Mitigating Legal Hallucinations via Symbolic Constraints and Analogical Precedents

Zixuan Huang[1], Yanxiang Ma[1], Luhan Wang[2], Yunke Wang[1], Duo Shi[3], Chang Xu[1]
[1]School of Computer Science, The University of Sydney, Australia,
[2]School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications, China,
[3]China Mobile Group Design Institute Co., Ltd., China
Correspondence: c.xu@sydney.edu.au

Abstract

With the growing potential of large language models (LLMs) in the legal domain, domain-specific finetuning and retrieval-augmented generation (RAG) methods have received widespread attention. However, current methods still suffer from hallucination risk and failing to resolve semantic drift and adapt to varying citation numbers. To address this, we propose Authoritative and Accurate Lawyer (AALawyer), a complementary dual-retriever framework based on the Legal Syllogism and the nature of different legal data. First, we introduce Symbolic Constrained Retrieval (SCR) for closed-set article retrieval, by constraining retrieval to the generative prediction. Second, we build Analogical Precedent Retrieval (APR) to retrieve open-set judicial precedents for reasoning with a newly collected large criminal dataset. Extensive experiments, including LawBench, our Hallucination Risk-Benchmark, and comprehensive ablation studies, demonstrate the effectiveness of AALawyer, which mitigates hallucinations while improving the explainability of legal reasoning. Code is available here.

1 Introduction

Recent studies have shown that LLMs can perform well in legal reasoning, text comprehension (Ma et al., 2021), and generation tasks (Zhong et al., 2020). However, legal texts are inherently challenging for general models due to complex terminology, specialized writing styles, and rigorous logical structures. As a result, LLMs often generate hallucinated outputs when applied directly to legal analysis tasks (Huang et al., 2025).

To mitigate hallucinations, current legal LLMs adopt the pipeline "Incremental Pretraining, Finetuning, and Retrieval-Augmented Generation (RAG)" (Huang et al., 2023; Yue et al., 2023; Zhou et al., 2024). Although their strategy improves factual foundation, one issue is the semantic gap between the abstract and the concrete in their dense retrieval, leading to Semantic Drift. We take a real-world "Intentional Injury" case retrieval result as an example, where the defendant used a crowbar to beat a victim during a village dispute . While the legal ground truth is *Article*[1] *234* (Intentional Injury), a standard dense retriever (Chen et al., 2024) falsely retrieves *Article 289* (Gathering crowds for beating, smashing, and looting) and *Article 235* (Negligent causing grievous injury). This failure occurs because the concrete description "using a crowbar to fight" shares high semantic overlap with the descriptive actions of "beating and smashing" in *Article 289*, overpowering the abstract legal concept in *Article 234*. Furthermore, the shared token of "injury" misleads the retriever to *Article 235*, ignoring the crucial logic between intentional and negligent, as well as the degree of injury.

Furthermore, determining the precise number of related articles remains a challenge. Legal LLMs rely on traditional static retrieval strategies (fixed top-k and thresholds) that fail to adapt to the various cases, which means a simple case may have a single related article, while a complex case relates multiples. While threshold-based filtering can alleviate this, it relies on heuristic hyperparameter tuning that may not generalize across diverse case types. They retrieve articles as a generic semantic matching task and fail to bridge the logic gap, often missing ground truths or introducing irrelevant noise to have a negative effect on downstream analysis (Yue et al., 2023; Huang et al., 2023).

To address these limitations, we propose AALawyer, a system that improves legal reasoning by aligning structure with the Legal Syllogism theory (MacCormick, 1994). We map the retrieval process to the core components of syllogism: the

[1]In this paper, the term "article" refers specifically to legal provisions or statutes (e.g., Article 234 of the Criminal Law), rather than academic publications, which is in line with conventional legal terminology.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13894–13920

Major Premise (Legal Norms), the Minor Premise (Fact Analysis) and the Conclusion. For the Major Premise, we introduce the Symbolic Constrained Retrieval (SCR) to strictly constrain the retrieval of legal norms. Complementarily, for the Minor Premise, we integrate Analogical Precedent Retrieval (APR) to retrieve similar precedent cases, providing reliable references to support the interpretation. This dual-RAG design effectively suppresses reasoning drift by both legal norms and factual analysis in verifiable sources.

AALawyer implements a dual-retrieval to balance different types of legal retrieval contents: (1) For closed-set articles : Leveraging the finite nature of legal articles, we introduce Symbolic Constrained Retrieval (SCR). Unlike traditional legal LLMs with retriever and generator, SCR parametrizes retrieval into the LLM backbone, sharing parameters to predict accurate article numbers. This design not only minimizes hallucination risk by grounding outputs in verifiable symbols but also enables adaptive citation, allowing the model to dynamically determine the number of cited articles based on case complexity. (2) For open-set reference cases: Since precedent retrieval involves an unbounded search space, we construct Analogical Precedent Retrieval (APR) using a newly collected large-scale criminal dataset. Recognizing that case analogy relies on descriptive similarity, this module employs dense retrieval to achieve semantically analogous judgments. This enhances explainability and transparency, providing legal practitioners with helpful and reliable reference cases similar to the input case, supporting in making final decisions.

We finetuned a legal LLM, Authoritative and Accurate Legal LLM (AA-LeLLM), as the backbone and integrated it with our dual-retrieval modules. Evaluations on extensive ablations, Lawbench and our proposed 4-dimensional HR-Benchmark demonstrate that AALawyer achieves state-of-the-art performance.

Our main contributions include the following.

- We propose a novel retrieve method SCR, formulating legal citation as a symbolic prediction task to mitigate the hallucination issues, semantic drift and achieve adaptive citation.

- We develop AALawyer, integrating SCR for rigorous article grounding and APR for retrieving analogous precedents to enhance comprehensiveness and explainability, which is supported by our newly constructed large-scale criminal case dataset.

- We conduct extensive experiments demonstrating that both our RAG strategies and the model achieve excellent improvement. Results on both public datasets and our proposed HR-Benchmark verify that AALawyer significantly reduces hallucination risks while improving analysis quality.

2 Related Works

2.1 Legal Large Language Models

The early applications of AI in the legal domain focused on information retrieval and extraction (Bommarito II et al., 2021; Ji et al., 2018), as well as legal prediction and question-resolution tasks (Ma et al., 2021; Ye et al., 2018; Yang et al., 2019; Kien et al., 2020; Zhong et al., 2020), greatly improve the efficiency of legal work.

To address domain-specific knowledge limitations and hallucination issues (Huang et al., 2025; Orgad et al., 2025; Rawte et al., 2023; Magesh et al., 2025; Colombo et al., 2024), legal domain LLMs emerged. In Chinese law, these include models based on incremental pretraining and multitask finetuning, such as LawGPT (Zhou et al., 2024), Lawyer LLaMA (Huang et al., 2023), Fuzi.Mingcha (Deng et al., 2023) and LexiLaw (Li et al., 2024); models leveraging finetuning combined with external information retrievals, such as DISC-LawLLM (Yue et al., 2023), ChatLaw (Cui et al., 2023), HanFei (He et al., 2023), and Wisdom-Interrogatory (Wu et al., 2024).

As the field continues to evolve, several standardized benchmarks (Fei et al., 2024; Yue et al., 2023) and legal datasets (Yao et al., 2022; Xiao et al., 2018; Yue et al., 2023; Deng et al., 2023) have been introduced to support this area.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) was first introduced by (Lewis et al., 2020). With the rise of LLMs, RAG has drawn increasing attention in addressing hallucination issues (Yao et al., 2023; Bang et al., 2023). In Gao et al. (2023), RAG is categorized into two types: Naive RAG and Advanced RAG. Naive RAG refers to the traditional Retrieve-and-Generate framework (Chen et al., 2017). In contrast, Advanced RAG achieves better performance by modifying the framework.

