Why LLMs Forget the Old Law: Locating and Patching New-Law Bias in Legal Reasoning

Anonymous ACL submission

Abstract

Leveraging Large Language Models (LLMs) to assist in complex legal decision-making can significantly enhance judicial efficiency. However, when handling intertemporal legal conflicts, where laws change over time, LLMs often struggle to accurately identify which legal statute applies, and their decision-making processes lack transparency. This paper employs mechanistic interpretability analysis to uncover the root cause of this failure: we find that while early layers capture relevant legal features, deep negative attention heads dominate the model's blind preference for later laws, overriding earlier rational judgments. To address this, we propose **InterLex**, a targeted intervention framework for judicial prediction. The framework ensures the precise application of legal statutes by suppressing the activation of specific attention heads and leveraging attention alignment in intermediate layers. Furthermore, we construct **InterLexBench**, an intertemporal legal conflict benchmark, comprising 500 real-world and synthetic cases across diverse legal scenarios. Experimental results on InterLexBench show that the framework outperforms the base model by 5.4 percentage points in Macro F1 score on charge prediction, while also attaining state-of-the-art performance on the existing public dataset. Our code and benchmark are available at https://anonymous.4open.science/r/lex-E550/.

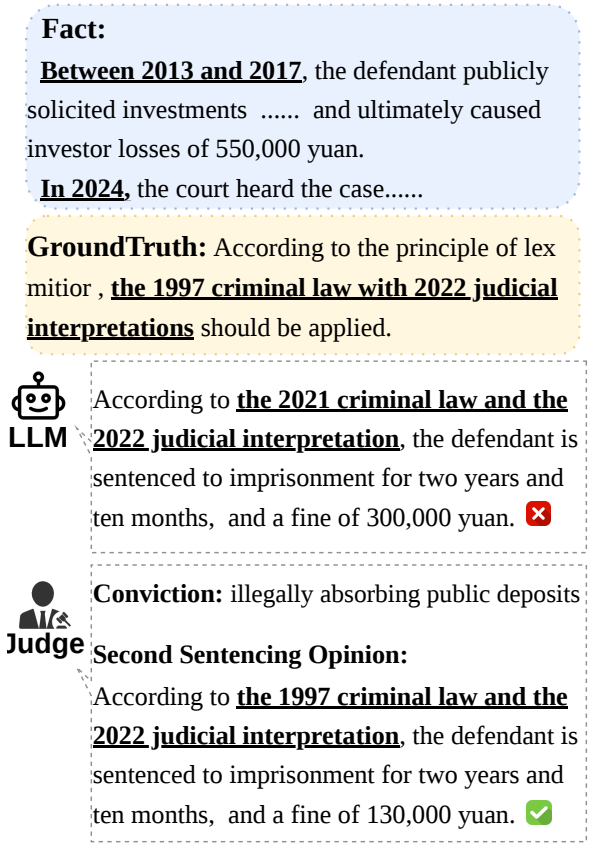


Figure 1: An example of how a case involving criminal acts committed between 2013 and 2017 is adjudicated under the legal statutes in effect in 2024. It contrasts the old law (lighter fine) with the new law (heavier fine) and highlights the LLM's failure.

1 Introduction

The rapid advancement of LLMs has established a new paradigm for intelligent legal applications, demonstrating remarkable capabilities in semantic understanding and generative reasoning, particularly in Legal Judgment Prediction (LJP) (Lai et al., 2024; Wei et al., 2025; Dina et al., 2025; Li et al., 2024b). However, when confronted with intertemporal legal conflicts, scenarios where criminal acts occurred under an older statute while the trial takes place after a newer law has taken effect, LLMs exhibit a systemic "new law bias". Specifically, LLMs tend to mechanically apply the most recent legal statutes without considering fundamental principles of criminal law, such as the *lex mitior* principle, which requires applying more lenient laws retroactively to benefit the defendant. This results in erroneous charges and disproportionately severe sentencing.

As illustrated in Figure 1, taking the crime of

illegally taking public deposits as an example, the principal penalties prescribed by the old and new laws are largely similar, yet the old law imposes a more lenient fine. Recent studies show that LLMs produce errors in legal tasks, and this failure stems not merely from insufficient retrieval, but from systematic biases within the models themselves, including inductive biases and uncritical acceptance of false premises (Dahl et al., 2024; Magesh et al., 2025). Current mainstream mitigation strategies, including domain-specific fine-tuning (Deng et al., 2023) and Chain-of-Thought prompting (Jiang and Yang, 2023; Yuan et al., 2024a), treat LLMs as end-to-end semantic matchers, failing to address the temporal-logical prioritization and normative hierarchy resolution required in legal conflicts. Notably, while Retrieval-Augmented Generation(RAG) can retrieve multiple versions of legal statutes, its lack of an intrinsic conflict resolution mechanism leads the model to treat old and new statutes as conflicting noise (Xu et al., 2024; Jin et al., 2024; Zhao et al., 2024; Gao et al., 2025), ultimately exacerbating reasoning errors.

This paper departs from the prevailing "black-box" paradigm and investigates the internal computational processes underlying LLM legal reasoning from a mechanistic interpretability perspective (Rai et al., 2024; Wang et al., 2024; Zhang et al., 2025a). Through comparative attention analysis across different legal statutes and cases, combined with causal intervention experiments (Heimersheim and Nanda, 2024; Meng et al., 2022; Wang et al., 2023), we identify the neural origins of the "new law bias": specific deep-layer attention heads dominate the final layers of the model, systematically suppressing certain semantic representations captured in earlier layers and forcibly overwriting them with features of the newer law (Jin et al., 2024). Based on this discovery, we propose **InterLex**, an inference-time intervention framework (Li et al., 2023; Wehner et al., 2025; Postmus and Abreu, 2024), which comprises two core components: Intertemporal conflict resolution operates by silencing the outputs of identified negative attention heads during forward propagation (Jin et al., 2024; Meng et al., 2022). This directly disrupts the signal pathway of the new law bias, reversing the selection probability in favor of the older law. Charge disambiguation leverages attention alignment metrics to detect confusable charges and performs fine-grained element-wise comparison based on the four constitutive elements of crime to achieve accurate charge determination. This framework requires no parameter fine-tuning; it reliably corrects model judgment biases solely through controlled interventions at inference time.

The main contributions are as follows:

- We propose a mechanistically grounded intervention method for legal reasoning, integrating activation patching with intermediate-layer alignment to systematically resolve statute selection and charge disambiguation in intertemporal legal conflict scenarios.
- We release an intertemporal legal conflict benchmark of 500 real judicial precedents, augmented with counterfactual temporal perturbations and colloquial style rewrites, filling a systematic evaluation void in this domain.
- On our intertemporal conflict benchmark, InterLex surpasses the base model by 5.4 Macro F1 points in charge prediction, and outperforms it by 2.2 percentage points in prison term prediction score, achieving the best results across five subtasks.

2 Related Work

2.1 Large Language Models and Agents in the Legal Domain

Recent research has developed LLMs for legal domains. In Chinese legal contexts, LawGPT_zh (Liu et al., 2023), Lawyer LLaMA (Huang et al., 2023), DISC-LawLLM (Yue et al., 2024), and ChatLaw (Cui et al., 2024) advance domain-specific capabilities from knowledge coverage toward specialized practice. Internationally, SaulLM (Colombo et al., 2024) and LawLLM (Shu et al., 2024) provide deep adaptations to different jurisdictions. Concurrently, LLM-based legal agents have emerged. Jing et al. (2025) propose role-based judicial deliberation to enhance judgment interpretability. AgentCourt (Chen et al., 2025) simulates adversarial courtroom debates, while SimCourt (Zhang et al., 2025b) constructs a framework for Chinese criminal trials.

However, existing works exhibit two key limitations. First, they overlook legal intertemporal conflicts, failing to handle cases where applicable law differs between the time of act and judgment. Second, their decision-making processes lack interpretability and diagnostic capabilities. Existing benchmarks (Chalkidis et al., 2022; Guha et al., 2023; Fei et al., 2024) focus primarily on static legal knowledge, insufficiently covering complex reasoning with legal changes.

2.2 Model Internal Mechanisms and Interpretability

Understanding Transformer internals is essential for diagnosing model reasoning. Layer-wise analysis has identified functional stages (Lad et al., 2024) and information bottlenecks in middle layers (Skean et al., 2025), which can be examined using tools such as the Logit Lens (nostalgebraist, 2020). Attention research has uncovered recurring circuits, including induction heads (Olsson et al., 2022) and successor heads (Gould et al., 2024). Lindsey et al. (2025) further show how low-level circuits compose into higher-level cognitive operations. Causal intervention methods (Vig et al., 2020; Geiger et al., 2024) move beyond correlation by enabling controlled manipulation of internal representations to establish mechanistic evidence. More recent work on context-based reasoning (Yüksekgönül et al., 2024; Halawi et al., 2024) identifies attention layers critical for integrating contextual information, yet these studies largely concern how models leverage parametric knowledge rather than how they resolve conflicts between context and prior knowledge. This work focuses on legal intertemporal reasoning as a specific and underexplored instance of such conflict, employing causal interventions to expose systematic biases and to design a training-free conflict detection mechanism.

3 Preliminaries

This section introduces the problem formulation and presents two findings that reveal failure modes in LLM-based legal reasoning. These findings directly motivate the intervention and disambiguation mechanisms developed in Section 4. Detailed definitions of the analytical tools used to probe model internals are provided in Appendix B.

3.1 Problem Description

Legal Judgment Prediction (LJP) predicts the judgment Y (charges and sentencing) for a case description $X = \{w_1, w_2, \ldots, w_n\}$ based on legal statutes $\mathcal{L}$. Unlike conventional LJP with static $\mathcal{L}$, real-world judicial practice requires evaluating X against intertemporal law versions $\mathcal{L} = \{l_{v1}, l_{v2}, \ldots\}$ using the temporal anchors of the criminal act and the principle of *lex mitior*. The objective is to learn:

$$Y = f(X, \{l_{v1}, l_{v2}, \ldots\}) \quad (1)$$

where f resolves legal conflicts by identifying the temporally applicable and most favorable statute version.

