, our pipeline is designed within the framework of the Civil Law system. Adaptation to Common Law systems, which focus on case law, demands further study. However, given that the adversarial principle (prosecution vs. defense) is a universal tenet of criminal justice, we argue that our Lawyer Agent's core mechanism—challenging narrative bias—remains valid across legal systems.

Second, our current Information Extraction (IE) module uses a generic SOAM (Subject-Object-Act-Mens Rea) schema to ensure scalability. Developing jurisdiction-specific schemas that represent the granular constitutive elements of individual crimes could further improve performance.

Third, introducing an adversarial agent inherently increases inference cost compared to single-pass baselines. However, our early exit mechanism mitigates this by terminating the process immediately upon finding clear innocence, achieving a lower expected token consumption than exhaustive debate methods (e.g., Debate-Feedback); see Appendix E for more details.

Finally, we address only the binary guilty / not-guilty decision. Charge and sentencing prediction should exhibit the same prosecutorial bias, and our framework extends naturally—offense-stage arguments bear on charge severity, culpability-stage arguments on sentencing—which we leave as our primary follow-up.

7 Ethical Considerations

The power of criminal justice, including the judgment of guilt, is constitutionally delegated exclusively to authorized human experts and public officials. This study does not advocate for AI to usurp this sovereign power or make autonomous verdicts. Instead, we explore the potential of AI as a Decision Support System to assist legal professionals in detecting overlooked innocence and ensuring procedural fairness. Licenses for all external resources are listed in Appendix I.

Concerning data privacy, all "Natural Innocent" judgments used in this study were collected from publicly available records released by the Supreme Court of Korea and have been strictly de-identified and anonymized to protect the privacy of all individuals involved. See Appendix B for detailed information.

Regarding the human expert evaluation detailed in Section 4.6, the protocol involved active law enforcement investigators assessing the legal validity of generated arguments. This evaluation was conducted as a minimal-risk professional task assessment. It did not involve the collection of sensitive personal information from the participants, nor did it pose any physical or psychological risks. Residual failure modes of the pipeline, which motivate keeping a human in the loop, are analyzed in Appendix H.2.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant (RS-2023-00280883, RS-2023-00222663); by the National Research Foundation, Korea, under project BK21 FOUR(Dept. of Data Science, SNU, No. 5199990914569); by the Korea Institute of Science and Technology Information (KISTI) in 2026 (No. (KISTI)K26L3M1C1), aimed at developing KONI (KISTI Open Neural Intelligence), a large language model specialized in science and technology; and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2025-02263754, Human-Centric Embodied AI Agents with Autonomous Decision-Making); by grant (25202MFDS003) from Ministry of Food and Drug Safety in 2025; by AI-BIO Research Grant through Seoul National University; Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2025-25442149, LG AI STAR Talent Development Program for Leading Large-Scale Generative AI Models in the Physical AI Domain).

References

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Hamid Alinejad-Rokny, Shiwen Ni, and Min Yang. 2025a. AgentCourt: Simulating court with adversarial evolvable lawyer agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5850–5865, Vienna, Austria. Association for Computational Linguistics.

Justin Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2025b. MAgICoRe: Multi-agent, iterative, coarse-to-fine refinement for reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32663–32686, Suzhou, China. Association for Computational Linguistics.

Xi Chen, Mao Mao, Shuo Li, and Haotian Shangguan. 2025c. Debate-feedback: A multi-agent framework for efficient legal judgment prediction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 462–470, Albuquerque, New Mexico. Association for Computational Linguistics.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 11:102050–102071.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009, Singapore. Association for Computational Linguistics.

Markus Dirk Dubber. 2005. Theories of crime and punishment in german criminal law. *The American journal of comparative law*, 53(3):679–707.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yi Feng, Chuanyi Li, and Vincent Ng. 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)*, pages 648–664, Dublin, Ireland. Association for Computational Linguistics.

Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, and Nanyun Peng. 2024. LLM self-correction with DeCRIM: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7773–7812, Miami, Florida, USA. Association for Computational Linguistics.

Mingxuan Gao and Kechang Ma. 2019. *Criminal Law (Ninth Edition)*. Peking University Press, Beijing. ISBN: 9787301307120.

Gemma Team. 2024. Gemma.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 417–421, New York, NY, USA. Association for Computing Machinery.

Huanghai Liu, Quzhe Huang, Qingjing Chen, Yiran Hu, Jiayu Ma, Yun Liu, Weixing Shen, and Yansong Feng. 2025. JUREX-4E: Juridical expert-annotated four-element knowledge base for legal reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3794–3814, Suzhou, China. Association for Computational Linguistics.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736, Copenhagen, Denmark. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *Preprint*, arXiv:1807.02478.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095, Online. Association for Computational Linguistics.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024a. LLMRefine: Pinpointing and refining large language models via fine-grained actionable feedback. In *Findings*

of the Association for Computational Linguistics: NAACL 2024, pages 1429–1445, Mexico City, Mexico. Association for Computational Linguistics.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024b. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2.5: A party of foundation models.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Weijie Yu, Zhongxiang Sun, Jun Xu, Zhenhua Dong, Xu Chen, Hongteng Xu, and Ji-Rong Wen. 2022. Explainable legal case matching via inverse optimal transport-based rationale extraction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 657–668, New York, NY, USA. Association for Computing Machinery.