Among the Advanced RAG, Modular RAG breaks away from the traditional pipeline, providing some powerful pipelines for reducing hallucinations and enhancing generation quality. Instead, it introduces additional modules such as the Search (Wang et al., 2023), Memory (Cheng et al., 2023b; Wang et al., 2022), Extra Generation (Yu et al., 2023a), Task-Adaptable (Cheng et al., 2023a; Dai et al., 2023), Alignment (Yang et al., 2023; Yu et al., 2023b; Ma et al., 2023), and Validation Module (Yu et al., 2024). These approaches modify the pipeline in various ways, including Rewrite-Retrieve-Generate (Ma et al., 2023), Generate-Generate (Yu et al., 2023a), Recite-Generate (Sun et al., 2023), ITER-Retrieve-Generate (Yang et al., 2024; Shi et al., 2024), Retrieve-Validate-Generate (Yan et al., 2024). There are some models using a Generate-Generate pipeline. GENRE (De Cao et al., 2020) employs an auto-regressive language model to directly generate entity names as intermediate retrieval targets. SEAL (Bevilacqua et al., 2022) generates n-gram phrases that are likely to appear in relevant documents and uses them as lexical cues for downstream retrieval. GENREAD (Yu et al., 2023a) directly generates contextual documents for given questions.

3 Preliminary

3.1 Retrieval-Augmented Generation Models

In the legal domain, the document collection consists of law-related texts such as legal articles, court judgments, and case descriptions. Traditionally, Retrieval-Augmented Generation(RAG) models follow the retriever-and-generator RAG pipeline (Lewis et al., 2020). It first uses a retriever η to retrieve some documents z from the given document collection. Then, z is treated as the latent variable that is marginalized via the top-k approximation. The top-k documents occupy the major probability that can approximate the full probability over all documents.

The R&G RAG can approximate the posterior probability of the output sequence y. Following Lewis et al. (2020), the posterior probability is

$$p_{\text{R\&G}}(y \mid x) \approx \sum_{z \in \text{top-}k(p_\eta(\cdot|x))} p_\eta(z \mid x)\, p_\theta(y \mid x, z), \quad (1)$$

where x denotes the input, $y = (y_1, \ldots, y_N)$ denotes the output sequence generated by the generator model, $p_\eta(z \mid x)$ is the relevance scoring function of the retriever, and $p_\theta(y \mid x, z)$ is the probability that the generator predicts output y. θ is the generator.

Unlike traditional retriever-and-generator pipeline (Lewis et al., 2020), the generator-and-generator RAG (G&G-RAG) replaces the retriever with a generator ϕ. Instead of selecting z, the generator ϕ generates a set of representative intermediate candidates g, such as n-gram phrases, entity names, or contextual documents. The posterior probability approximated from G&G-RAG can be defined as,

$$p_{\text{G\&G}}(y \mid x) \approx \sum_{g \in \text{top-}k(p_\phi(\cdot|x))} p_\phi(g \mid x)\, p_\theta(y \mid x, g). \quad (2)$$

3.2 Large Language Model in Law

Existing legal LLMs (Huang et al., 2023; Yue et al., 2023; Zhou et al., 2024) are typically developed following the process of "Incremental Pretraining, Finetuning, and RAG". First, based on an open-source LLM θ, incremental pretraining is conducted on legal general-purpose datasets D_{law} and other general-domain datasets D_{general}. It is unsupervised training. Then, according to a specific legal task, the model is fine-tuned on the corresponding supervised dataset $D_{\text{law\_task}}$. After fine-tuning, model θ_0 obtains task-adapted parameters and is used as a generator in the RAG pipeline. Finally, a retrieve-and-generate RAG is employed to construct a prompt used in the final legal analysis.

However, existing approaches suffer from hallucination on incorrect citations caused by semantic drift and static retrieval limitations (as illustrated in Sec 1) and hallucinated content caused by generation noises (Huang et al., 2025; Yue et al., 2023; Huang et al., 2023).

4 Methodology

4.1 Symbolic Constrained Retrieval

To measure the hallucination, we first define the optimization objective. We use the accuracy $\text{Acc}(\theta)$ and the authenticity $\text{Auth}(\theta)$ to measure the hallucination risk. $\text{Acc}(\theta)$ denotes the accuracy on whether the model cites the correct article number, and $\text{Auth}(\theta)$ measures the correctness of the content in the cited article. Then we define the hallucination risk as

Definition 1. *The Hallucination Risk of the model θ is*

$$H(\theta) = 1 - Acc(\theta) \cdot Auth(\theta), \quad (3)$$

where $Acc(\theta), Auth(\theta) \in [0, 1]$, and detailed metrics are shown in Appendix A.1.1

There are two challenges depending on previous approach. Dense retrieval suffers from semantic drift due to the semantic gap. Moreover, generative approaches suffer from hallucinations due to the high-dimensional output space. We recognize that legal articles form a closed set, unlike the open-ended nature of general text. Therefore, we propose Symbolic Constrained Retrieval (SCR) to reformulate retrieval as a closed-set symbolic mapping. By training the model to map facts to symbols, we bridge the semantic gap to mitigate the semantic drift. By compressing the output into low-dimensional symbolic identifiers, we constrain the generation to eliminate content hallucination.

Firstly, we integrate the article number prediction task into a combined legal dataset for finetuning, enabling the model parameters to learn both downstream legal analysis objectives and article number retrieval. The model is an Authoritative and Accurate Legal LLM (AA-LeLLM) θ_0.

Secondly, we format the user's input x with a prompt for criminal article number prediction. Since articles have a finite number, we model this as a closed-set classification task rather than open retrieval. We restrict the distribution space of g to the finite set of valid symbolic identifiers N. As illustrated in Figure 1, the model predicts a sequence of identifiers $\hat{N}$. This symbolic prediction allows the model to leverage the internal legal logic learned during finetuning while maintaining a strict output constraint. Each predicted number $n_i \in \hat{N}$ can retrieve the corresponding law content c_i from the law article database DB_{law}. We assemble n_i and c_i as $a_i = (n_i, c_i)$ where $a_i \in A$.

Finally, the original input x and all retrieved articles A are incorporated into a second prompt for legal analysis. We switch the head of p_θ into a generation head to generate the final analysis output y. When utilized in generation tasks, equipped with the generation head, AA-LeLLM can be defined as θ_0^{gen}. We define the function of SCR as

$$p_{\text{SCR}}(y \mid x) \approx p_{\theta_0}^{\text{cls}}(\hat{N} \mid x)\, p_{\theta_0}^{\text{gen}}(y \mid x, A), \quad (4)$$

where

$$A = \left\{ (n_i, c_i) \mid n_i \in \hat{N},\ c_i = DB_{\text{law}}[n_i] \right\}.$$

Algorithm 1 Symbolic Constrained Retrieval

Input: Legal case input description: x; Legal Article Database: $\texttt{DB}_{\text{law}} = \{(n_i, c_i)\}_{i=1}^{m}$; Prompt for article number prediction: $\texttt{Format}_1$; Prompt for final analysis: $\texttt{Format}_2$; Pretrained model parameters: θ; General legal task dataset: $D_{\text{law\_other}}$; Specific legal task dataset: $D_{\text{law\_specific}}$

Output: Final legal analysis y

$D_{\text{law\_task\_all}} = D_{\text{law\_other}} \cup D_{\text{law\_specific}}$;
$\theta_0 = \texttt{Finetune}(\theta,\ D_{\text{law\_task\_all}})$;
$x_1 = \texttt{Format}_1(x)$;
$\hat{N} \sim p(n \mid x_1; \theta_0)$;
for all $n_i \in \hat{N}$ **do**
  $c_i = \texttt{Retrieve}(\texttt{DB}_{\text{law}}, n_i)$;
  $a_i = (n_i, c_i)$;
end for
$x_2 = \texttt{Format}_2(x,\ \{a_i\}_{i \in \hat{N}})$;
$\hat{y} \sim p(y \mid x_2; \theta_0)$;
return $\hat{y}$;

The details of SCR are shown in Algorithm 1.