3.2 Finding 1: New Law Bias in Deep-Layer Attention Heads

We construct a probing dataset of $\sim$260 intertemporal legal conflict samples requiring the *lex mitior* principle. When the newer statute is legally correct, Qwen3-32B achieves 78.1% accuracy; when the older statute is correct, accuracy drops to 46.3%, below the 50% random baseline, revealing a systematic new law bias.

We locate this bias by comparing the contribution scores of the top-5 positive and top-5 negative heads (defined in Appendix B) across the two conditions. When the newer statute is correct, positive heads dominate. When the older statute is correct, both positive and negative head contributions weaken, but positive heads shrink by $\approx$78%, negative heads by only $\approx$66% (Figure 4). This differential attenuation inverts the balance, enabling negative heads to drive the incorrect prediction.

3.3 Finding 2: Alignment Symmetry as a Confusion Indicator

We compare alignment patterns for two groups, both evaluated on Qwen3-32B: a simple task group comprising 260 samples that contrast correct statutes with random distractors, and the aforementioned intertemporal legal conflict group of approximately 260 samples. Alignment scores quantify how strongly legal statutes attend to case facts during reasoning (see Appendix B for the formal definition). In simple tasks, the correct statute exhibits consistently higher alignment than the distractor across layers. In conflict tasks, alignment curves of the two competing statutes are nearly identical across all layers.

This contrast reveals that the alignment between candidate statutes serves as a reliable indicator of model decision certainty. When the model is confident, the alignment to the correct statute dominates; when confused between similar statutes, alignment scores converge.

4 Methodology

Building on the analyses in Section 3, we propose **InterLex**, an *interventional legal adjudication framework* that decomposes legal judgment prediction into four sequential stages, as illustrated in

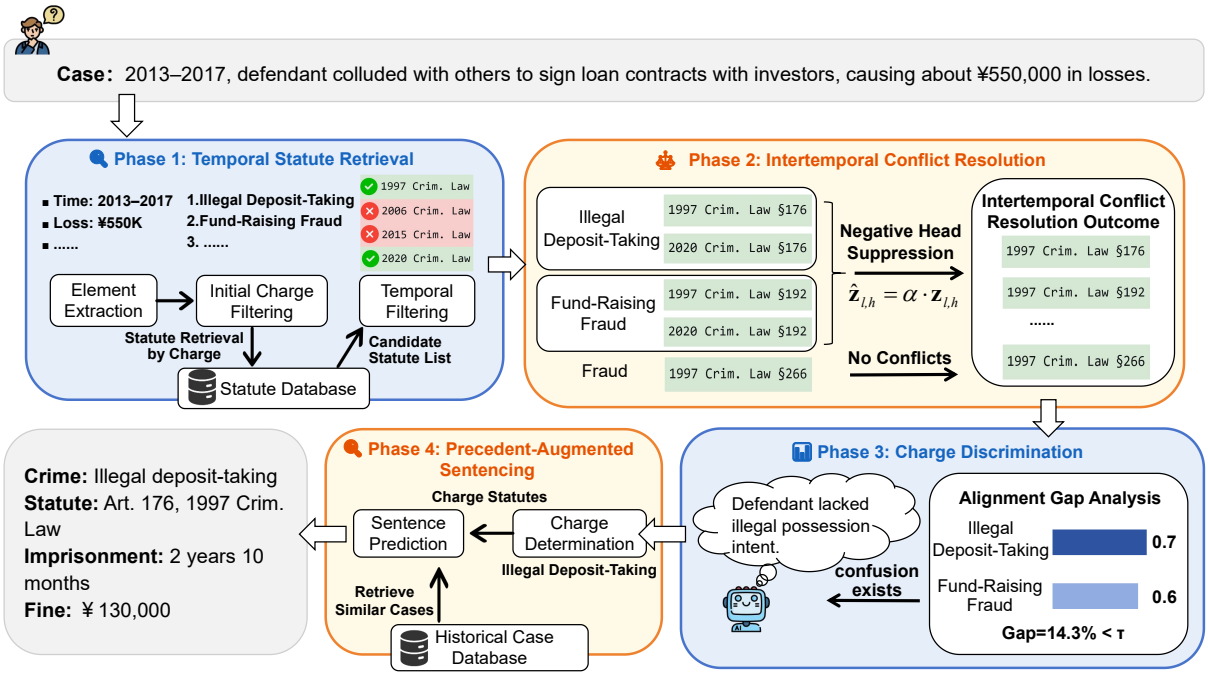


Figure 2: Overview of the **InterLex** framework. Given a case description, InterLex performs legal judgment prediction in four sequential stages: (1) structured element extraction with temporal statute retrieval, (2) intertemporal conflict resolution via targeted activation intervention, (3) charge disambiguation with four-element adjudication, and (4) precedent-augmented sentencing prediction.

Figure 2: (1) temporal statute retrieval with structured case element extraction, (2) intertemporal conflict resolution via targeted activation intervention, (3) charge disambiguation with four-element adjudication, and (4) precedent-augmented sentencing prediction.

4.1 Temporal Statute Retrieval

Given X, we prompt an LLM to extract structured elements:

$$(t_{\text{act}},\ t_{\text{trial}},\ \mathbf{v}_X) = \text{LLM}_{\text{extract}}(X) \tag{2}$$

where t_{act} and t_{trial} are the crime commission time and trial time, and $\mathbf{v}_X$ is a sentencing element vector encoding aggravating/mitigating factors. When temporal information is absent, t_{trial} defaults to the document date with t_{act} estimated via a two-year retrospective window.

Candidate charge identification. The LLM produces a Top-K candidate charge list $\mathcal{C} = \text{LLM}_{\text{top-}K}(X)$. We set $K{=}5$ based on recall analysis conducted on Qwen3-32B using CAIL2018 (Xiao et al., 2018): increasing K from 1 to 5 yields a 5.4-point recall gain, while further expansion to $K{=}20$ provides only 0.8% additional gain, suggesting that the marginal utility of enlarging the candidate set diminishes rapidly beyond $K{=}5$. A detailed comparison across two backbone models and four values of K is provided in Appendix D.

Temporal filtering. For each $c \in \mathcal{C}$, all historical versions are retrieved and filtered to retain the statute effective at t_{act} and t_{trial}:

$$\mathcal{L}' = \bigcup_{c \in \mathcal{C}} \{ l^{(c)}_{t_{\text{act}}},\ l^{(c)}_{t_{\text{trial}}} \} \tag{3}$$

This yields a charge-grouped set of at most two statute versions per charge.

4.2 Intertemporal Conflict Resolution

When a charge group $\mathcal{L}'_c$ contains statute versions from different legislative periods, an intertemporal conflict arises. As shown in Finding 1 (§3.2), such conflicts trigger a systematic new law bias originating from specific deep-layer negative heads. We propose a targeted intervention to suppress these bias signals.

Targeted intervention. Inspired by the causal tracing paradigm of (Meng et al., 2022), we directly intervene on the identified negative heads. For the h-th head at layer l with output $\mathbf{z}_{l,h}$, the intervened output is:

$$\hat{\mathbf{z}}_{l,h} = \alpha \cdot \mathbf{z}_{l,h}, \quad \forall\, (l,h) \in \mathcal{H}_{\text{neg}} \tag{4}$$

where $\mathcal{H}_{\text{neg}}$ is the set of top-k negative heads ranked by $|\Delta C_{l,h}|$ (Eq. 9) and $\alpha \in (0, 1)$ is a suppression factor.

Parameter selection. We conduct ablation experiments varying $\alpha \in [0.1, 0.5, 1, 2, 5]$ over different head-set sizes (top-3, top-5, top-10). Detailed experimental setup and results are presented in Appendix F. Two key findings emerge: (i)Negative heads carry bias signals: for negative heads, as α decreases from 1 to 0.1, accuracy shows an overall upward trend, confirming that these heads emit signals that compete with correct decision paths. Suppressing them helps restore the flow of positive features. (ii)Positive heads show a vulnerability: Suppressing positive heads ($\alpha < 1$) naturally degrades performance, while over-amplifying them ($\alpha > 1$) also leads to a gradual accuracy decline, suggesting that representational saturation can disrupt decision balance. Based on these observations, we intervene only on negative heads and adopt $\alpha = 0.1$ applied to the top-5 negative heads ($k = 5$).

For each charge group $\mathcal{L}'_c$ that contains conflicting statute versions, we scale the activations of the heads in $\mathcal{H}_{\text{neg}}$ by a factor of α, perform a controlled forward pass, and then select the statute version with the highest posterior probability:

$$l^*_c = \text{argmax}_{l \in \mathcal{L}'_c} P\big(l \mid X,\ \{\hat{\mathbf{z}}_{l,h}\}_{(l,h) \in \mathcal{H}_{\text{neg}}}\big) \quad (5)$$

Charge groups containing only a single statute version are left unmodified.

4.3 Charge Disambiguation

After intertemporal conflict resolution, each candidate charge is associated with a single statute. When multiple candidates remain, the framework must determine the final charge. As shown in Finding 2 (§3.3), alignment symmetry between competing statutes reliably indicates model confusion, motivating the following detection.

Mid-layer alignment metric. Our analysis shows that middle layers perform cross-segment semantic integration most effectively. As a concrete example, in Qwen3-32B we select the mid-layer set $\mathcal{M} = \{40, 42, 44, 46, 48, 50\}$ and define the average alignment over these layers as:

$$\bar{A}_c = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \text{Align}^{(m)}(\mathcal{S}_c,\ \mathcal{T}) \quad (6)$$

where $\text{Align}^{(m)}$ is defined in Eq. (8).