Kepu Zhang, Haoyue Yang, Xu Tang, Weijie Yu, and Jun Xu. 2025a. Beyond guilt: Legal judgment prediction with trichotomous reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1815–1826, Suzhou, China. Association for Computational Linguistics.

Kepu Zhang, Weijie Yu, Zhongxiang Sun, and Jun Xu. 2025b. Syler: A framework for explicit syllogistic legal reasoning in large language models. CIKM '25, page 4117–4127, New York, NY, USA. Association for Computing Machinery.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

A Prompt Templates

Prompt 1: SOAM Schema Extraction

```
[System]
You are a legal information extraction
assistant. Your task is to extract SOAM
elements only from the given facts. Do
not fabricate or infer missing information.
Output a strict JSON object only, without
any other text.

[User]
<ROLE>
You are a SOAM element extractor.
</ROLE>
<RULES>
1) Extract only from the facts; if missing,
fill with an empty string.
2) Each slot must contain two fields: value
and evidence (or just value).
3) Output only a JSON object, the top-level
keys must be: Act, MensRea, Object, Subject
</RULES>
<INPUT>
{case_text}
</INPUT>
<OUTPUT_JSON_EXAMPLE>
{{
"Act": {{"value": "...", "evidence":
"..."}},
"MensRea": {{"value": "...", "evidence":
"..."}},
"Object": {{"value": "...", "evidence":
"..."}},
"Subject": {{"value": "...", "evidence":
"..."}}
}}
</OUTPUT_JSON_EXAMPLE>
```

SOAM Extraction Prompt Prompt 1 shows the zero-shot template used to extract the Subject, Object, Act, and Mens Rea elements from the case description, together with the supporting evidence span for each element.

Prompt 2: Initial Judgement Prompt

```
[System]
You are a criminal law expert. Please
strictly output the specified LJP block
as required. Do not output article
numbers, case numbers, or superfluous text
irrelevant to the format.

[User]
<CASE_FACTS>
{case_text}
</CASE_FACTS>
{ie_block}
<QUESTION>
Based solely on the facts above and IE
elements, please determine whether the
conduct meets the Constituent Elements of
the crime (Offense).
[Scope of Judgment]
```

```
1.    **Determine  Only**:   Whether  the
conduct meets the objective and subjective
constituent elements defined by criminal
law.
2. **Strictly Forbidden**: Considering any
justification grounds (e.g., Self-Defense)
or excuse grounds (e.g., Insanity).
3.  Even if justification/excuse grounds
exist, if the constituent elements are met,
you must label **YES** for this stage.
</QUESTION>
[Offense]
- label: YES or NO
- grounds:
1) ...
2) ...
```

Initial Judgment Prompt Prompt 2 shows the prompt template used for the Initial Judgment (Judge 1). This template is applied across the Offense, Unlawfulness, and Culpability stages. The *Question* block explicitly adapts its instructions for each stage: Checking constituent elements for Offense, evaluating justification grounds (e.g., self-defense) for Unlawfulness, and determining the responsibility capacity (e.g., insanity) for Culpability, ensuring that the model focuses strictly on the legally relevant criteria of the current stage.

Prompt 3: Defense(Offense)

```
[User]
{case_text}
</CASE_FACTS>
{ie_block}
[Offense Initial Judgment]
{judge_block}
<QUESTION>
You are a Defense Lawyer. Please read the
[Offense Initial Judgment] and write a
defense opinion for the defendant.
Goal: Find mitigating circumstances such
as "Constituent Elements Not Met (Offense
NO)" or "Attempt/Discontinuation".
Strategy (In Dubio Pro Reo):
1.  **Challenge Elements**: Raise factual
or legal doubts regarding the nature of
the act, causation, or subjective elements
(Intent/Negligence).
2.   **Attempt & Discontinuation**:  If
the act was not completed or voluntarily
abandoned, point this out.
3. **Targeted Rebuttal**: You must directly
cite original text from the case facts
(Evidence) to support your rebuttal.
[Stage Scope]
1.  **Defend Only**: Issues regarding the
applicability of Constituent Elements.
2.   **Strictly Forbidden**:  Mentioning
Self-Defense, Necessity, Insanity, Age,
etc., which belong to subsequent stages.
3.  If the case only involves the above
subsequent issues, output "No Objection"
for this stage.
Please    output    "Defense    Opinion
(Defense-Offense)".  If the prosecution's
judgment is unassailable (or points belong
to later stages), output "No Objection".
</QUESTION>
[Defense-Offense]
- point1: ...
```

Defense at the Offense Stage Prompt 3 instructs the Lawyer to challenge the establishment of constituent elements or argue for attempt, discontinuation, or reasonable doubt under *In Dubio Pro Reo*, while sharing the same System Prompt as the Initial Judge.