SCR uses a single model θ_0 to handle both retrieval and output generation effectively. This enables a Generate-Retrieve-Generate pipeline, which differs from previous pipelines as shown in Figure 2. It assigns semantic meaning to the symbols in its parameters, effectively learning a compact representation for retrieval. This helps the dimension of the feature space to stay low. Intuitively, a lower-dimensional feature space is less likely to cause hallucinations. We formalize this intuition by deriving an upper bound on the risk,

Lemma 1. *Denote θ and θ_0 as the traditional LLM and SCR. Then the hallucination risk of SCR $H(\theta_0) = 1 - Acc(\theta_0)$ can be upper bounded as*

$$1 - Acc(\theta_0) \leq 1 - Acc(\theta) \cdot Auth(\theta), \quad (5)$$

where $1 - Acc(\theta) \cdot Auth(\theta) = H(\theta)$. The equality holds only when all generated article content is perfectly accurate and authentic, which cannot be achieved in practice. Proof is shown in the Appendix A.1.1.

This Lemma shows that replacing intermediate candidates with closed-set symbolic identifiers, the hallucination of SCR is guaranteed to be lower than G&G-RAG models.

4.2 Analogical Precedent Retrieval

While SCR handles the closed-set of legal articles, legal reasoning also requires referencing precedent

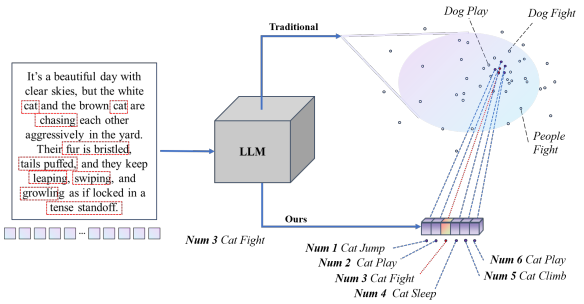


Figure 1: A conceptual comparison between the traditional generation-based method and our classification generation method in the output space.

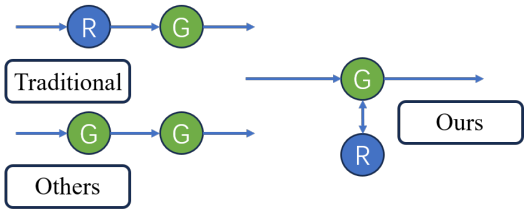


Figure 2: The pipeline comparison between traditional legal LLM RAG (corresponding to Eq. 1), generative RAG (corresponding to Eq. 2), and our proposed SCR.

cases to assist in the judgment's comprehensiveness and explainability. Unlike articles, judicial precedents form an open-set with millions of cases, making the classification approach of SCR impossible. However, since precedent retrieval involves mapping concrete descriptions to concrete descriptions, it does not suffer from the semantic drift in article retrieval. Therefore, we have designed the Analogical Precedent Retrieval (APR), a dense retrieval system that can provide relevant precedent court cases to augment the analysis.

APR needs to be supported by a large number of criminal cases. To address the lack of suitable data, we first collected a large-scale dataset DB_{case} comprising 176k criminal court cases from the public sources. Upon collecting the dataset, we use an embedding model (Xiao et al., 2024) to encode the criminal case dataset into a vector dataset $\{\mathbf{v}_i\}_{i=1}^{n}$, where each case $c_i \in \texttt{DB}_{\text{case}}$ corresponds to a vector $\mathbf{v}_i$. Then, the user input x is encoded into $\mathbf{v}_x$ and matched against the case vector dataset using Euclidean distance (Johnson et al., 2019) to retrieve the top-k most relevant cases $\{c_i\}_{i\in\mathcal{C}}$. Next, we construct the prompt x_1 by combining x, $\{c_i\}_{i\in\mathcal{C}}$, and (if available) A from SCR. Finally, the prompt x_1 is fed into the finetuned model p_θ to generate the final analysis y. Algorithm 2 outlines the APR retrieval process in the Appendix A.1.2.

4.3 Authoritative and Accurate Lawyer

In the above methods, SCR mitigates the hallucination risk by symbolic mapping, and APR provides the analogical references to make legal reasoning more comprehensive and professional. To integrate, we propose the AALawyer framework. As shown in Figure 3, the framework corresponding three parts of Legal Syllogism theory:

Stage 1: Major Premise (SCR) The input legal case e is fed into our backbone model AA-LeLLM to perform symbolic mapping, predicting the relevant article number(s) $\hat{N}$. These predicted article numbers are then used to retrieve the corresponding article content C from the SCR-Database DB_{law}. Finally, we integrate e, N, and C into the final input for downstream processing. This stage is shown in the figure in orange.

Stage 2: Minor Premise Enrichment (APR). We first encode the input case e into a dense vector representation. Then, we search our case vector library APR-Database DB_{case} to retrieve the top-k most relevant case documents. These retrieved documents are then appended to the final input as auxiliary context for enhanced reasoning. This stage is shown in the figure in blue and red.

Stage 3: Conclusion. The final input, comprising the original case e, the predicted article number(s) N, the article content C, and the top-k similar cases, is fed into AA-LeLLM θ_0 to generate the final legal case analysis. In this stage, AA-LeLLM explains the answer with the retrieved relevant articles and analogous previous judgments, thereby providing professional and explainable legal reasoning with low hallucination. This stage is shown in the figure in green.

5 Hallucination Risk-Benchmark

Existing legal benchmark evaluations mostly focus on assessing the model weights alone, lacking a method to evaluate overall system performance. Meanwhile, usual human expert evaluations are too resource-consuming. In order to evaluate the RAG and the effectiveness of our entire system, we leverage other LLMs (Guo et al., 2025) for scoring. Models with larger parameters have stronger logical judgment and analytical professionalism, and can simulate the evaluation of our answers by experts in the legal field, which can be demonstrated in (Zheng et al., 2023; Yue et al., 2023).

We build the Hallucination Risk-Benchmark (HR-Benchmark) to measure the overall perfor-