Confusion detection and disambiguation. To operationalize this metric for decision-making, we compute the relative alignment gap for each candidate charge, normalizing against the maximum to ensure comparability across samples of varying sequence lengths:

$$\text{Gap}(c) = \frac{\bar{A}_{\max} - \bar{A}_c}{\bar{A}_{\max}}, \quad \bar{A}_{\max} = \max_{c'} \bar{A}_{c'} \quad (7)$$

A charge is flagged as confused if its alignment is within a relative threshold τ of the maximum, forming the confusion set $\mathcal{C}_{\text{conf}} = \{c \in \mathcal{C} : \text{Gap}(c) < \tau\}$. Using a grid search on the probing dataset, we set τ^*=0.30 (recall 71.4%, false positive rate 28.6%).

The disambiguation strategy then branches on the size of this confusion set. When $|\mathcal{C}_{\text{conf}}|$=1, the top-aligned charge is directly accepted: $c^* = \text{argmax}_c \bar{A}_c$. When $|\mathcal{C}_{\text{conf}}|$>1, alignment alone is insufficient to distinguish the candidates, and we invoke structured four-element analysis based on Chinese criminal law theory. For each charge in $\mathcal{C}_{\text{conf}}$, the LLM evaluates four constitutive elements, namely subject qualification (criminal capacity, special status), subjective mens rea (intent /negligence, specific purpose), protected legal interest (property, public safety), and objective conduct (means, causation, result), with each element scored on $[0, 1]$. The charge with the highest composite score is selected; cases where all candidates exhibit substantial mismatches are flagged for manual review. The full definitions of the four elements, the prompt template, and the output schema are provided in Appendix E.

4.4 Precedent-Augmented Sentencing

Given the determined charge c^* and applicable statute $l^*_{c^*}$, this stage bridges legal characterization and quantitative sentencing via retrieval-augmented generation. We first filter the case corpus to retain only those cases sharing the same charge, obtaining $\mathcal{D}_{c^*} = \{d \in \mathcal{D} \mid \text{charge}(d) = c^*\}$. From this subset, we apply a temporal constraint that keeps only cases decided within the effective period of the applicable statute, i.e., $\text{eff}(l^*_{c^*}) \leq \text{date}(d) \leq \text{date}(X)$, simulating real judicial practice where judges reference prior precedents under the same legal regime. The retained cases are then ranked by element similarity to the current case, measured via the cosine similarity between sentencing element vectors: $\text{Sim}(X, d) = \cos(\mathbf{v}_X, \mathbf{v}_d)$. We select the top-3 most similar cases as the reference set $\mathcal{R}_X$.

Finally, the case facts, the applicable statute, and the reference set $\mathcal{R}_X$ are jointly composed into a prompt for the LLM, which outputs a structured sentencing recommendation comprising the principal punishment (category and duration), accessory punishments, fines, and an explanatory reasoning chain.

5 Experiments

5.1 Benchmark Construction

Existing legal judgment prediction datasets (e.g., CAIL2018 (Xiao et al., 2018)) assume a static legal environment and lack cross-version statute annotations, making them unsuitable for evaluating temporal legal reasoning. We construct an *intertemporal legal conflict benchmark* based on real case data, employing two data augmentation strategies:

Counterfactual temporal perturbation. From the PKULaw case database, we collect real judgments spanning key legislative amendment dates. For each sample with case facts X and original timestamps, we inject counterfactual times, forcing the model to adjudicate between the statute effective at the time of the act and the one effective at the time of trial. Groundtruth labels are determined by law graduates following the *lex mitior* principle, verified through double-blind annotation with senior expert review; only unanimously agreed samples are retained.

Textual style perturbation. To assess robustness to non-standard inputs, we rewrite a subset of formal judicial texts into colloquial narratives via LLM-guided paraphrasing, simulating real-world legal consultation scenarios. Rewrites are constrained to preserve all legally decisive facts, filtered by BERTScore thresholds, and manually spot-checked for semantic fidelity.

The benchmark covers 182 distinct charge categories, exhibiting a long-tailed distribution typical of real-world judicial corpora, with 57.1% of charge categories appearing only once. This skew, together with the prevalence of multi-charge joint indictments, poses a realistic challenge for fine-grained temporal legal reasoning.

5.2 Evaluation Metrics

We define five sub-tasks spanning charge determination and sentencing. For classification tasks we use macro F1 score; for regression tasks we adopt Normalized Log Distance (NLD) to handle highly skewed value distributions. Given true value y and prediction $\hat{y}$, the per-sample log distance is $d(y, \hat{y}) = |\ln(y+1) - \ln(\hat{y}+1)|$, and $\text{NLD} = (\ln y_{\max} - \bar{d})/\ln y_{\max} \in [0, 1]$, where $y_{\max}$ is the task-specific ceiling, set to 300 months for prison terms (corresponding to the 25-year statutory maximum for fixed-term imprisonment under Chinese criminal law) and 10^6 CNY for fines (covering all observed amounts in our dataset). If a prediction exceeds $y_{\max}$, it is treated as completely incorrect. Higher NLD indicates better accuracy.

- **Charge prediction** (F1): multi-label classification of applicable charges.
- **Principal punishment type** (F1): classification of punishment category (fixed-term imprisonment, life imprisonment, death penalty, etc.).
- **Prison term prediction** (NLD): regression on sentence duration in months.
- **Accessory punishment** (F1): classification of supplementary penalties.
- **Fine prediction** (NLD): regression on fine amount.

5.3 Baselines

We compare against three categories of baselines; detailed descriptions of each model and method are provided in Appendix A.

General-purpose LLMs. GPT-4o (OpenAI et al., 2024), QwQ-32B (Team, 2025), DeepSeek-V3.2 (DeepSeek-AI et al., 2025) and DeepSeek-R1 (Guo et al., 2025).

Legal-domain models. DISC-LawLLM (Yue et al., 2024), Fuzi-Mingcha (Deng et al., 2023), LawGPT_zh (Liu et al., 2023), LexiLaw (Li et al., 2024a), TaiLing (DUTIR-LegalIntelligence, 2024), and Wisdom-Interrogatory (Wu et al., 2024).

Other advanced methods. Search-o1 (Li et al., 2025), MALR (Yuan et al., 2024b), and ATRIE (Luo et al., 2025).

Implementation details. We employ two types of language models, which serve distinct roles across the four stages of **InterLex**. DeepSeek-R1 functions as the primary reasoning engine: it performs structured element extraction and candidate charge generation (Stage 1) and final sentencing prediction (Stage 4). Qwen3-32B provides activation intervention for intertemporal conflict resolution (Stage 2) and mid-layer alignment extraction for confusion detection (Stage 3). All inference is performed with temperature set to 0.1. Ex-

Table 1: Results on the intertemporal legal conflict benchmark. Best results are in **bold**; second-best are underlined.

	Method	Chg. F1↑	Prin. F1↑	Term NLD↑	Acc. F1↑	Fine NLD↑
General	GPT-4o	.297	.857	.817	.773	.871
	QwQ-32B	.578	.844	.818	.740	.797
	DS-V3.2	.580	.865	.859	.773	.884
	DS-V3.2(CoT)	.601	.877	.842	.786	.835
	DeepSeek-R1	.569	.877	.847	.780	.867
Legal	DISC-LawLLM	.342	.701	.207	.422	.063
	Fuzi	.151	.656	.445	.221	.063
	LawGPT_zh	.026	.292	.236	.123	.012
	LexiLaw	.152	.766	.065	.571	.139
	TaiLing	.382	.526	.485	.143	.016
	Wisdom-Int.	.200	.279	.159	.091	.029
Others	Search-o1	.602	.805	.774	.740	.768
	MALR	.526	.682	.618	.591	.571
	ATRIE	.559	.883	.852	.799	.844
	Ours_{QwQ}	.614	.860	.843	.779	.845
	Ours	**.623**	**.885**	**.869**	**.815**	**.891**

Note: Note: DS stands for DeepSeek, and CoT stands for Chain-of-Thought.

periments are conducted on NVIDIA A800 80GB GPUs.

5.4 Main Results

Table 1 presents the results on our intertemporal legal conflict benchmark. Our framework achieves the best performance across all five sub-tasks, outperforming its base model DeepSeek-R1 by 5.4 F1 points on charge prediction and 2.2 NLD points on prison term prediction. GPT-4o achieves competitive sentencing scores but exhibits notably poor charge prediction (F1=0.297), likely due to insufficient coverage of Chinese criminal law taxonomies.

DISC-LawLLM leads them with a charge prediction F1 of only 0.342, still trailing general LLMs by a wide margin. This shortfall stems primarily from their limited base model size(7B–13B) and training data that lack intertemporal conflict scenarios.

Among the other advanced methods, ATRIE is strong in sentencing(term NLD=0.852) but weak in charge prediction, while Search-o1's retrieval strategy lacks conflict-aware design. In contrast, after replacing the base model with QwQ-32B, the charge prediction F1 score rises from 0.578 to 0.614 and the prison term NLD from 0.818 to 0.843, demonstrating that the framework can consistently improve upon its base model.

We further evaluate on LawBench (Fei et al., 2024), a general-purpose Chinese legal benchmark. As shown in Table 2, our method achieves state-of-the-art results across charge, statute and prison term prediction, confirming generalizability beyond intertemporal conflict scenarios.

Table 2: Results on the LawBench general legal benchmark. Best results are in **bold**; second-best are underlined.

	Method	Charge F1↑	Statute F1↑	Term NLD↑
General	GPT-4o	.420	.525	.826
	QwQ-32B	.599	.732	.843
	DeepSeek-R1	.614	.728	.867
	Qwen-7B	.202	.199	.746
Legal	DISC-LawLLM	.412	.324	.758
	Fuzi-Mingcha	.560	.252	.772
	LawGPT_zh	.323	.292	.236
	LexiLaw	.398	.132	.781
	TaiLing	.435	.319	.684
	Wisdom-Int.	.351	.026	.731
Others	Search-o1	.621	.757	.724
	MALR	.571	.534	.617
	ATRIE	.608	.745	.810
	Ours	**.652**	**.790**	**.880**

5.5 Ablation Study

Table 3 reports the contribution of each module. The three modules target complementary aspects of the pipeline:

Intertemporal Conflict Resolution(Stage 2). This stage resolves statute version conflicts and thus directly affects sentencing. Without it, the framework defaults to the newest statute, lowering prison term NLD by 3.5 points and fine NLD by 2.1 points. This confirms that incorrect statute application propagates into sentencing error.