Prompt 4: Defense(Unlawfulness)

```
[User]
<CASE_FACTS>
{case_text}
</CASE_FACTS>
{ie_block}
[Illegality Initial Judgment]
{judge_block}
<QUESTION>
You are a Defense Lawyer. Please read the
[Illegality Initial Judgment] and write a
defense opinion for the defendant.
Goal:  Argue for Self-Defense,  Necessity,
or other Justification grounds.
Strategy:
1.   **Exploit Favorable Circumstances**:
Focus  on  descriptions  like  "Victim's
fault", "Imminent danger", "Unavoidable".
2.   **Challenge Strict Standards**:  If
the prosecution sets the bar too high for
Defense/Necessity, argue that the urgency
of the situation should be considered.
[Stage Scope]
1. **Defend Only**: Justification grounds
such as Self-Defense, Necessity, Justified
Act.
2.   **Strictly Forbidden**:  Issues of
Responsibility  such  as  Insanity,  Age
(leave for Culpability).
3. If the case involves only responsibility
issues,  output  "No  Objection"  for  this
stage.
Please    output    "Defense    Opinion
(Defense-Illegality)".    If  there  are
clearly no justification grounds (or they
belong  to  responsibility  issues),  output
"No Objection".
</QUESTION>
[Defense-Illegality]
- point1: ...
```

Defense at the Unlawfulness Stage Prompt 4 directs the Lawyer to identify justification grounds (e.g., self-defense or necessity) and to argue for justification even in legally ambiguous situations.

Prompt 5: Defense(Culpability)

```
[User]
<CASE_FACTS>
{case_text}
</CASE_FACTS>
{ie_block}
[Culpability Initial Judgment]
{judge_block}
<QUESTION>
You are a Defense Lawyer. Please read the
[Culpability Initial Judgment] and write a
defense opinion for the defendant.
Goal: Argue for lack of Responsibility
Capacity (e.g., Insanity, Minor) or lack
of Expectability.
Strategy:
1. **Highlight Special States**: Focus
on Age, History of Mental Illness,
Intoxication (if pathological), Duress,
etc.
2. **Humanitarian Perspective**: If the
defendant was in a desperate situation,
argue the inevitability of their conduct.
[Stage Scope]
1. **Defend Only**: Criminal Responsibility
Capacity (Insanity, Age, Expectability).
Please output "Defense Opinion
(Defense-Culp)". If there are no relevant
circumstances, output "No Objection".
</QUESTION>
[Defense-Culp]
- point1: ...
```

Defense at the Culpability Stage Prompt 5 prompts the Lawyer to argue for a lack of responsibility by highlighting special states such as insanity, minor age, or lack of expectability.

Prompt 6: Final Judgement Aggregation

```
[System]
You are a criminal law expert with 15 years
of experience. You should make strict,
restrained, and verifiable judgments based
on facts.
You will see the Prosecution's Initial
Judgment and the Defense Points. Adopt only
those defense points that are consistent
with the facts, and provide a final
judgment.
<CORE_RULES>
1) Judge based ONLY on the facts appearing
in <CASE_FACTS>.
2) Adopt ONLY defense points consistent
with the facts; ignore any points beyond
the facts.
3) Output must strictly follow the
specified format: label + grounds.
</CORE_RULES>


[User]
<CASE_FACTS>
{case_text}
</CASE_FACTS>
{ie_block}
[Prosecution - Initial Judgment]
{judge_initial}
[Defense Opinion]
{lawyer_defense}
<QUESTION>
Adopt only defense points consistent with
the facts, and provide the final judgment
for {block_name}.
</QUESTION>
[{out_block}] # stage name
- label: {label_line}
- grounds:
1) ...
```

Final Judgment Aggregation Prompt 6 aggregates the Initial Judgment and the Lawyer's defense arguments and applies rigorous fact-checking to produce the final verdict for the current stage.

B About "Natural Innocent" Dataset

B.1 Construction Pipeline

The **Natural Innocent** dataset is constructed from criminal judgments publicly released by the South Korean government, specifically from the *Korean Law Information Center* and the courts. All government judgments undergo a strict anonymization process to protect personal privacy.

To ensure label integrity, Class 1 (Not Guilty) contains only cases with a final verdict of **Not Guilty (Acquittal)**, while Class 0 (Guilty) contains cases with a final guilty verdict, added to enable guilty-side metrics and balanced evaluation. Furthermore, as securing the detailed description of facts (prosecutorial charges) is central to our input construction, we restricted our scope to **first-instance criminal judgments**. In the South Korean legal system, the first instance serves as the fact-finding trial (*Tatfrage*), whereas the second and third instances focus on legal interpretation (*Rechtsfrage*). Therefore, first-instance judgments provide the most in-depth description of factual relationships and the court's reasoning regarding the charges.

From the pool of first-instance acquittals, we filtered cases using **keyword matching** and categorized them into the three stages of the trichotomous framework. We identified whether the court's reasoning contained specific keywords or adjudicated specific legal issues relevant to each stage:

- **Offense-level Acquittals:** We primarily targeted the idiomatic judicial phrase "*lack of proof of crime*" (insufficiency of evidence).