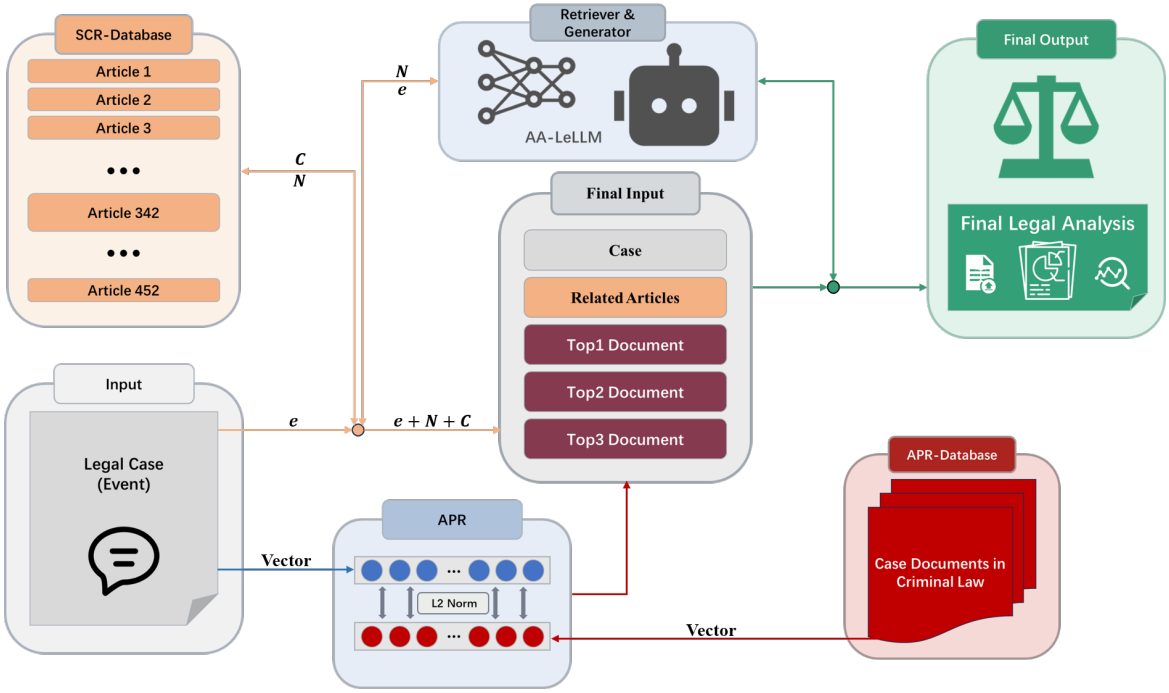


Figure 3: Overall architecture of the AALawyer, consisting of SCR (orange line), APR (blue and red line), and the AA-LeLLM generation stage (green line).

mance of the system by focusing on the measurement of modeling hallucination and other aspects. In detail, we evaluate the generated answers from four dimensions they are hallucination score **Hallu**, professionalism **Prof**, informativeness **Info**, and explainability **Expa**. **Hallu**↓ is the Hallucination Risk defined in Definition 1. With a lower hallucination score, the model has better performance. **Prof** is the score of the legal professionalism of the analysis. **Info** assesses the richness of information in the analysis. **Expa** refers to the transparency of legal analysis, whether it is supported by exact materials. We also calculate the average score (**Avg.**) to measure the overall performance of the model. The average score is calculated by averaging the value of $(1-$ **Hallu**$)$ and the other three scores. The scoring range is defined from 0% to 100%.

In the evaluation set, we randomly selected 200 cases from the dataset of CAIL2018 (Xiao et al., 2018). To avoid instability in evaluation caused by randomness, we set $temperature$ to 0. When using the API, the evaluations must be conducted at the same period to prevent any potential model updates, ensuring that the evaluation is performed using the same model with identical parameters.

6 Experiment

6.1 Experiment Setup

We trained the model by incremental pretraining and finetuning, added two RAG processes, and evaluated its performance on Lawbench and HR-Benchmark. The details of datasets and metrics are in the Appendix. All tasks were run on four NVIDIA GeForce RTX 4090 GPUs.

Incremental Pretraining. The rank r is set to 8, the scaling factor α is set to 16, and $dropout$ is set to 0. LoRA adaptation is applied across all Transformer layers. The optimizer is AdamW with a learning rate lr of 5×10^{-5}, and a cosine learning rate scheduler is employed. Gradient accumulation is performed with 8 steps and a batch size of 2 per device, while the cut-off length is set to 2048. The model training is conducted with bfloat16 precision.

Finetuning. The per-device batch size is set to 1, while all other settings remain consistent with Incremental Pretraining. We choose DeepSeek-7B as the base model for AA-LeLLM. The model is incrementally pretrained and finetuned using multitask datasets for both legal classification and analysis tasks. The details are shown in the Appendix.

Evaluation. In Lawbench, we set the maximum truncation length for inference to 2048 with runs on 500 samples. In HR-Benchmark, $temperature$

Model	Criminal		Full Domain				Average
	FAP	CP	DFI	MDI	ITI	ED	
DeepSeek-7B (Guo et al., 2025)	16.86	28.89	27.20	39.69	35.80	58.35	34.47
LawGPT-beta1.1-7B (Zhou et al., 2024)	0.15	15.68	4.95	6.85	2.40	14.94	7.50
LexiLaw-6B (Li et al., 2024)	13.15	39.99	3.30	15.60	22.80	15.30	18.36
HanFei-7B (He et al., 2023)	2.64	30.96	6.39	30.44	30.20	14.73	19.23
Wisdom-Interrogatory-7B (Wu et al., 2024)	32.84	35.09	7.84	36.72	21.00	15.98	24.91
Fuzi-Mingcha-7B (Deng et al., 2023)	25.19	55.93	19.59	28.46	18.60	16.90	27.45
Qwen3-8B (Team, 2025)	73.21	51.88	46.00	54.40	38.80	65.86	55.02
Internlm3-8B (Cai et al., 2024)	82.92	55.02	37.20	52.29	41.20	66.38	55.83
AA-LeLLM-7B(Ours)	88.84 $\pm$ 0.34	57.93 $\pm$ 0.60	20.10 $\pm$ 0.50	46.29 $\pm$ 1.19	39.00 $\pm$ 1.52	81.36 $\pm$ 0.07	55.59

Table 1: Scores of models with comparable parameter sizes on LawBench classification tasks.

Category	Method	Parameter	F1-Score	Precision	Recall
Dense Retrieval	BGE-v1.5 (Xiao et al., 2024)	$k^\dagger = 1$	24.26	26.60	18.19
		$\tau^\dagger = 0.97$	16.73	10.92	31.33
	BGE-m3 (Chen et al., 2024)	$k^\dagger = 1$	52.68	59.40	40.63
		$\tau^\dagger = 0.56$	36.77	25.89	55.95
	SAILER (Li et al., 2023)	$k^\dagger = 1$	63.63	72.20	49.38
		$\tau^\dagger = 0.63$	48.49	34.10	59.10
Sparse Retrieval	BM25	$k^\dagger = 1$	35.19	40.20	27.50
		$k = 2$	30.36	25.50	34.88
		$k = 3$	26.37	19.33	39.67
Generative Retrieval	SCR (Qwen3-8B)	$\tau^*_\theta(x)$	73.20	67.70	74.83
	SCR (Internlm3-8B)	$\tau^*_\theta(x)$	82.92	43.81	78.39
	SCR (AALawyer)	$\tau^*_\theta(x)$	**88.94**	**89.67**	**84.26**

Table 2: Performance comparison of different retrieval methods on the FAP task. Results are shown for different parameter: top-k (k) and thresholds (τ). The dagger (†) indicates the optimal value for the parameter. The optimal parameter selections are shown in the Appendix.

is 0. Judges are $Deepseek\text{-}Chat$ and $Qwen\text{-}Plus$ and running on 200 randomly selected samples.

6.2 Main Results

As shown in Table 1, the Criminal Article Classification task (FAP), achieves a significant improvement of 71.98% over the base DeepSeek-7B model, reaching the best classification performance among all compared models. Similarly, the Criminal Case Classification task (CP) also improves by 29.04%. For full-domain legal classification tasks, Marital Disputes Identification (MDI) improves by 6.60%, Issue Topic Identification (ITI) by 3.20%, and Event Detection (ED) by 23.01%. Although Dispute Focus Identification (DFI) drops by 7.10%, this is expected since it belongs to the full legal domain and is not aligned with our training goal, which is specifically focused on criminal law. Even so, the performance remains within a reasonable range.

To explicitly evaluate the retrieval part, we compare different retrieval paradigms in Table 2. Traditional dense retrievers in legal LLMs exhibit lower precision even in $k = 1$, indicating that they often retrieve wrong articles due to the semantic drift between abstract and concrete. In contrast, our SCR achieves 89.67% precision and 84.26% recall. This demonstrates that by reformulating retrieval as a symbolic mapping task, AALawyer effectively eliminates the semantic gap in article retrieval.