Charge Disambiguation(Stage 3). This stage determines the final charge based on the previously selected statute. Removing it and selecting the charge with the highest mid-layer alignment score directly reduces the charge F1 score by 3.6 points, suggesting that while alignment scores offer useful guidance, finer disambiguation remains necessary.

Precedent-Augmented Sentencing(Stage 4). This stage provides sentencing anchors from historical

Table 3: Ablation study on the intertemporal conflict benchmark. We individually remove Stage 2 (Intertemporal Conflict Resolution), Stage 3 (Charge Disambiguation), and Stage 4 (Precedent-Augmented Sentencing). Full stage descriptions are in Section 4.

Configuration	**Chg.** F1	**Prin.** F1	**Term** NLD	**Acc.** F1	**Fine** NLD
Full framework	.623	.885	.869	.815	.891
w/o stage2	–	.868	.843	.792	.870
w/o stage3	.587	.872	.849	.795	.883
w/o stage4	–	.863	.841	.787	.859

precedents. Since the charge has already been finalized in earlier stages, removing this stage affects only the sentencing metrics. In this configuration, the fine NLD declines by 3.2 points, demonstrating that LLMs struggle with sentencing without reference cases for calibration.

The results confirm that each module addresses a distinct failure mode, and their synergistic combination yields consistent improvements across all sub-tasks.

The results indicate that each stage addresses a distinct reasoning error, and their synergistic combination yields consistent improvements across all sub-tasks.

5.6 Case Study

To intuitively demonstrate the reasoning advantages of our method, we select a typical intertemporal conflict case and compare it with DeepSeek-V3.2 (CoT), as shown in Table 6.

Case Overview. Between 2013 and 2017, the defendant Wu conspired with others to publicly solicit investments under the pretext of various projects, promising monetary returns, resulting in approximately 550,000 RMB in losses to investors. Wu was identified as an accessory. The case was tried in 2024, during which Amendment (XI) to the Criminal Law had revised the statutory penalties for the relevant charges, constituting a typical intertemporal conflict scenario.

Reasoning Comparison. The differences between the two methods are reflected in three aspects: (1) *Charge Prediction*: DeepSeek-V3.2 equates "causing financial losses" with "intent of illegal possession," incorrectly predicting the charge as fundraising fraud. Our method, through candidate charge recall and alignment analysis, identifies the absence of circumstances such as absconding with funds or reckless squandering, and correctly determines the charge is illegally absorbing public deposits. (2) *Statute Application*: Although DeepSeek-V3.2 performs a *lex mitior* principle to Article 192, this analysis is built upon an incorrect charge, producing a cascading error. Our method places conflict resolution ahead of charge determination, independently resolving statute versions for all candidate charges, thereby preventing error propagation by design. (3) *Sentencing*: DeepSeek-V3.2 lacks external references and sets the fine at 100,000 RMB at its discretion. Our method retrieves three analogous cases as sentencing anchors and accurately predicts a prison term of 34 months. The deviation in the predicted fine may stem from factors that are difficult to model, such as regional economic disparities across the retrieved cases and the defendant's personal financial circumstances. These factors are often considered in judicial practice but are not present in the case fact descriptions provided as model input.

6 Conclusion

This work addresses the critical yet underexplored problem of intertemporal legal conflicts in legal judgment prediction. We systematically analyze the internal mechanisms of large language models (LLMs) for legal reasoning, revealing a systematic "new law bias" and its causal pathway via component attribution. Building on these mechanistic insights, we propose an interpretable and intervenable framework. This framework integrates targeted intervention to counteract the bias, charge discrimination based on intermediate layer alignment analysis, and similar-case retrieval for sentencing reference.

Extensive experiments on our newly constructed Intertemporal Legal Conflict Benchmark and the general LawBench demonstrate the superiority of our method over strong baselines, including general and legal-specific LLMs. Ablation studies confirm the contribution of each core module.

Limitations

Our study currently focuses on intertemporal conflicts; extending the framework to other conflict types (e.g., conflicts between general and special laws) is a key direction. The intervention strength also requires empirical tuning. Future work will explore adaptive intervention, integration with structured legal knowledge graphs, and the construction

of larger-scale, multi-domain benchmarks to advance trustworthy legal AI. We note that our framework is intended as an assistive tool and should not replace judicial discretion; erroneous predictions could lead to unfair sentencing if used without human oversight.

References

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. 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. 2025. 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.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2024. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *Preprint*, arXiv:2306.16092.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

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.

Nasa Zata Dina, Sri Devi Ravana, and Norisma Idris. 2025. Legal judgment prediction using natural language processing and machine learning methods: A systematic literature review. *SAGE Open*.

DUTIR-LegalIntelligence. 2024. Tailing. https://huggingface.co/DUTIR-LegalIntelligence/tailing.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Linfeng Gao and 1 others. 2025. Probing latent knowledge conflict for faithful retrieval-augmented generation. *arXiv preprint arXiv:2510.12460*.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning, 1-3 April 2024, Los Angeles, California, USA*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2024. Successor heads: Recurring, interpretable attention heads in the wild. In *The Twelfth International Conference on Learning Representations*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Re, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2024. Overthinking the truth: Understanding how language models process false demonstrations. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *Preprint*, arXiv:2305.15062.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the 19th International Conference on Artificial Intelligence and Law (ICAIL)*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215.

Huihao Jing, Wenbin Hu, Hongyu Luo, Jianhui Yang, Wei Fan, Haoran Li, and Yangqiu Song. 2025. Maslegalbench: Benchmarking multi-agent systems in deductive legal reasoning. *Preprint*, arXiv:2509.24922.

Vedang Lad, Wes Gurnee, and Max Tegmark. 2024. The remarkable robustness of LLMs: Stages of inference? In *ICML 2024 Workshop on Mechanistic Interpretability*.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2024. Large language models in law: A survey. *AI Open*, 5:181–196.

Haitao Li, Qingyao Ai, Qian Dong, and Yiqun Liu. 2024a. Lexilaw: A scalable legal language model for comprehensive legal understanding.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun LIU. 2024b. Lexeval: a comprehensive chinese legal benchmark for evaluating large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5420–5438, Suzhou, China. Association for Computational Linguistics.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023. Xiezhi:chinese law large language model. https://github.com/LiuHC0428/LAW_GPT.

Kangcheng Luo, Quzhe Huang, Cong Jiang, and Yansong Feng. 2025. Automating legal interpretation with LLMs: Retrieval, generation, and evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4015–4047, Vienna, Austria. Association for Computational Linguistics.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2025. Hallucination-free? assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*. ArXiv:2405.20362.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

nostalgebraist. 2020. Interpreting GPT: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. LessWrong.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Joris Postmus and Steven Abreu. 2024. Steering large language models using conceptors: Improving addition-based activation engineering. *arXiv preprint arXiv:2410.16314*.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the US legal system.

In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 4882–4889. ACM.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, volume 33, pages 12388–12401. Curran Associates, Inc.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. Knowledge mechanisms in large language models: A survey and perspective. *arXiv preprint arXiv:2407.15017*.

Jan Wehner and 1 others. 2025. Representation engineering for large-language models: Survey and research challenges. *arXiv preprint arXiv:2502.17601*.

B. Wei, Y. Yu, L. Gan, and 1 others. 2025. An llms-based neuro-symbolic legal judgment prediction framework for civil cases. *Artif Intell Law*.

Yiquan Wu, Yuhang Liu, Yifei Liu, Ang Li, Siying Zhou, and Kun Kuang. 2024. wisdominterrogatory. Available at GitHub, version 1.0.

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. *CoRR*, abs/1807.02478.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8541–8565.

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. 2024a. Can large language models grasp legal theories? Enhance legal reasoning with insights from multi-agent collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. 2024b. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7577–7597, Miami, Florida, USA. Association for Computational Linguistics.

Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2024. Lawllm: Intelligent legal system with legal reasoning and verifiable retrieval. In *Database Systems for Advanced Applications - 29th International Conference, DASFAA 2024, Gifu, Japan, July 2-5, 2024, Proceedings, Part V*, volume 14854 of *Lecture Notes in Computer Science*, pages 304–321. Springer.

Mert Yüksekgönül, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2024. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Zhang and 1 others. 2025a. Unboxing the black box: Mechanistic interpretability for algorithmic understanding of neural networks. *arXiv preprint arXiv:2511.19265*.

Kaiyuan Zhang, Jiaqi Li, Yueyue Wu, Haitao Li, Cheng Luo, Shaokun Zou, Yujia Zhou, Weihang Su, Qingyao Ai, and Yiqun Liu. 2025b. Chinese court simulation with llm-based agent system. *Preprint*, arXiv:2508.17322.

Yu Zhao, Xiaotang Du, and 1 others. 2024. Analysing the residual stream of language models under knowledge conflicts. *arXiv preprint arXiv:2410.16090*.

A Detailed Descriptions of Baselines

This appendix provides detailed descriptions of the baselines summarized in Section 5.3.

A.1 General-Purpose LLMs

General-purpose LLMs, pretrained on massive corpora and equipped with strong semantic understanding capabilities, have demonstrated outstanding performance across a wide range of natural language processing tasks. We adopt the following representative models as baselines:

- **GPT-4o** (OpenAI et al., 2024): An end-to-end multimodal autoregressive model developed by OpenAI, capable of generating text, image, and audio outputs. Compared with its predecessors, it achieves substantial improvements on non-English text understanding, vision, and speech tasks while offering faster response times.

- **QwQ-32B** (Team, 2025): A reasoning-enhanced model that supports a context length of 32k tokens and attains performance comparable to GPT-4 on tasks such as mathematical reasoning and code generation.