We also collected cases based on specific contentious issues inherent to individual crimes, such as distinguishing fraud from simple **debt default**, the **public interest** aspect in defamation, or determining whether failing to return **mistakenly remitted funds** constitutes embezzlement.

- **Unlawfulness-level Acquittals:** We collected cases containing keywords related to justification grounds, including "*Excessive Defense*," "*Emergency Evacuation*," "*Justifiable Act*," "*Self-Defense*," and "*Victim's Consent*."

- **Culpability-level Acquittals:** We used keywords related to responsibility and attribution, such as "*Theory of Expectability*," "*Capacity to Discern*," "*Criminal Responsibility*," and "*Mistake of Law*."

The specific keywords and their distribution are detailed in Table 7. All resulting stage assignments were verified by a police investigator.

Category	Keyword (Translation)	Count
Offense	Debt Default (Fraud)	192
	Lack of Proof	189
	Property of Others	187
	Public Interest (Defamation)	186
	Negation of Causality	151
	Simple Debt Default	106
	Return (Embezzlement)	102
	Potential Loss	92
	Custodian Status	79
	Pecuniary Loss	68
Unlawfulness	Not against Social Rules	175
	Self-Defense	118
	Victim's Consent	51
	Justifiable Act	49
	Emergency Evacuation	20
	Excessive Defense	10
	Act due to Duty	8
Culpability	No Intent	87
	Mistake of Justification	29
	Capacity to Decide	21
	Awareness of Illegality	20
	Capacity to Discern	20
	Lawful Act (Expectability)	19
	Expectability	15
	Mistake of Law	5
	Medical Treatment	5

Table 7: Detailed statistics of keyword-based filtering for the Natural Innocent dataset. Keywords are categorized into the three stages of the trichotomous reasoning framework. Top keywords only for Offense and Culpability.

B.2 Overview & Statistics

This section summarizes the dataset: its provenance and scale (Table 8), the distribution of verdict types under the trichotomous framework (Table 9), the ten most frequent charges (Table 10), and the token length of case descriptions (Table 11).

Item	Description
Source	Real-world Korean criminal judgments
Verdict Type	Guilty & Not Guilty (Acquittal)
Total Cases	3,412
Jurisdiction	South Korea
Crime Domains	Fraud, Violence, Narcotics, etc.
Purpose	Balanced FGR/FNR evaluation

Table 8: Overview of the Natural Innocent dataset.

Class	Legal Ground	Count	Ratio (%)
Class 0	Guilty	1,291	37.8
Class 1	Offense Not Established	1,464	42.9
Class 2	Justification	431	12.6
Class 3	No Culpability	226	6.6
Total		3,412	100.0

Table 9: Distribution of verdict types under the trichotomous framework (Offense–Unlawfulness–Culpability).

Rank	Charge	Count	Ratio (%)
1	Fraud	657	19.3
2	Assault	249	7.3
3	Aggravated Fraud	241	7.1
4	Embezzlement	163	4.8
5	Agg. Assault / Injury	129	3.8
6	Occupational Embezzlement	111	3.3
7	Property Damage	110	3.2
8	Injury	109	3.2
9	Defamation	100	2.9
10	Cyber Fraud	98	2.9

Table 10: Top-10 offense distribution in the Natural Innocent dataset. All charges are normalized to offense-level categories and reported in English.

Statistic	Tokens
Tokenizer	Qwen-2.5-7B
Mean	781
Median	575
25th Percentile	301
75th Percentile	1,030
IQR	301 – 1,030

Table 11: Token-length statistics of case descriptions in the Natural Innocent dataset.

B.3 Synthetic vs. Natural Innocent

Figure 4 shows the difference between Synthetic Innocent and Natural Innocent.

On the morning of September 1, 2013, around 9:00 AM, the defendant Chang quarreled with his wife, Victim Chen, in their rental room because he suspected her of having an affair.

Chang noticed Chen showing signs of suicidal intent. **For the purpose of preventing her from jumping off the building, he choked her by the neck** and pinned her onto the bed attempting to stop her.

He continued choking her until she stopped reacting, then let go. He then **locked the door** and left the rental room, finding her unresponsive but thinking she had fallen asleep, **and fearing someone else might enter and hurt her.**

Around 10:00 AM that day, Chang returned to the room and found her having difficulty breathing. He immediately called for an ambulance and tried CPR, then called the police for help, but ultimately failed to save her life.

(A) Synthetic Innocent

The defendant and the victim traveled together in a car driven by the defendant. While in the car, they began arguing for unknown reasons, and the defendant stopped the car on the road near A City Hall.

Around 04:50 on June 11, 2011, after the car stopped at the victim's request, the victim kicked the passenger side door while getting out. When the defendant asked, "Why are you kicking the door?", the victim, who was physically stronger and had a more aggressive nature, assaulted the defendant first by punching him in the right eye and chest area.