In Table 3, compared with our base model DeepSeek-7B, our AALawyer reduces hallucination risk by 37.6%, improves professionalism by 13.4%, informativeness by 41.9%, and explainability by 43.4%, and achieves an overall average score increase by 34.1%. Furthermore, compared with our AA-LeLLM without RAG, our AALawyer reduces the hallucination risk by 12.2%, improves 10.3% professionalism, 57.1% informativeness, and 47.3% explainability, and achieves an overall increase in the average score of 31.7%.

Overall, AA-LeLLM demonstrates excellent performance on our target classification task, FAP, enabling SCR to resolve semantic drift and hallucination issues while maintaining high accuracy, resulting in effective legal reasoning. Moreover, AA-LeLLM also shows competitive performance on other non-target classification tasks in various datasets, achieving an average score improvement of 21.12% compared to the base model. This high-

Baseline	Method	Hallu↓	Prof	Info	Expa	Avg.
DS-7B	Vanilla	82.4	58.8	40.0	44.6	40.2
	SCR	85.4	46.4	39.4	44.2	36.2
	APR	63.0	66.6	81.6	78.8	66.0
	AALawyer	81.4	55.0	74.0	70.8	54.6
AA-LeLLM	Vanilla	57.0	61.9	24.8	40.7	42.6
	SCR	**39.6**	70.6	37.9	67.8	59.1
	APR	53.9	60.9	74.7	79.5	65.3
	AALawyer	44.8 $\pm$ 1.8	**72.2** $\pm$ 0.2	**81.9** $\pm$ 0.2	**88.0** $\pm$ 0.2	**74.3**

Table 3: Comparison of our AA-LeLLM against DeepSeek-7B (DS-7B) on HR-Benchmark.

Baseline	Method	Avg.Acc	Avg.Auth	H(θ)↓
DS-7B	Vanilla	0.29	0.44	0.88
Qwen3-8B	Vanilla	0.59	0.70	0.58
	SCR	0.59	0.94	0.44
Internlm3-8B	Vanilla	0.71	0.69	0.51
	SCR	0.71	0.95	0.33
AA-LeLLM	Vanilla	0.76	0.49	0.63
	SCR	0.76	0.93	**0.29**

Table 4: Hallucination risk calculation by metrics.

lights its robustness and effectiveness as a legal classifier. It serves as a qualified module for the retriever component of SCR in our framework. In the system, SCR and APR both show strong performance on HR-Benchmark, SCR mainly improves **Hallu**↓, **Prof** and **Expa**, while APR mainly improves **Info** and **Expa** of legal reasoning. The low **Hallu**↓ of SCR also proves Lemma 1.

6.3 Ablation Studies

6.3.1 Adaptive Citation Capability

A single case may correspond to multiple legal articles due to the inherent complexity of legal cases. Traditional multi-label classification methods often rely on fixed thresholds τ as $\hat{A} = \{j \mid \hat{y}_j \geq \tau\}$ or by selecting the top-k scoring labels as $\hat{A}_{\text{top-}k} = \arg\text{top-}k_{j=1,\ldots,K}\ \hat{y}_j$. Both have problems. As shown in Table 2, when $\hat{y}_n \approx \tau$, it is unclear whether that label should be included, and the optimal value of τ can vary by case. Alternatively, top-k strategies avoid threshold tuning but might include low-confidence labels, risking irrelevant results. In contrast, SCR treats citation as an symbolic generation task. The model naturally learns when to stop generating symbols via the $< EOS >$ token, allowing it to decide how many symbolics to output without relying on rigid decision boundaries, thus avoiding the pitfalls of traditional methods.

This adaptive threshold enables more reliable predictions across diverse inputs, as

$$\hat{N} = \{n \mid p_\theta^{\text{cls}}(x)[n] \geq \tau_\theta^*(x)\}. \tag{6}$$

6.3.2 Calculation of Hallucination Risk

To further prove our Lemma 1, we selected 150 cases from the dataset of CAIL2018 (Xiao et al., 2018) and calculate the hallucination risk H(θ) on the latest criminal law as

$$\text{H}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(1 - \text{Acc}_i(\theta) \cdot \text{Auth}_i(\theta)) \tag{7}$$

The results are shown in Table 4. We calculate the average score of $\text{Acc}_i(\theta)$ and $\text{Auth}_i(\theta)$, using the metric

$$\text{Auth}_i(\theta) = \frac{(1+\beta^2)\cdot \text{LCS}(g_\theta(x_i), c_{y_i})}{\text{len}(g_\theta(x_i)) + \beta^2 \cdot \text{len}(c_{y_i})}, \tag{8}$$

where we set $\beta = 0$, and

$$\text{Acc}_i(\theta) = \frac{2 \cdot |f_\theta(x_i) \cap y_i|}{|f_\theta(x_i)| + |y_i|}, \quad f_\theta(x_i), y_i \subseteq A \tag{9}$$

The Hallu↓ is calculated according to Formula 7, rather than directly from $Avg.Acc$ and $Avg.Auth$. The result shows that our SCR has a score of 0.93 in $Avg.Acc$, indicating that it effectively mitigates the Auth(θ) component of the hallucination risk. And SCR reduces hallucination risk by 59% compared with base model, which shows an excellent performance and achieves the goal of our SCR. Furthermore, these metric is grounded in a mathematically defined hallucination risk formula, enabling a more objective and reliable assessment of hallucination. The results are corresponding to the result tables in Main Result part, which also validate the reliability of our proposed HR-Benchmark.

6.3.3 Performance on Long-tail Problem

We conducted a further evaluation on the long-tail data. We identified legal articles appearing fewer than 154 times from the 154,592 samples in the training set as long-tail articles. We then evaluated the F1-score on the corresponding long-tail subset (5.2%) of the LawBench test set, with the results presented in Table 6.

The base model without finetuning suffers from severe long-tail issues. And the dense retriever SAILER, which is finetuned on the legal corpus, only achieves 0.5385. In contrast, our model maintains robustness, reaching a high score of 0.7436. This demonstrates the effectiveness of our training method and dimensionality reduction.

Confidence Threshold (τ)	Retention Rate	Avg. F1	F1 Gain
Baseline	100.0%	0.8962	-
$\tau \geq 0.50$	99.8%	0.8980	+0.18%
$\tau \geq 0.80$	99.6%	0.8998	+0.36%
$\tau \geq 0.90$	98.0%	0.9037	+0.75%
$\tau \geq 0.95$	87.0%	0.9306	+3.44%
$\tau \geq 0.98$	68.6%	0.9445	+4.83%

Table 5: Performance verification of the SCR module under different confidence thresholds.

Model	Overall Data	Long-tail Data
DS-7B	0.1686	0.0256
SAILER	0.6363	0.5385
SCR	0.8894	**0.7436**

Table 6: Performance comparison on overall and long-tail datasets.

Dimension	Pearson (r)	p-value
Hallucination **Hallu**↓	0.8981^{***}	< 0.001
Informativeness **Info**	0.6661^{***}	< 0.001
Professionalism **Prof**	0.6503^{***}	< 0.001
Explainability **Expa**	0.7240^{***}	< 0.001
Average Score	$\mathbf{0.9074}^{***}$	$< \mathbf{0.001}$

Table 7: Meta-evaluation of correlation between human and LLM-as-a-judge.

6.3.4 RAG without Finetuning

Table 3 shows the non-finetuned base model DeepSeek-7B performance with each RAG. The results are consistent with our prior theory. SCR performs poorly on DeepSeek-7B because it needs to be used combined with special finetuning process, corresponding to Algorithm 1. And it shows a good performance with APR, because APR is supported by a huge and professional criminal law dataset, making it highly effective for criminal law tasks regardless of model weights. This demonstrates that it is a robust and adaptable RAG method in the criminal law domain.