- **DeepSeek-V3.2** (DeepSeek-AI et al., 2025): A large-scale Mixture-of-Experts language model that reduces computational complexity via a sparse attention mechanism and substantially scales post-training compute through a scalable reinforcement learning framework, achieving performance on par with GPT-5 across a range of reasoning benchmarks. We evaluate it under two settings: direct inference and inference with Chain-of-Thought (CoT) prompting.

- **DeepSeek-R1** (Guo et al., 2025): A reasoning model trained within a reinforcement learning framework that explicitly produces complete reasoning chains, achieving performance close to OpenAI's o1 model on complex logical reasoning tasks.

A.2 Legal-Domain Models

Legal-domain models, adapted to legal corpora through domain-specific training, are better able to understand legal terminology and capture legal reasoning patterns. We select the following six representative Chinese legal LLMs:

- **DISC-LawLLM** (Yue et al., 2024): Built upon Baichuan-13B, it constructs its training data using a legal-syllogism strategy and integrates a dynamically updated legal knowledge base, supporting both legal reasoning and knowledge retrieval.

- **Fuzi-Mingcha** (Deng et al., 2023): Formalizes legal judgment analysis as four subtasks—statute retrieval, generation of constitutive elements of a crime, statute interpretation, and judgment prediction—within a judicial-syllogism reasoning framework.

- **LawGPT_zh** (Liu et al., 2023): Fine-tuned from ChatGLM-6B using the LoRA method on approximately 200k legal question-answering pairs, and further augmented with a retrieval module.

- **LexiLaw** (Li et al., 2024a): Fine-tuned from ChatGLM-6B on legal-domain datasets, with a particular focus on legal consultation scenarios.

- **TaiLing** (DUTIR-LegalIntelligence, 2024): Built upon the Qwen-7B series, providing judicial services such as text verification, information extraction, and sentencing assistance.

- **Wisdom-Interrogatory** (Wu et al., 2024): Built upon Baichuan-7B, with 40 GB of legal-domain knowledge injected through continued pretraining, followed by instruction tuning on 100k samples.

A.3 Other Advanced Methods

- **Search-o1** (Li et al., 2025): A reasoning framework based on an agentic retrieval-augmented generation mechanism. It enables the model to autonomously identify knowledge gaps during multi-step reasoning and to dynamically invoke external search for on-demand, iterative knowledge retrieval and integration, thereby mitigating the limited domain knowledge of LLMs in specialized fields.

- **MALR** (Yuan et al., 2024b): A multi-agent legal reasoning framework that designs multiple specialized agents with distinct roles (e.g., a statute-retrieval agent, a fact-analysis agent, and a reasoning-synthesis agent) to perform complex legal analysis through collaborative division of labor, mirroring the multi-party coordination found in real-world judicial practice.

- **ATRIE** (Luo et al., 2025): A retrieval-augmented legal reasoning method that focuses on the understanding and application of abstract concepts in legal texts (e.g., the scope of the term "household" in "household burglary"). It retrieves real cases from the validation set to obtain concrete interpretations of abstract legal concepts and, at inference time, combines such interpretations with rule-based

reasoning to improve the model's accurate application of legal norms.

B Analytical Tools for Model Probing

This appendix provides formal definitions of the analytical tools referenced in Section 3. We consider a decoder-only Transformer Φ with L layers and H attention heads per layer. For an input sequence, $A_{i,j}^{(l,h)}$ denotes the attention weight from token i to token j at layer l and head h.

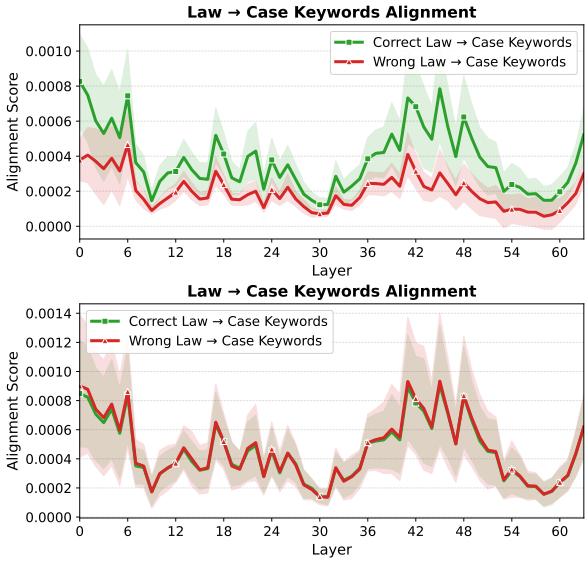


Figure 3: **Top:** Simple task (correct statute vs. random distractor). **Bottom:** Conflict task (correct statute vs. competing statute). Statute–fact alignment scores across layers for simple and intertemporal conflict tasks. In simple tasks, the correct statute exhibits consistently higher alignment than the distractor. In conflict tasks, alignment curves of the two competing statutes converge across all layers, indicating model confusion.

B.1 Statute–Fact Alignment

To quantify how strongly legal statutes attend to case facts during legal reasoning, we define the Statute–Fact Alignment score. Given legal statute token indices $\mathcal{S}$ and case fact token indices $\mathcal{T}$, the alignment at layer l is:

$$\text{Align}^{(l)}(\mathcal{S},\mathcal{T}) = \frac{1}{H|\mathcal{S}||\mathcal{T}|}\sum_{h=1}^{H}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{T}} A_{i,j}^{(l,h)}\,\mathbb{I}(i>j) \tag{8}$$

where $\mathbb{I}(i > j)$ enforces the causal mask. This score measures the density of information flow from case facts to legal statutes.

B.2 Component Attribution via Logit Lens

To identify which component drives the model's decisions, we apply Logit Lens (nostalgebraist, 2020) to decode intermediate representations. Let $\mathbf{p}_{l,h} = \mathbf{W}_U \cdot \mathbf{W}_{l,h}^{\text{out}} \cdot \mathbf{z}_{l,h}$ be the logit vector obtained by projecting the output of the h-th attention head in layer l through the unembedding matrix $\mathbf{W}_U$, where $\mathbf{z}_{l,h}$ is the head output before projection and $\mathbf{W}_{l,h}^{\text{out}}$ is the output projection matrix. The direct logit attribution of this head to the target prediction is defined as:

$$\Delta C_{l,h} = (\mathbf{p}_{l,h})_{t_{\text{correct}}} - (\mathbf{p}_{l,h})_{t_{\text{wrong}}}. \tag{9}$$

Heads with $\Delta C_{l,h} > 0$ are **positive heads** (favoring the correct answer); those with $\Delta C_{l,h} < 0$ are **negative heads** (biased toward the incorrect answer).

C Figures for Findings

This part provides supplementary figures supporting the findings in Section 3. Figure 4 shows the contribution scores of attention heads under the two conditions, and Figure 3 compares statute–fact alignment between simple and conflict tasks.

D Recall Analysis for Candidate Charge Identification

This appendix provides the full empirical justification for setting K=5 in the candidate charge identification step described in Section 4. To determine the optimal value of K, we use the charge prediction task from CAIL2018 (Xiao et al., 2018) as a validation set and quantitatively evaluate recall performance under different retrieval depths. We report results on two backbone LLMs, Qwen3-32B and DeepSeek-V3.2, with Hit Rate, Precision, Recall, and F1 measured at $K \in \{1, 5, 10, 20\}$. Results are summarized in Table 4.

Both backbones exhibit a similar pattern of diminishing marginal returns as K grows. For DeepSeek-V3.2, increasing K from 1 to 5 raises the hit rate from 86.20% to 94.60%, an 8.4-point improvement; further expanding K from 5 to 20 yields only an additional 0.8 points (94.60% $\rightarrow$ 95.40%). Qwen3-32B follows the same trend: its hit rate reaches 89.00% at K=5 and gains merely 0.8 points when K is further increased to 20.

Larger values of K not only substantially increase the token cost of downstream stages in the framework, but may also introduce long-tail noise

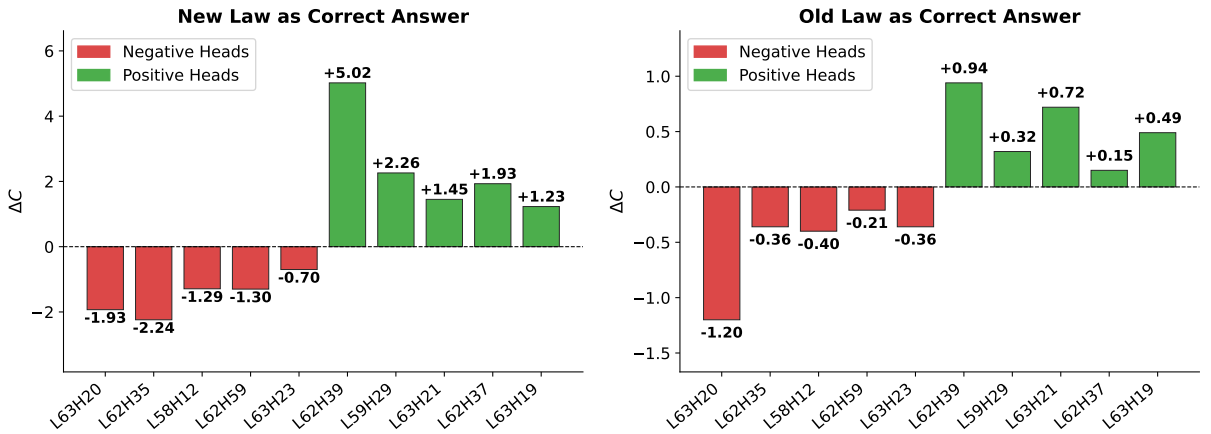


Figure 4: Contribution scores of the Top-5 positive and Top-5 negative attention heads under two conditions. **Left:** the newer statute is the correct answer; **Right:** the older statute is the correct answer. Positive heads decay by ≈78%, negative heads by ≈66% when the older statute is correct, inverting the influence balance.

Table 4: Candidate charge recall and classification performance of Qwen3-32B and DeepSeek-V3.2 on CAIL2018 under varying retrieval depths K.