In response, to restrain the victim while being assaulted, the defendant grappled with him. **In the process of falling together, he caused the victim to fall to the ground.**

As a result, the left side of the victim's head struck the ground directly, causing a fatal impact. The victim was transported to a hospital but passed away days later due to brain hemorrhages caused by the head trauma.

(B) Natural Innocent

Figure 4: Comparison of Synthetic vs. Natural Innocent narratives. **(Left)** Synthetic cases usually contain forced, **implausible scenarios with explicit exculpatory cues**, risking shortcut learning. **(Right)** Natural Innocent cases present realistic, logical narratives where innocence must be *inferred* from implicit contextual details (e.g., incidental injury during self-defense), requiring genuine legal reasoning.

Synthetic Innocent (Left) The synthetic innocent case shows clear evidence of coerced counterfactual rewriting. Although preventing a suicidal act could plausibly justify physical restraint, the narrative escalates into implausible behavior—continuing to choke until the victim becomes unresponsive and then locking the door out of fear that someone else might enter. These actions break common-sense causal and behavioral expectations and read as overtly injected exculpatory cues rather than organically arising facts. Such unnatural constructions make innocence highly salient and explicit, which risks shortcut learning during fine-tuning: models may learn to associate conspicuous, unrealistic signals with acquittal rather than develop genuine legal reasoning.

Natural Innocent (Right) The natural innocent case presents a temporally coherent, realistic narrative in which innocence must be inferred rather than stated. The victim initiates the assault; the defendant attempts to restrain him; and the fatal injury occurs incidentally as both fall, without any explicit claim about intent. Exculpatory clues—self-defensive context, proportional response, and accidental outcome—are embedded implicitly in the sequence of events. This forces models to integrate context and causality to infer non-culpability, making the case substantially harder and more faithful to real-world acquittals than synthetic innocence with explicit cues.

C Detailed Case Study (Qualitative Analysis)

Input Facts. Figure 5 shows the input fact description of a real-world criminal case. Following a gambling-related dispute, the victim returned to the defendant's residence armed with a knife and explicitly threatened to kill him. The victim initiated the attack and stabbed the defendant, after which the knife was removed, and the defendant attempted to restrain the victim while repeatedly calling the police. Despite being subdued, the victim continued to struggle and issue death threats, and later died due to neck compression while being restrained until police arrival.

SOAM Schema Extraction. To provide operational clarity on our schema and address potential conceptual ambiguity, we demonstrate the step-by-step mapping of the narrative shown in Figure 5. Our extraction module maps the unstructured facts into the following elements:

- **Subject** (S)**:** The defendant.
- **Object** (O)**:** Victim C.
- **Actus Reus** (A)**:** The defendant "grabbed both of the victim's hands, knocked him to the floor, and continued to press the victim's neck with his knee for about 10 minutes." (This captures strictly the objective physical conduct and the resulting death).
- **Mens Rea** (M)**:** The defendant "called 112

From around 14:00 on November 3, 2019, to around 01:00 the next day, the defendant engaged in gambling, specifically 'Go-Stop', dozens of times at his home in Seogwipo-si B for the purpose of spending time or socializing. Victim C (male, 76 years old), who lost money playing cards as mentioned above, became enraged, argued with D and others, went to his own home, and then returned to the defendant's living room around 02:26 on November 4, 2019, carrying a knife (total length 19.5 cm, blade length 9 cm).

He shouted, "Since D isn't here, I'll kill you [the defendant] instead," charged at the defendant threatening to kill him, and thrust the knife at the defendant's abdomen, inflicting a superficial laceration and bruising of about 2 cm on the defendant's abdomen and a superficial laceration of about 3 cm on his left elbow.

When the victim returned with the knife and thrust it at the defendant's abdomen, someone grabbed the victim's hand, took the knife away, and threw it onto the sofa. The defendant then grabbed both of the victim's hands, knocked him to the floor, and while pressing his knee against the neck of the victim lying on the floor, called 112 around 02:26 stating, "An acquaintance is making a disturbance with a knife, please come quickly." Six minutes later, around 02:32, he reported again stating, "It's been a while since I reported, why haven't you come yet? Please come quickly."

As the victim continued to struggle and threaten to kill the defendant even after being subdued, the defendant continued to press the victim's neck with his knee for about 10 minutes until the police arrived at the scene around 02:37 on the same day, causing the victim to lose consciousness on the spot. The victim died of asphyxiation due to neck compression while receiving treatment at the I emergency room around 04:46 on the same day.

(A) Input Fact Description

Figure 5: Input fact description of the case involving an armed attack and subsequent restraint.

> around 02:26... [and] reported again stating, 'It's been a while since I reported, why haven't you come yet?'." (This captures the subjective state: the intent was to restrain the attacker until help arrived, not a malicious intent to kill).

In real-world criminal adjudications, *Mens Rea* is rarely stated explicitly. Instead, it is inextricably embedded within the context. As demonstrated above, our pipeline handles this edge case by instructing the extraction module to capture contextual actions (e.g., calling the police twice) that implicitly reveal the subjective state. By isolating these inferred subjective elements (M) from the fatal physical outcome (A), the SOAM schema prevents the model from blindly equating a lethal consequence with malicious intent. This structured breakdown provides the exact factual anchor the Lawyer Agent needs to construct the defense arguments later shown in Figure 7.