6.3.5 Verification Step via Confidence Scores in SCR

We conducted an experiment on the verification step as shown in Table 5. We extracted the confidence scores (average token log-probabilities) and found that confidence is indeed correlated with prediction accuracy. By selecting different thresholds to exclude predictions, we can obtain higher accuracy in the final application. There is a trade-off between the Retention Rate and the Gain here. The trade-off curve is shown in Figure 4.

6.3.6 Human-expert Meta-evaluation on HR-Benchmark

Regarding the subtle legal hallucinations that may caused by LLMs as judges, we conducted a small-scale "meta-evaluation". We randomly sampled 20% of the HR-Benchmark test results on AALawyer and invited annotators holding formal degrees in law to participate in manual, blind evaluations. We calculated the Pearson correlation coefficient (r) between human scores and LLM scores.

As shown in Table 7, the average score achieved a Pearson correlation coefficient of 0.9074. Additionally, all individual dimensions showed highly significant positive correlations ($p < 0.001$). Even for the subjective and complex dimensions, they all reached a correlation greater than 0.65. This demonstrates the reliability of our automated metrics in reflecting the real judgments of human experts.

More ablations are shown in the Appendix A.2.

7 Conclusion

In this work, we addressed the hallucination risk, semantic drift and static retrieval limitations in legal LLMs. We proposed AALawyer, a framework inspired by Legal Syllogism theory combined with a dual-retriever module. We introduced SCR, which fundamentally reformulates article retrieval as a generative prediction task in a closed set, ensuring logical consistency. Furthermore, we complementarily designed an APR for open-set case retrieval by introducing a new large-scale criminal dataset. To validate our approach, we also introduced a 4-dimensional Hallucination Risk-Benchmark. Extensive experiments demonstrate that AALawyer achieves superior performance, providing reliable legal reasoning.

Limitations

There are several limitations in our current work.

First, model scale and data constraints. Due to the computational constraints, our experiments were conducted on a 7B model within the Criminal Law domain. While the 7B scale is sufficient to demonstrate the effectiveness of our dual-retriever framework, larger-scale models could potentially yield superior reasoning performance. Furthermore, we only focus on criminal law means that there is the potential to explore our methods to other legal domains, such as Civil or Administrative Law.

Second, dependency on finetuning. Our ablations indicate that the SCR module relies on domain-specific finetuning to achieve symbolic mapping. Updating parametric knowledge is more complex than non-parametric methods. When real-world legal statutes are amended, the framework may require online finetuning or continual learning to update the legal knowledge base.

Third, the scale of the HR-Benchmark test set. We chose 200 cases due to constraints on resources. Although our experiments demonstrate that this subset highly conforms to the distribution of naturally collected legal datasets, a larger scale can further solidify the evaluation.

Ethical Considerations

This work involves the collection and analysis of publicly available legal data. We strictly adhere to data privacy principles. All personally information regarding victims, witnesses, and other sensitive parties in the dataset has been anonymized to protect privacy.

Furthermore, we emphasize that the proposed system is designed solely for research purposes and as an learning assistant for legal practicers. It should not be used as a substitute for human legal judgment or consultation. We bear no responsibility for any consequences resulting from the use of this model in real-world legal decision-making.

Acknowledgments

This work was supported in part by the Australian Research Council under Projects DP240101848 and FT230100549.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems*, 35:31668–31683.

Michael J Bommarito II, Daniel Martin Katz, and Eric M Detterman. 2021. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. In *Research handbook on big data law*, pages 216–227. Edward Elgar Publishing.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 1870–1879. Association for Computational Linguistics (ACL).

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023a. Uprise: Universal prompt retrieval for improving zero-shot evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12337.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2023b. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36:43780–43799.

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, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

N De Cao, G Izacard, S Riedel, and F Petroni. 2020. Autoregressive entity retrieval. In *ICLR 2021-9th International Conference on Learning Representations*, volume 2021. ICLR.

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.

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.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2.

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.

Wanwei He, Jiabao Wen, Lei Zhang, Hao Cheng, Bowen Qin, Yunshui Li, Feng Jiang, Junying Chen, Benyou Wang, and Min Yang. 2023. Hanfei-1.0.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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

Lu Ji, Zhongyu Wei, Xiangkun Hu, Yang Liu, Qi Zhang, and Xuan-Jing Huang. 2018. Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3703–3714.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. Sailer: structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1035–1044.

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

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.

Neil MacCormick. 1994. *Legal reasoning and legal theory*. Clarendon Press.

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*, 22(2):216–242.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM_Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7339–7353.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3170–3179.

Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. 2023. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv preprint arXiv:2308.11761*.

Yiquan Wu, Yuhang Liu, Yifei Liu, Ang Li, Siying Zhou, and Kun Kuang. 2024. wisdominterrogatory.

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.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv e-prints*, pages arXiv–2401.

Diji Yang, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. 2024. Im-rag: Multi-round retrieval-augmented generation through learning inner monologues. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 730–740.

Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023. Prca: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5364–5375.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4085–4091.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. Leven: A large-scale chinese legal event detection dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 183–201.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023a. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 14672–14685.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023b. Augmentation-adapted retriever improves generalization of language models as generic plug-in. *arXiv preprint arXiv:2305.17331*.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *Preprint*, arXiv:2309.11325.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, pages 9701–9708.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model. *Preprint*, arXiv:2406.04614.

A Appendix

A.1 Theoretical Analysis

A.1.1 Proof of Lemma 1

By definition, the hallucination risk for the traditional model θ is

$$H(\theta) = 1 - \text{Acc}(\theta) \cdot \text{Auth}(\theta).$$

$$\text{Acc}(\theta) = \frac{2 \cdot |f_\theta(x) \cap y|}{|f_\theta(x)| + |y|}, \quad f_\theta(x) \subseteq A, \quad y \subseteq A$$

$$\text{Auth}(\theta) = \frac{(1+\beta^2) \cdot \text{LCS}(g_\theta(x), c_y)}{\text{len}(g_\theta(x)) + \beta^2 \cdot \text{len}(c_y)}$$

Thus, our objective to reduce hallucination risk is

$$\min_\theta \quad H(\theta) = 1 - \text{Acc}(\theta) \cdot \text{Auth}(\theta).$$

Both $\text{Acc}(\theta)$ and $\text{Auth}(\theta)$ must be increased to reduce $H(\theta)$.

The overall SCR prediction process can be expressed as

$$p_{\text{SCR}}(y \mid x) \approx p_{\theta_0}^{\text{cls}}(\hat{N} \mid x)\, p_{\theta_0}^{\text{gen}}(y \mid x, A)\,,$$

where

$$A = \left\{ (n_i, c_i) \mid n_i \in \hat{N},\ c_i = DB_{\text{law}}[n_i] \right\}.$$

For our SCR model θ_0, because the article content c_i is retrieved from a ground-truth database rather than generated, we can assume the authenticity is perfect, $\text{Auth}(\theta_0) = 1$. The risk for our model thus simplifies as

$$H(\theta_0) = 1 - \text{Acc}(\theta_0).$$

To prove the lemma, $H(\theta_0) \leq H(\theta)$, we need to show that:

$$1 - \text{Acc}(\theta_0) \leq 1 - \text{Acc}(\theta) \cdot \text{Auth}(\theta)$$

$$\text{Acc}(\theta_0) \geq \text{Acc}(\theta) \cdot \text{Auth}(\theta). \tag{10}$$

We now justify this premise by analyzing the optimization objectives.