Metric (%)	K=1	K=5	K=10	K=20
Qwen3-32B				
Hit Rate	83.60	89.00	89.40	89.80
Precision	83.60	18.04	9.06	4.58
Recall	82.93	88.67	89.07	89.73
F1	83.13	29.90	16.42	8.71
DeepSeek-V3.2				
Hit Rate	86.20	94.60	95.20	95.40
Precision	86.20	19.32	9.82	5.13
Recall	85.53	94.57	95.63	98.03
F1	85.73	31.99	17.77	9.73

that interferes with subsequent reasoning. We therefore adopt K=5 as the operating point, which strikes the best balance between coverage and computational efficiency. We note that the sharp drop in Precision and F1 at larger K is expected: these metrics penalize the inclusion of additional candidates, whereas the purpose of this stage is to maximize recall for the downstream filtering modules. Hit Rate and Recall are therefore the metrics of primary interest here.

E Details of Four-Element Adjudication

This part complements the four-element adjudication step described in Section 4.3 by providing (i) the full legal definitions of the four constitutive elements under Chinese criminal law theory, (ii) the prompt template used to elicit element-wise analysis from the LLM, and the structured output schema.

E.1 Definitions of the Four Constitutive Elements

Under Chinese criminal law theory, the constitution of a crime comprises the following four elements:

- **Subject of the crime.** Whether the actor satisfies the subject requirements of the charge, including criminal responsibility capacity (e.g., age, mental competence) and any special status required by the offense (e.g., the official capacity required for duty-related crimes).
- **Subjective element.** Whether the actor's mental state matches the form of culpability required by the charge, including the determination of negligence, the distinction among different forms of intent, and the presence of any specific criminal purpose (e.g., the purpose of unlawful possession).
- **Object of the crime.** Whether the legal interest infringed by the case coincides with the interest protected by the charge, such as property rights, personal rights, or public safety.
- **Objective element.** Whether the conduct, means, target, and consequences satisfy the objective requirements of the charge, including the concrete form of the act, causation, and the resulting harm.

For each candidate charge in the confusion set $\mathcal{C}_{\text{conf}}$, the model performs an element-by-element analysis using the prompt template in Appendix E.2, and produces one of the following adjudication outcomes based on the overall fit across the four elements:

- **Confirmed charge.** The charge c^* that best fits the case is explicitly identified, together with the basis of the determination.
- **Flagged for manual review.** When the case facts exhibit substantial mismatches with *all* candidate charges, the case is labeled "not guilty" under the current candidate set and flagged for manual review.

Four-Element Analysis Prompt Template

You are a judge. Based on the following case facts, please analyze whether the case satisfies the constitutive elements of **[Charge]**.

[Case Facts]
Case description

[Applicable Statute]
The best-matching statute version for this charge

Please analyze each of the four constitutive elements in turn and assign a matching score between 0 and 1:

1. **Subject:** Does the actor satisfy the subject requirements of this charge? (Consider age, criminal responsibility capacity, special status, etc.)
2. **Subjective element:** Does the actor's mental state satisfy the requirements of this charge? (Consider intent/negligence, criminal purpose, etc.)
3. **Object:** Does the legal interest infringed in the case correspond to the interest protected by this charge?
4. **Objective element:** Do the conduct, means, and consequences satisfy the objective requirements of this charge?

[Output Format] (Strictly follow the JSON format below)

```
{
  "charge": "<charge name>",
  "subject": {"score": 0.0-1.0, "reason":
       "<analysis>"},
  "subjective": {"score": 0.0-1.0, "
      reason": "<analysis>"},
  "object": {"score": 0.0-1.0, "reason":
      "<analysis>"},
  "objective": {"score": 0.0-1.0, "reason
      ": "<analysis>"}
}
```

Figure 5: Prompt template for four-element analysis. The placeholders **[Charge]**, **[Case Facts]**, and **[Applicable Statute]** are filled in for each candidate charge in $\mathcal{C}_{\text{conf}}$. The original prompt is in Chinese; the English version shown here is provided for the reader's convenience.

In confusion cases, the system preserves the complete four-element analysis for every candidate charge, including the per-element conclusions and the underlying reasoning, so that legal practitioners can verify the result. This design supports the rigor and traceability required of legal decision-making. We note that in the experimental evaluation reported in the main text, only the model's finally confirmed charge is used for accuracy computation.

E.2 Prompt Template

The structured prompt template used for four-element analysis is shown in Figure 5. The model is instructed to adopt the perspective of a judge and to evaluate each element independently with a numerical score in $[0, 1]$ together with a textual justification.

E.3 Output Schema and Aggregation

The JSON object returned by the model contains, for each of the four elements, a numerical score in $[0, 1]$ and a free-form textual justification. Let $s_e^c \in [0, 1]$ denote the score assigned to element $e \in \{\text{subject}, \text{subjective}, \text{object}, \text{objective}\}$ for candidate charge c. The composite score for charge c is the unweighted mean of the four element scores, $\bar{s}^c = \frac{1}{4}\sum_e s_e^c$, and the confirmed charge is selected as $c^* = \arg\max_{c \in \mathcal{C}_{\text{conf}}} \bar{s}^c$.

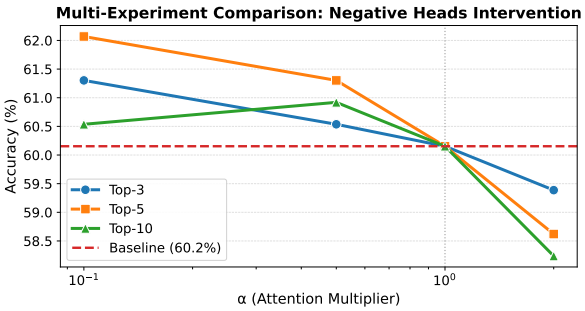
F Attention Head Suppression

To validate the mechanism and determine optimal parameters, we conduct ablation experiments on the Qwen3-32B model using the complex task validation set from Section 3. We select top-3, top-5, and top-10 positive and negative heads based on their contribution scores, and vary $\alpha \in \{0.1, 0.5, 1, 2, 5\}$. The experiments reveal a clear asymmetry in how the two types of attention heads respond to α.

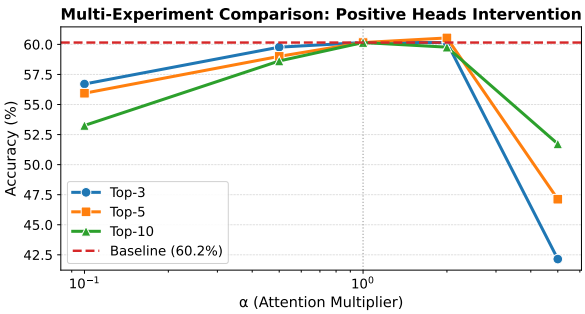
Effect of suppressing negative heads. Experimental results show that for $0 < \alpha < 1$, accuracy generally improves as α decreases (i.e., under stronger suppression). This confirms that negative heads likely carry bias signals that compete with correct decision signals. Suppressing these heads helps restore the flow of positive features.

Effect of suppressing positive heads. For $0 < \alpha < 1$, suppressing positive heads naturally degrades performance. However, unlike negative heads, continuously amplifying positive heads beyond $\alpha > 1$ gradually leads to a decrease in accuracy. This suggests that over-amplifying a single

signal may disrupt the representational space of the residual stream and affect decision balance.



(a) Negative head intervention



(b) Positive head intervention

Figure 6: Experiments on Qwen3-32B. Accuracy curves for different numbers of key attention heads under varying intervention strength α. The red dashed line indicates the baseline accuracy.

F.1 Transferability of the Temporal Conflict Resolution Method

The parameters α of the targeted intervention mechanism described above were calibrated on the Qwen3-32B model. It is therefore worth exploring whether the method can transfer to models of different sizes or architectures. This section evaluates transferability along two dimensions: cross-scale and cross-architecture.

F.1.1 Transferability Across Model Scales

We replicate the attention head contribution analysis and intervention experiments on Qwen3-8B and Qwen3-14B, discussing three aspects: structural distribution of key components, directional consistency of intervention effects, and parameter sensitivity.

Table 5 compares the normalized layer distribution of key attention heads across the three models. For all three models, the key decision heads are concentrated in the last 15%–25% of layers, and the median position shifts deeper as model size increases. This consistent distribution pattern indicates that the phenomenon of deep-layer attention heads assuming decision-making functions is not accidental to a specific model but a common characteristic of this model family on this type of task.

Table 5: Comparison of normalized layer distribution of key attention heads

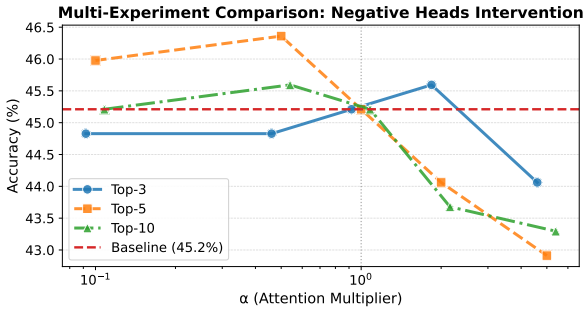
Model	Layers	Layer Range	Norm. Range	Median
Qwen3-8B	36	24–35	67%–97%	∼89%
Qwen3-14B	40	30–39	75%–98%	∼92%
Qwen3-32B	64	54–63	84%–98%	∼97%

Figures 7 and 8 show the intervention results on Qwen3-8B and Qwen3-14B, respectively. Regarding directional consistency, the three model scales generally exhibit similar trends, though with local deviations.