[Illegality]
label: NOT_JUSTIFIED
grounds:
Lack of Imminence: While the victim initially attacked with a knife, the defendant successfully subdued him, threw the knife away, and pinned him to the floor. At this point, the "imminent danger" of the knife attack had effectively ended. The victim was an elderly man (76 years old) and was already physically suppressed by the defendant.

Excessive Defense: The defendant continued to press his knee against the victim's neck for approximately **10 minutes** even after the victim was subdued. Pressing the neck of an elderly person for such a prolonged period, leading to asphyxiation and death, clearly **exceeds the limits of necessary defense**.

Conclusion: The act of prolonged choking went beyond mere protection and resulted in death. Therefore, it constitutes Excessive Defense and is **Not Justified**.

(B) Model's First Judgement in 'Unlawfulness' stage

Figure 6: Model's initial judgment at the Unlawfulness stage, concluding that the act was not justified.

Initial Judgment. As shown in Figure 6, the base model initially judged the defendant's actions as *Not Justified* at the Unlawfulness stage. The model reasoned that the imminent danger had ended once the knife was removed and that maintaining neck compression for an extended period constituted excessive defense.

[Defense-Illegality]
point1: Imminent Danger Persisted
Evidence: *"...the victim continued to struggle and threaten to kill the defendant even after being subdued..."*

Argument: The Prosecution claims the danger ended. However, the facts explicitly state the victim was still struggling and shouting valid death threats ("I will kill you"). For the defendant, facing a knife-wielding attacker who had just stabbed him (causing lacerations), the fear and danger were continuous. Looking back comfortably from a desk (Hindsight) ignores the terrifying reality of the struggle.

point2: Necessity of the Act
Evidence: *"Called 112... called again... 'Why aren't you here yet? Come quickly.'"*

Argument: The defendant was begging for police help while holding down a lethal attacker. He did not choke the victim to kill, but to hold him until police arrived. The 10-minute duration was defined by the police arrival time (02:37), not the defendant's malice. He had no other means to safely restrain a struggling man who promised to kill him.

point3: Victim's Fault
Evidence: *"Victim C... returned with a knife... shouted 'I will kill you'..."*

Argument: The victim initiated a lethal encounter. The defendant acted purely defensively against an unprovoked, armed invasion of his home.

(C) Lawyer Agent's Defense in 'Unlawfulness' stage

Figure 7: Lawyer Agent's stage-specific defense arguments highlighting imminence, necessity, and victim fault.

Lawyer Agent Intervention. Figure 7 illustrates the Lawyer Agent's intervention at the Unlawfulness stage. Without presenting new facts, the agent reorganizes existing evidence into legally salient defense arguments, emphasizing the persistence of imminent danger, the necessity of restraint while awaiting a delayed police response, and the victim's fault as the initiator of a lethal armed attack.

[Illegality-Final]
label: JUSTIFIED
grounds:
Victim's Fault: According to the case facts, Victim C returned to the defendant's home with a knife to threaten and attack after a dispute. This series of actions clearly shows the victim was at fault and directly threatened the defendant's life.

Imminent Danger: During the knife attack, the defendant was in extreme danger. The measures taken to stop the attack (suppression) were a reasonable reaction to eliminate the immediate danger in a desperate situation. The victim's continued struggle sustained this imminence.

Last Resort (No Other Choice): Facing a lethal threat and waiting for police who were delayed, the defendant had no other choice to ensure his safety but to maintain suppression. Taking necessary defensive measures was blameless. **Conclusion**: The defendant's actions meet the conditions for Justifiable Defense. Although the specific operation might have physically exceeded limits (resulting in death), the intent and context define it as **Self-Defense**, negating illegality.

(D) Model's Revised Final Judgement

Figure 8: Model's revised final judgment accepting justification after defense intervention.

Revised Judgment. After incorporating the Lawyer Agent's defense arguments, the model revises its judgment, as shown in Figure 8, concluding that the defendant's conduct was *Justified*. The revised decision recognizes continued imminence,

Backbone	Method	Avg Calls	Input Tok.	Output Tok.	Latency (s)
Qwen-2.5-7B	Baseline D	4.0	1,830	1,392	**3.35**
	Self-Refine	4.0	2,724	2,952	4.86
	Debate	8.0	16,167	12,997	16.08
	OBJECTION	7.17	6,149	3,285	**6.77**
Llama-3.1-8B	Baseline D	4.0	1,830	637	**3.63**
	Self-Refine	4.0	2,724	1,351	5.45
	Debate	8.0	9,151	5,947	18.70
	OBJECTION	5.57	3,717	1,168	**5.77**

Table 12: Efficiency comparison on the CAIL dataset. **Avg Calls**, **Tokens**, and **Latency** for OBJECTION are weighted averages representing the early-exit distribution. OBJECTION is significantly faster than the multi-agent Debate baseline, thanks to fewer expected interactions.

lack of reasonable alternatives, and last-resort necessity, thereby negating unlawfulness through justifiable self-defense.