Therefore, the original 2-dimensional optimization is reduced to a single-dimensional objective, shown as

$$\min_{\text{Acc, Auth} \in [0,1]} \quad H(\theta) = 1 - \text{Acc} \cdot \text{Auth}$$
$$\Longrightarrow \quad \min_{\text{Acc} \in [0,1]} \quad H(\theta_0) = 1 - \text{Acc}.$$

So the optimized gradient

$$\begin{aligned} \nabla_\theta H(\theta) &= \frac{\partial H(\theta)}{\partial \text{Acc}(\theta)} \cdot \frac{\partial \text{Acc}(\theta)}{\partial \theta} \\ &\quad + \frac{\partial H(\theta)}{\partial \text{Auth}(\theta)} \cdot \frac{\partial \text{Auth}(\theta)}{\partial \theta} \\ &= -\Bigg(\text{Auth}(\theta) \cdot \frac{\partial \text{Acc}(\theta)}{\partial \theta} \\ &\quad + \text{Acc}(\theta) \cdot \frac{\partial \text{Auth}(\theta)}{\partial \theta} \Bigg) \end{aligned} \tag{11}$$

becomes

$$H(\theta_0) = 1 - \mathrm{Acc}(\theta_0)$$

$$\Rightarrow \quad \nabla_{\theta_0} H(\theta_0) = -\frac{\partial \mathrm{Acc}(\theta_0)}{\partial \theta_0}.$$

The training of the traditional model θ involves a two-dimensional optimization, where the gradient is a composite signal from two potentially conflicting objectives to improve accuracy and authenticity simultaneously.

In contrast, the training of our model θ_0 is a simpler, single-dimensional problem focused solely on maximizing $\mathrm{Acc}(\theta_0)$. The gradient is a more direct signal. Given that the optimization of θ_0 is more focused and stable, it is reasonable to conclude that it can more effectively achieve a final accuracy $\mathrm{Acc}(\theta_0)$ that surpasses the $\mathrm{Acc}(\theta) \cdot \mathrm{Auth}(\theta)$ which has the more complex training process. Therefore, the premise Equation 10 is justified, which completes the proof.

A.1.2 Algorithm of APR

Shown in Algorithm 2.

Algorithm 2 Analogical Precedent Retrieval

Input: User input description: x; Legal case database: $\mathtt{DB}_{\text{case}} = \{c_1, c_2, \ldots, c_n\}$; All legal case vectors: $\{\mathbf{v}_i\}_{i=1}^n$; Encoder: $\mathtt{Enc}(\cdot)$; Prompt formatter: $\mathtt{Format}$; Generator: $p_\theta(y \mid \cdot)$; Top-k: k; Maximum token length: $T_{\max} = 2048$; *(Optional)* Relevant legal article set: $A = \{(n_i, c_i)\}_{i=1}^m$

Output: Final legal analysis $\hat{y}$

$\mathbf{v}_x = \mathtt{Enc}(x)$;
$\mathcal{C} = \text{k-argmin}_{i=1}^n \|\mathbf{v}_x - \mathbf{v}_i\|_2$;
for all $i \in \mathcal{C}$ **do**
  $c_i = \mathtt{Retrieve}(\mathtt{DB}_{\text{case}}, i)$;
end for
$x_1 = \mathtt{Format}(x, \{c_i\}_{i \in \mathcal{C}}, A)$;
$\hat{y} \sim p_\theta(y \mid x_1)$, s.t. $\mathtt{TokenLen}(x_1) \leq T_{\max}$;
return $\hat{y}$;

A.2 Additional Experiments

A.2.1 Cross-Model Verification on HR-Benchmark

To eliminate the self-preference bias, we conduct cross-verification evaluation evaluation in Table 10. We use another state-of-the-art general LLM, *Qwen-Plus*, as judge to evaluate the Table 3 again.

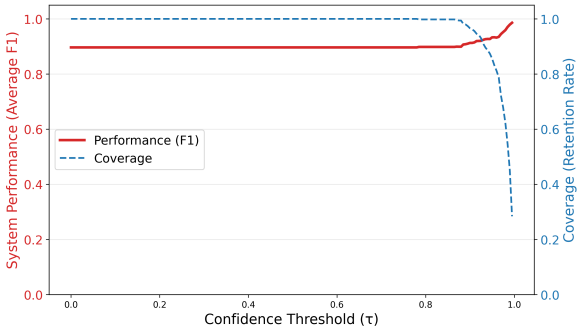


Figure 4: Performance curve of SCR under different confidence thresholds.

Compared with the evaluation conducted by *Deepseek-Chat*, We observe that the evaluation trends are highly consistent with the analysis in our main results and Section 6.3.4.

On our finetuned AA-LeLLM, the SCR module significantly reduces the Hallu↓ to 28.0%, and AALawyer framework achieves the highest overall score 79.3%. This confirms that our method's performance improvement is robust and stable across different judges.

As shown in the DS-7B, applying SCR without finetuning leads to a performance drop and APR improve overall performance. This perfectly aligns with our conclusion in Section 6.3.4 that SCR relies on the symbolic mapping capability acquired through finetuning, while APR serves as a plug module are still suitable to non-finetuned models.

These part demonstrate that the each part of AALawye rather than self-preference bias.

A.2.2 Analysis on Inference Constraints

To enforce the "closed-set" constraint during SCR inference, our primary approach relies on soft constraints learned via training likelihoods (as detailed in Appendix A.4.2). To investigate whether explicitly masking invalid tokens during decoding could further improve robustness, we conducted an additional experiment employing a hard constraint via a logits processor.The results showed that both the soft constraint and the hard constraint yielded an identical F1-Score of 89.65. This finding indicates that our shared-head training mechanism effectively captures the task distribution, inherently suppressing invalid outputs without the need for manual logits manipulation during inference. Nevertheless, implementing such a hard constraint can provide a theoretical guarantee of validity for practical deployment.

A.2.3 Sample Scale and Representativeness of HR-Benchmark

We chose the 200 cases because of the limited assessment resources (api). We further conducted a distribution correlation test between our 200-sample test dataset on HR-Benchmark and the CAIL2018 validation set (parent set) of 17,131 samples. For the frequency distribution of relevant articles, the cosine similarity is 0.9090 and Pearson correlation is 0.8848 ($p < 0.001$, exact $p = 6.2018 \times 10^{-62}$). This shows that our test set distribution is not artificially selected, and highly conforms to the distribution of the naturally collected dataset. Additionally, the Table 8 demonstrates that the proportion of major criminal charges is also highly consistent. The data distributions are compared in Figure 7.

A.2.4 Comparison of End-to-end Inference Latency

We conducted a comprehensive comparison on the end-to-end inference latency on a single NVIDIA GeForce RTX 4090 GPU. The efficiency results are in Table 9.

Although SCR slightly increases the inference time compared to single-step RAG (BGE-M3) due to an extra inference step, this cost is worthwhile for the significantly high F1 score of 88.94 in Table 2.

A.2.5 Selection of optimal threshold

As shown in Figure 6, we evaluate the F1-score for each threshold of the RAG models, and finally present the optimal values in the main evaluation. The optimal thresholds were tested to be 0.97 for BGE-v1.5, 0.56 for BGE-m3 and 0.63 for SAILER.

A.2.6 APR scale

To evaluate the retrieval effectiveness of our large-scale legal case dataset within the APR framework, we designed an ablation study to measure performance across varying data scales. We find that original score of vectors similarity S_{ori} will produce high scores even for irrelevant queries, so we test the base noise ϵ and calculate a Actual Score S_{act} as

$$S_{\text{act}} = \frac{S_{\text{ori}} - \epsilon}{1 - \epsilon} \tag{12}$$

To determine our base noise ϵ, we measured the system performance using 10 meaningless text strings. We calculated the average top-10 similarity score for each string over 100 independent runs. This test yielded a mean score of 0.6652 ± 0.0196. To ensure a effective (positive) result in our subsequent calculations, we defined ϵ as the lower bound of this noise range: $\epsilon = 0.6456$.