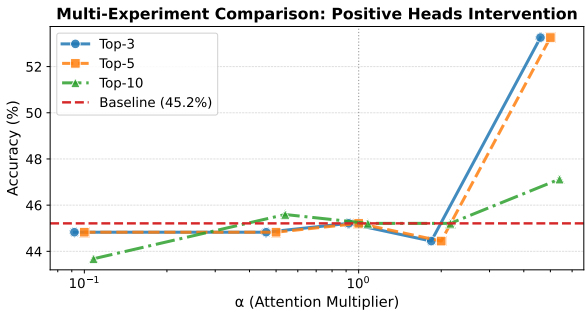
For negative heads, the 14B and 32B models show clear accuracy improvements under strong suppression ($\alpha = 0.1$); the 8B model also exhibits the same directional trend for Top-5 and Top-10, but Top-3 remains slightly below baseline in the suppression region, indicating that the benefit of suppressing negative heads is unstable when intervening on only a few heads in a small-scale model. Although the 14B model shows a clear overall trend, its Top-5 also falls slightly below baseline at moderate suppression strength ($\alpha = 0.5$). For positive heads, both 14B and 32B models show decreased accuracy in the suppression region, consistent with theoretical expectations; however, the 8B model is almost unaffected in the suppression region. Overall, the directional trends of interventions are largely consistent across the three models, but the frequency and magnitude of local deviations increase as model scale decreases, suggesting that the functional boundaries of attention heads become more blurred in smaller models.

F.1.2 Transferability Across Architectures

All the above analyses are limited to the Qwen3 family, which uses grouped query attention. To test the applicability of the targeted intervention method under different architectures, we replicate the full intervention experiment on DeepSeek-V2-Lite. DeepSeek-V2-Lite employs a mixture-of-experts architecture with multi-head latent attention (MLA), which compresses key-value representations via low-rank projection to significantly reduce KV cache overhead during inference. This architecture is fundamentally different from the

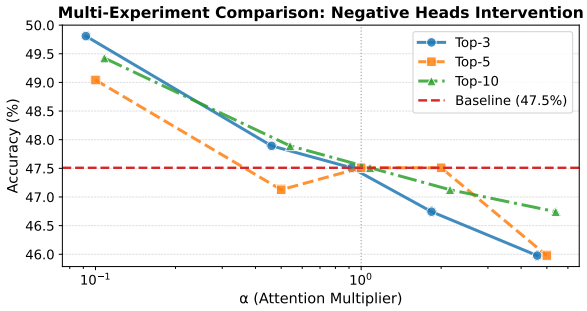


(a) Negative head intervention

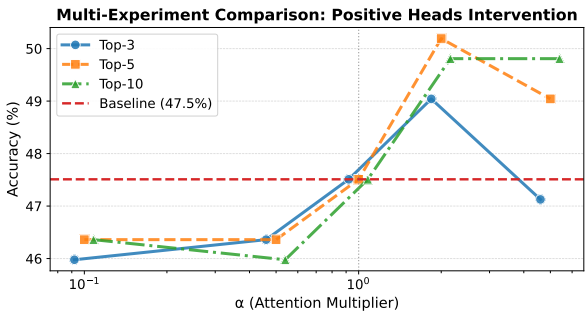


(b) Positive head intervention

Figure 7: Intervention effects on different numbers of key attention heads on Qwen3-8B



(a) Negative head intervention



(b) Positive head intervention

Figure 8: Intervention effects on different numbers of key attention heads on Qwen3-14B

Qwen3 family, making it an ideal testbed for cross-architectural transferability.

As shown in Figure 9, following the same procedure as for the Qwen3 family, we first perform attention head contribution analysis on calibration samples to identify positive and negative heads. We then conduct intervention experiments on the top-3, top-5, and top-10 head sets with $\alpha \in \{0.1, 0.5, 1, 2, 5\}$. The baseline accuracy is 44.8%.

Negative head intervention results. For negative heads on DeepSeek-V2-Lite (Figure 9a), in the suppression region ($0 < \alpha < 1$), all three intervention sets yield significant accuracy improvements: at $\alpha = 0.1$, Top-5 performs best with accuracy ∼56.7%; Top-3 follows with ∼56.3%; Top-10 reaches ∼54.3%, corresponding to improvements of approximately 11.9, 11.5, and 9.5 percentage points above baseline, respectively. As α increases from 0.1 to 1, accuracy declines. In the amplification region ($\alpha > 1$), accuracy fluctuates around baseline, approximately 45%–47%.

Positive head intervention results. Notably, the intervention effects for positive and negative heads on DeepSeek-V2-Lite exhibit an approximately mirror-symmetric structure. The main gain from negative head intervention is concentrated in the suppression region, with only small fluctuations around baseline in the amplification region. Positive head intervention shows the opposite pattern: accuracy remains slightly above baseline (approximately 45%–47%) in the suppression region, while it jumps substantially in the amplification region, reaching ∼57.0% for Top-5 at $\alpha = 5$. These two types of intervention confirm that positive and negative heads indeed perform opposing decision functions under the MLA architecture.

G Case Study Comparison Table

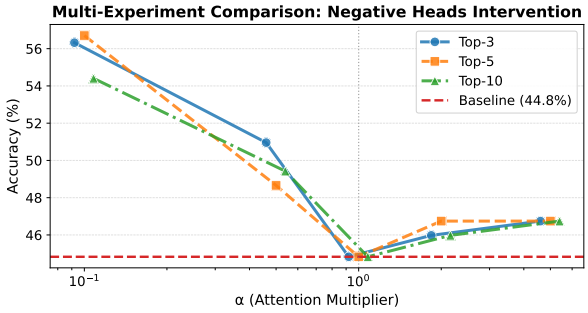
Table 6 provides the full reasoning comparison for the case study discussed in Section 5.6.

H Necessity and Sufficiency Verification of Key Negative Heads

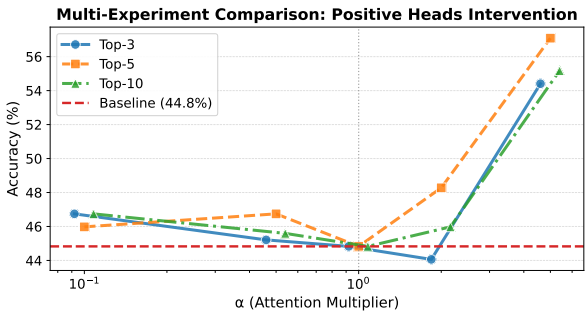
Two causal questions remain after the parameter ablation: (1) whether joint suppression of the identified negative heads exerts a causal effect on accuracy beyond what direct logit attribution can establish, and (2) whether each head within the set contributes a non-trivial portion of this effect. The probing set ($N = 261$) combined with the interven-

Table 6: Comparison of reasoning processes between DeepSeek-V3.2 (CoT) and our method. Red indicates errors; teal indicates correct reasoning.

Stage	DeepSeek-V3.2 (CoT)	Ours
Charge	Concludes Art. 192 [Fundraising Fraud] applies: "defrauded funds via fictitious projects with intent of illegal possession." Predicted: Fundraising Fraud.	Recalls 5 candidates; alignment scores: Illegal Absorption (0.0007) > Fundraising Fraud (0.0004). No absconding or reckless squandering found. Predicted: Illegal Absorption of Public Deposits.
Conflict Resol.	Applies favor-the-defendant to Art. 192 only: old law "5–10 yrs"; new law "7+ yrs or life." Applies old law—logic correct but wrong statute.	Applies the 1997 Criminal Law. Old and new laws differ in sentencing; the 1997 version has a lower fine ceiling and the case does not reach the higher tier added in 2020, so the 1997 version is more favorable.
Sentencing	Old-law range for fundraising fraud: 5–10 yrs. Base 5 yrs, reduced to 3 yrs (accessory). No case reference; fine 100k RMB.	Range: ≤3 yrs, fine 20k–200k RMB. 3 analogous cases retrieved. Final: 34 months, probation 36 months, fine 80k RMB.
Final Result	Fundraising Fraud, 36 months, fine 100k RMB	Illegally Absorbing Public Deposits, 34 months, probation 36 months, fine 80k RMB
Ground Truth	Illegally Absorbing Public Deposits, 34 months, probation 36 months, fine 130k RMB	



(a) Negative head intervention



(b) Positive head intervention

Figure 9: Intervention effects on different numbers of key attention heads on DeepSeek-V2-Lite

tion strength ($\alpha = 0.1$) yields a small number of decision flips by design, so accuracy-level per-head tests are not expected to be statistically resolvable at this sample size. We therefore use this section to establish the direction and magnitude of the intervention at the accuracy level, and defer per-head causal resolution to the logit-level path patching analysis in Section I, which measures continuous logit shifts rather than discrete accuracy flips.

H.1 Experimental Setup

Both experiments use Qwen3-32B on the full set of 261 complex-task samples with a fixed intervention strength $\alpha = 0.1$. The intervention targets are the top-5 negative heads ranked by direct logit attribution: $\mathcal{H}_{\text{neg}} = \{$L62H35, L63H20, L58H12, L62H59, L63H23$\}$.

Leave-One-Out (LOO) necessity test. For each head $(l_i, h_i) \in \mathcal{H}_{\text{neg}}$, we restore its intervention coefficient to $\alpha = 1$ while keeping the other four heads suppressed, and record the resulting accuracy $\text{Acc}_{-(l_i,h_i)}$. The per-head necessity score is

$$\text{Necessity}(l_i, h_i) = \text{Acc}_{\text{joint}} - \text{Acc}_{-(l_i,h_i)}. \quad (10)$$

A positive value indicates that removing the intervention on that head degrades the overall effect.

Single-head sufficiency test. For each head (l_i, h_i), we apply $\alpha = 0.1$ suppression to that head alone while leaving all others unmodified, and record the accuracy $\text{Acc}_{(l_i,h_i)}$. The per-head sufficiency score is

$$\text{Sufficiency}(l_i, h_i) = \text{Acc}_{(l_i,h_i)} - \text{Acc}_{\text{baseline}}. \quad (11)$$

Statistical testing. We report one formal test, on the joint effect: an exact McNemar test on paired per-sample correctness between baseline and joint suppression, with an exact conditional 95% confidence interval for the discordant-pair proportion. We do not apply formal hypothesis testing to the per-head comparisons: the expected number of decision flips per head under $\alpha = 0.1$ is at the noise floor of binary correctness on $N = 261$ samples,

and accuracy-level per-head tests at this scale lack the power to support either positive or negative conclusions. Per-head effects are therefore reported as descriptive effect sizes (percentage-point differences) only.