D Additional Implementation Details

Backbone Large Language Models We employed the following open-weight large language models in our experiments. The primary backbone model was `Qwen2.5-7B-Instruct` (Alibaba Cloud). For robustness, we additionally evaluated `Llama-3.1-8B-Instruct` (Meta) and `Gemma2-9B-it` (Google).

Inference Infrastructure All inference experiments were conducted using `vLLM` (v0.11.0) for high-throughput model serving via an OpenAI-compatible API server. Experiments were run on a workstation equipped with a single NVIDIA RTX PRO 6000 Blackwell GPU with 96GB VRAM. Models were loaded with automatic precision (typically `bfloat16`) to balance numerical stability and throughput.

Inference Hyperparameters To guarantee equitable comparison across all pipelines, we used consistent generation settings throughout the experiments: Temperature was set to 0.0 (greedy decoding for reproducibility), The maximum number of generation tokens was 2,048, The maximum context length was 8,192 tokens, and GPU memory utilization was capped at 0.90 to reserve capacity for the vLLM key–value cache.

LJPIV Fine-tuning Baseline Implementation For the LJPIV baseline, we fine-tuned `Qwen/Qwen2.5-7B-Instruct` using LoRA (Low-Rank Adaptation). Fine-tuning was performed in a separate environment equipped with a single NVIDIA RTX A6000 GPU (48GB VRAM).

Fine-tuning Configuration We used the PEFT library (v0.17.1) with the LoRA method. LoRA was applied to the following target modules: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The LoRA rank was set to 8, with LoRA alpha set to 32 and dropout set to 0.1.

Training Hyperparameters Optimization was performed using AdamW (PyTorch implementation) with a learning rate of 5×10^{-5} and a cosine learning rate scheduler. The batch size was 16 per device with gradient accumulation steps set to 2, resulting in an effective batch size of 32. Models were trained for three epochs using `bfloat16` precision, with a maximum sequence length of 2,048 tokens.

E Inference Efficiency Analysis

To evaluate the computational practicality of our framework, we measured the average processing time (latency) and token consumption per case on the **CAIL** dataset. Experiments were conducted on a single `NVIDIA RTX PRO 6000 blackwell` GPU using the `vLLM` backend.

We compared **OBJECTION** against three baselines: Baseline D (Chain-of-Thought), Self-Refine, and Debate (Multi-Agent). Table 12 summarizes the results across two backbones.

Efficiency Gains via Early Exit As shown in Table 12, the **Debate** baseline requires a fixed number of interactions (8.0 calls), leading to consistently high latency. In contrast, **OBJECTION** utilizes an early-exit mechanism, substantially decreasing the expected computational cost based on the verdict distribution:

- **Llama-3.1:** Since the model frequently dismisses charges at the Offense stage (62.68%),

the average API calls drop to **5.57**, shortening latency to **5.77s**.

- **Qwen-2.5:** With a higher conviction rate, the model averages **7.17** calls with a latency of **6.77s**.

This dynamic computation enables **OBJECTION** to achieve latency reductions of approximately **69.1%** (Llama) and **57.9%** (Qwen) compared to the Debate baseline. This shows that our structured agent system offers a superior efficiency-performance trade-off compared to unstructured debate frameworks.

F Ablation on LJPIV-CAIL dataset

Table 13 provides the full ablation study results on the LJPIV-CAIL benchmark, supplementing the main analysis in Section 4.4. The results follow the same trend observed in the ELAM dataset, confirming the integral role of our Lawyer Agent.

Table 13: Ablation Results (w/ `Qwen 2.5`) on LJPIV-CAIL dataset. The Lawyer Agent is effective in a low False Guilty Rate (FGR).

Configuration	G-F1	NG-F1	FGR↓	FNR↓
OBJECTION (Full)	**0.75**	**0.80**	**10.71**	33.93
(-) Schema	0.68	0.76	13.21	42.14
(-) Lawyer Agent	0.66	0.69	27.86	37.14
(-) 3-Step Structure	0.35	0.57	31.07	72.14
Ours w/ Normal Critic	**0.75**	0.59	55.71	**5.71**
LJPIV	0.80	0.83	9.29	27.14
LJPIV + Lawyer Agent	0.02	0.67	0.00	98.93

G Statistical Significance Testing

Protocol. All experiments use greedy decoding ($T = 0.0$), so the results carry no sampling variance and run-level variance testing is inapplicable. For every backbone–dataset setting we apply an exact McNemar test to the paired verdicts on the innocent subset, which tests FGR directly. All comparisons use the same per-case predictions reported in Table 2.

Results. As reported in Table 14, OBJECTION reduces FGR significantly in 20 of the 27 comparisons. Against the single-pass baselines the effect is near-uniform: 9 of 9 against Self-Refine and 8 of 9 against Baseline D, with the sole exception on Natural Innocent with Llama-3.1, where Baseline D already over-acquits (FNR 45.08%) and therefore starts from an atypically low FGR. Against Debate-Feedback the outcome is backbone-dependent—3 significant wins, 3 losses, and 3 ties—which is consistent with the volatility reported in Section 4.4: on Llama-3.1 the debate collapses into indiscriminate acquittal.

H Extended Experiments

H.1 Coverage across Acquittal Grounds

As shown in the Lawyer Agent prompt (Appendix A), the prompt presents a set of legal doctrines appropriate to each acquittal phase. One might object that this is too explicit a hint. We therefore tested the pipeline on acquittal grounds

Table 14: Exact McNemar tests on FGR. Each cell gives ΔFGR (OBJECTION minus the baseline, in points; negative favors OBJECTION) and the p-value. **Bold** marks a significant advantage for OBJECTION at $\alpha = .05$.

Setting	vs. Baseline D		vs. Self-Refine		vs. Debate	
	Δ	p	Δ	p	Δ	p
CAIL / Qwen	**−42.5**	**9.2e−35**	**−38.6**	**1.1e−27**	**−11.4**	**.018**
CAIL / Llama	**−20.0**	**2.3e−10**	**−28.6**	**2.7e−11**	+20.0	7.5e−15
CAIL / Gemma	**−31.1**	**3.4e−24**	**−33.6**	**1.7e−24**	+3.9	.215
ELAM / Qwen	**−33.2**	**4.4e−24**	**−34.4**	**9.4e−20**	**−3.8**	**.047**
ELAM / Llama	**−22.4**	**9.0e−11**	**−36.0**	**4.8e−18**	+18.4	4.7e−12
ELAM / Gemma	**−29.4**	**4.0e−20**	**−37.8**	**2.6e−25**	−1.8	1.000
Natural / Qwen	**−67.8**	**<1e−300**	**−69.6**	**<1e−300**	**−26.3**	**3.1e−13**
Natural / Llama	+9.7	2.0e−4	**−32.2**	**6.9e−128**	+38.7	1.5e−155
Natural / Gemma	**−60.3**	**<1e−300**	**−49.8**	**2.0e−257**	−7.1	.389

that do not appear in the prompt, and the performance generalized and held up. Table 15 breaks down FGR by the ground on which the court actually acquitted; the reduction is large and consistent across all three stages, including the Offense stage, which is dominated by evidentiary insufficiency rather than by any named doctrine.

Table 15: FGR (%) by acquittal ground on the Natural Innocent dataset (Qwen-2.5-7B).

Acquittal Ground	n	Baseline D	Ours
Offense	1,464	85.79	**17.08**
Unlawfulness	431	86.31	**17.40**
Culpability	226	72.57	**12.83**
All	2,121	84.49	**16.69**

H.2 Failure Mode Analysis

Table 16: Distribution of the 404 false acquittals on the Natural Innocent dataset (Qwen-2.5-7B).

Breakdown	n	%
By exit stage		
Offense	370	91.6
Unlawfulness	21	5.2
Culpability	13	3.2
By charge type		
Assault / bodily injury	171	42.3
Fraud / embezzlement	104	25.7
Other	129	31.9

Table 16 breaks down the 404 false acquittals produced on Natural Innocent with Qwen-2.5-7B. The errors concentrate at the Offense stage, and roughly 68% of those exits show a verdict–reasoning inconsistency—the reasoning finds the elements satisfied while the emitted label is `NO`—a self-consistency failure of the backbone that is markedly rarer on the commercial APIs (Table 4). The errors are also skewed by charge type: assault cases typically involve over-crediting a self-defense narrative in a mutual altercation, and fraud cases an uncorroborated negation of intent, both of which bear on public safety and motivate the human-in-the-loop deployment described in Section 7.

I Licenses and Attribution

External resources. Table 17 summarizes the external datasets, models, and software used in this work. All resources are used for research and evaluation purposes only.

Resource	Type	License	Use in this work
CAIL2018 (Xiao et al., 2018)	Dataset	MIT	In-domain LJP benchmark.
ELAM (via LJPIV) (Yu et al., 2022)	Dataset	MIT	Out-of-domain LJP benchmark.
Qwen-2.5-7B-Instruct (Yang et al., 2024)	LLM	Apache 2.0	Primary inference backbone.
Llama-3.1-8B-Instruct (Dubey et al., 2024)	LLM	Llama 3.1 Community	Robustness evaluation backbone.
Gemma-2-9B-it (Gemma Team, 2024)	LLM	Gemma Terms of Use	Robustness evaluation backbone.
vLLM v0.11.0	Software	Apache 2.0	LLM inference server.
PEFT v0.17.1	Software	Apache 2.0	LoRA fine-tuning.
Self-Refine (Madaan et al., 2023)	Baseline	Apache 2.0	Generic critic baseline.
LJPIV (Zhang et al., 2025a)	Baseline	Apache 2.0	Fine-tuned SOTA baseline.
Debate-Feedback (Chen et al., 2025c)	Baseline	Apache 2.0	Multi-agent debate baseline.

Table 17: Licenses and intended use of external resources. All are used solely for research purposes.