H.2 Results and Analysis

The baseline accuracy is 60.2%; joint suppression of all five negative heads yields 62.1%, a difference of $+1.9$ percentage points. The McNemar test gives $b_{10} = 2$ and $b_{01} = 7$ discordant pairs (exact $p = 0.18$), where b_{10} counts samples correct under baseline but incorrect under intervention and b_{01} counts the reverse. The discordant-pair imbalance favors the intervention (7 recovered vs. 2 lost) but does not reach significance at this sample size. Per-head descriptive effects are summarized in Table 7.

Table 7: Per-head sufficiency and necessity effects for the top-5 negative heads (Qwen3-32B, $\alpha = 0.1$, $N = 261$). Δ values are accuracy differences in percentage points, reported descriptively.

Head	Sufficiency Δ (pp)	Necessity Δ (pp)
L62H35	+0.4	+0.4
L63H20	0.0	+0.8
L58H12	+1.1	+0.8
L62H59	0.0	+0.8
L63H23	+0.4	0.0

Joint causal effect. The joint intervention shifts accuracy by $+1.9$ percentage points in the expected direction, consistent with the parameter-ablation sweep in the main text. The small discordant-pair count reflects that at $\alpha = 0.1$, the induced logit reweighting flips only borderline-decision samples while preserving confidently-correct predictions. The non-significant p-value is therefore power-limited rather than effect-limited, with the binary per-sample outcome too coarse to resolve the chosen intervention magnitude.

Per-head contributions. The per-head effect sizes range from 0 to 1.1 percentage points and sit at the noise floor of accuracy-level measurement on this set. We therefore do not interpret them as evidence of relative head importance. Qualitatively, the pattern is consistent with a distributed effect rather than a single dominant head: four of five heads show a non-negative LOO effect and three of five show a non-negative sufficiency effect.

Additivity. The sum of single-head sufficiency effects matches the joint effect to within measurement resolution ($+1.9$ vs. $+1.9$ percentage points). Because neither the individual nor the joint accuracy effects are statistically distinguishable from zero at this sample size, we report this agreement descriptively and refrain from drawing structural conclusions about additivity or inter-head interaction.

I Causal Contribution of Key Heads

The leave-one-out and single-head experiments in Section H establish the joint causal effect of the negative-head set at the accuracy level. Accuracy, however, collapses the output distribution into a binary signal and cannot resolve the relative contribution of individual heads at our sample size. We therefore complement those results with a path patching experiment (Wang et al., 2023), which yields a per-sample continuous measurement of each head's direct causal contribution to the logit difference.

I.1 Experimental Setup

Counterfactual construction. For each item in the 261-sample complex-task set, we construct a corrupted counterpart by swapping the content of options A and B in the prompt, with the correct-answer label flipping accordingly.

Three-pass protocol. For each (clean, corrupted) pair we execute three forward passes: (i) a *clean pass* on the clean prompt, caching the value-weighted output $\mathbf{z}_{l,h}^{\text{clean}}$ of every attention head at every token position and recording the logit difference $\text{ld}_{\text{clean}} = \text{logit}(t_{\text{correct}}) - \text{logit}(t_{\text{wrong}})$; (ii) a *corrupted pass* on the corrupted prompt, caching $\mathbf{z}_{l,h}^{\text{corr}}$ at every position and recording $\text{ld}_{\text{corrupted}}$; (iii) a *patched pass* on the clean prompt with the target head (l, h) overwritten at every token position, $\mathbf{z}_{l,h}^{(t)} \leftarrow \mathbf{z}_{l,h}^{\text{corr},(t)}$ for all $t \in \{1, \ldots, T\}$, while all other heads execute normally; we record $\text{ld}_{\text{patched}}$. The full-position protocol captures both direct contributions and any contribution mediated through earlier token positions, avoiding the misclassification of heads whose causal influence is routed through intermediate positions.

Path contribution metric. We define the path contribution score as the fraction of the full shift from clean to corrupted that is attributable to the patched head:

$$\text{PathContrib}_{l,h} = \frac{\text{ld}_{\text{patched}} - \text{ld}_{\text{clean}}}{\text{ld}_{\text{corrupted}} - \text{ld}_{\text{clean}}}. \qquad (12)$$

A value near $+1$ means the patched head alone reproduces the full corrupted effect; a value near

0 indicates negligible direct contribution. Under this metric, negative heads are expected to yield PathContrib < 0 and positive heads PathContrib > 0.

Valid-sample filtering. We retain only pairs for which both the clean and corrupted passes yield the expected predictions. This ensures the denominator in Eq. (12) is semantically meaningful. When the clean-to-corrupted shift fails to flip the model's prediction, the denominator no longer represents the causal axis we wish to decompose, and the resulting ratio is ill-defined.

I.2 Results and Analysis

Table 8 reports the path contribution scores for the top-10 negative and top-10 positive heads identified by ΔContribution in the main text, comparing the full-position protocol against a last-position-only ablation.

Table 8: Path contribution scores for key attention heads (Qwen3-32B). The candidate set is the top-10 negative and top-10 positive heads ranked by |ΔContribution| from the main-text analysis. Within each block, rows are ordered by Full Mean. "Full" reports the primary full-position results, "Last" a last-position-only ablation. Expected signs: negative for negative heads, positive for positive heads. "–" indicates the head was not included in the last-position ablation run.

Type	Head	Full Mean	Full Std	Last Mean	Sign
Negative	L63H20	−0.100	0.043	−0.096	✓
	L63H23	−0.085	0.033	−0.086	✓
	L62H35	−0.061	0.035	−0.055	✓
	L58H45	−0.040	0.046	−0.033	✓
	L62H36	−0.018	0.027	−0.017	✓
	L60H20	−0.015	0.026	−0.011	✓
	L62H38	−0.004	0.030	–	✓
	L62H62	+0.004	0.028	−0.001	×
	L62H59	+0.008	0.038	+0.003	×
	L58H12	+0.010	0.038	+0.009	×
Positive	L62H39	+0.090	0.057	+0.085	✓
	L63H16	+0.073	0.056	+0.075	✓
	L58H40	+0.059	0.061	–	✓
	L63H19	+0.047	0.030	+0.051	✓
	L63H40	+0.040	0.042	+0.035	✓
	L62H37	+0.027	0.035	+0.015	✓
	L63H43	+0.026	0.029	+0.023	✓
	L59H29	+0.020	0.040	+0.012	✓
	L62H34	+0.010	0.036	+0.004	✓
	L63H21	−0.030	0.037	−0.034	×

Causal confirmation of core heads. Seven of the top-10 negative heads and nine of the top-10 positive heads carry the expected sign, confirming the causal picture inferred from ΔContribution. Within the negative block, L63H20 (-0.100), L63H23 (-0.085), and L62H35 (-0.061) combine large magnitude with low coefficient of variation, identifying them as stable, sample-consistent causal contributors. A symmetric pattern holds for the positive block, where L62H39 ($+0.090$), L63H16 ($+0.073$), and L58H40 ($+0.059$) emerge as the corresponding correction-signal providers.

Robustness to patching protocol. The full-position and last-position-only columns yield nearly identical scores ($|\Delta| < 0.013$ across all rows where both are available, with most differences below 0.005). The causal signal of each key head is therefore written predominantly into the final-token residual stream, with negligible mediation through earlier positions, ruling out the possibility that the observed effects are artifacts of a particular patching choice.

Heads with reversed signs. In the path patching results, four heads exhibit signs opposite to those expected from the ΔContribution ranking: L62H62, L62H59, L58H12 (expected negative) and L63H21 (expected positive). At first glance this appears to negate their negative role, but the per-head intervention experiments at $\alpha = 0.1$ reported in Appendix H show otherwise. L58H12 and L62H59 both belong to the top-5 set used in the joint suppression experiment. Suppressing L58H12 alone yields a $+1.1$ percentage point accuracy gain and contributes $+0.8$ percentage points in the leave-one-out(LOO) test; suppressing L62H59 alone produces no measurable effect, yet it also contributes $+0.8$ percentage points in the LOO test. In other words, these two heads indeed act as negative contributors at the intervention level, whereas their mean values under the marginal measurement of path patching are near zero or even slightly positive.

We regard this discrepancy as non-contradictory; it reflects a fundamental difference in what the two attribution methods measure. Path patching perturbs a single head while holding the rest of the structure unchanged and measures the average effect on the correct-answer logit, which is a marginal measurement. When the output of a head is highly correlated with that of other negative heads, its marginal contribution is compensated for by the remaining heads in the set. Even if that head is removed, the model can maintain a similar bias signal through the correlated pathways, which suppresses the measured mean magnitude and can even push it to either side of zero due to sample noise. This exactly accounts for the observation that the path contribution magnitudes of L58H12 and L62H59

are below 0.011 and on the same order as their standard deviations.

In contrast, the joint suppression experiment at $\alpha = 0.1$ measures the effect of a joint intervention: when all the correlated negative heads are suppressed simultaneously, the compensatory pathways are cut off together, allowing the originally masked negative effect to emerge. This also explains the observation that suppressing the top-5 negative heads yields a larger gain than suppressing the top-3. If we follow the marginal magnitude ranking of path patching, the heads added beyond the top-3 (including L58H12 and L62H59) already have small contribution scores, so the marginal gain should diminish; yet the actual gain grows larger, indicating that these heads share a causal pathway with the top-3, mutually compensate when measured in isolation, and deliver their full effect only under joint intervention.

L63H21 represents a different scenario. It ranks among the top positive heads in the ΔContribution ranking, yet under both path patching protocols it consistently yields a negative mean of -0.030, and it does not belong to the negative set of the joint suppression experiment. The agreement between the two protocols rules out procedural artifacts and suggests that the correlation between ΔContribution and this head's output direction is more likely driven by shared upstream signals than by the head's own causal contribution direction.