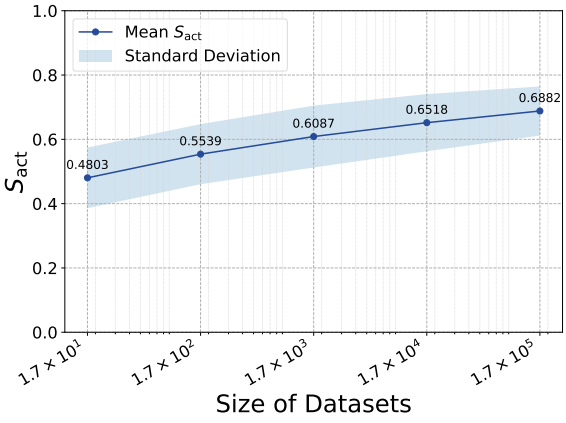


Figure 5: The effectiveness of our large-scale datasets.

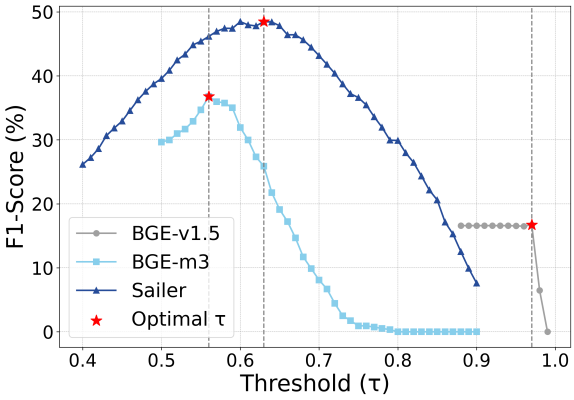


Figure 6: The optimal thresholds of RAG models (with a fixed $top\text{-}k = 5$).

Then we measured S_{act} by running the experiment 20 times for 5 law cases randomly selected from the China Court Website. This evaluation was performed on subsets of our entire dataset, randomly sampled at scales of 17, 170, 1,700, 17,000, and 170,000 documents at each running time, with the average top-10 S_{act} being calculated. The results in Figure 5 shows that a larger dataset leads to the retrieval of cases that are more similar to the query, which demonstrate the effectiveness of our large-scale datasets in APR.

To ensure cross-court and cross-time generalization, APR is supported by an extensive corpus. This dataset comprises 176k real-world cases from 2001 to 2020 and covering 25 provinces, providing sufficient diversity for robust analogical retrieval.

Major Criminal Charges	Valid Set (17k)	HR-Benchmark (200)
Theft (Top-1)	5.4%	5.0%
Drug Trafficking/Smuggling (Top-2)	4.8%	4.5%
Intentional Injury (Top-3)	3.4%	3.6%

Table 8: Distribution comparison of top-3 major criminal charges between the original CAIL2018 validation set and our sampled HR-Benchmark.

Method	Average Latency (s/query)
BGE-M3	1.7717
SCR (Ours)	2.2718
BGE-M3 + APR	3.9388
SCR + APR (Ours)	4.5619

Table 9: End-to-End inference latency comparison on the 200-case HR-Benchmark subset.

Baseline	Method	Hallu↓	Prof	Info	Expa	Avg.
DS-7B	Vanilla	87.5	65.0	56.5	56.6	50.2
	SCR	88.8	51.7	50.9	46.2	42.5
	APR	44.8	83.5	82.0	82.3	75.8
	AALawyer	75.5	60.4	68.4	58.8	53.0
AA-LeLLM	Vanilla	42.1	81.9	52.9	64.3	64.3
	SCR	**27.2**	**88.5**	58.4	73.8	73.4
	APR	44.1	74.9	70.9	76.3	69.5
	AALawyer	31.8	83.8	**75.0**	**81.7**	**77.2**

Table 10: HR-Benchmark with $Qwen$ as judge.

A.2.7 Training without SCR

For the completeness of the experiment, we excluded the target task FAP from the finetuning, resulting in M_{23}, as shown in Figure 8a. We can observe that the results of other tasks only have minor fluctuations. This further demonstrates the effectiveness of our proposed SCR training approach, which does not negatively impact the performance of other finetuning tasks. The primary focus of this work is to validate the effectiveness of SCR. So we are not focus on task-specific data augmentation or finetune for other task objectives. If further work intends to improve performance on other non-oriented tasks, expanding the finetuning data or methods would be a viable strategy. And integrating our SCR retrieval module into the training pipeline does not interfere with the performance of other task objectives.

We trained multiple versions of AA-LeLLM and finally selected M_{20} as our final model. The comparisons are shown in Figure 8. The training differences, detailed results, and further analyses are in the Appendix A.2.9.

A.2.8 Training-Free Deployment with SOTA LLMs

We have observed that as large models evolve, SOTA models are approaching the performance of our fine-tuned AA-LeLLM on the FAP task (around 80%), as shown in Table 1. This indicates the possibility of a training-free SCR, where general models already have the necessary legal prediction and analysis capabilities without specific fine-tuning. To assess this, we conducted an ablation study on SOTA models using the latest version of the criminal law. We find an advantage of our SCR method that when laws are modified, we only need to update our legal article retrieval database. Since the crime corresponding to an article number generally remains unchanged during modified, this approach avoids the hard process of re-annotating data when it becomes a new version of law. This is proved by the result in Table 11 that our baselines (DS-7B and AA-LeLLM) maintained trends consistent with our main tables with old version of law contents.

The recent SOTA level model Qwen3-8B showed a different trend: SCR improved overall performance but unexpectedly increased the Hallu↓. We check the result and find that this stems not from a failure in legal reasoning, but from the model's inability to follow the unseen prompt format for the information-sparse symbolic prediction task. This suggests that finetuning remains necessary for now, and a better-designed prompt could be key for future training-free deployment. Meanwhile, our Table 4 shows that SOTA models are becoming increasingly accurate at the prediction task, reaching a high level, suggesting that as foundation models evolve, they may acquire sufficient latent legal knowledge to make SCR a truly training-free method.

Until such training-free deployment becomes fully viable, finetuning remains a crucial component of our methodology. We acknowledge that updating parametric knowledge is indeed more complex than non-parametric methods. However, under the AALawyer framework, this cost is fully man-

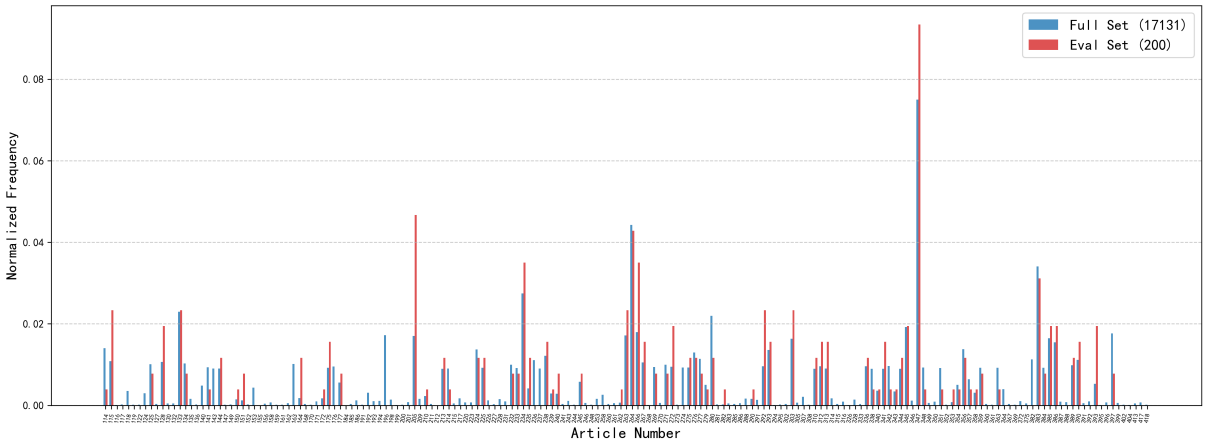
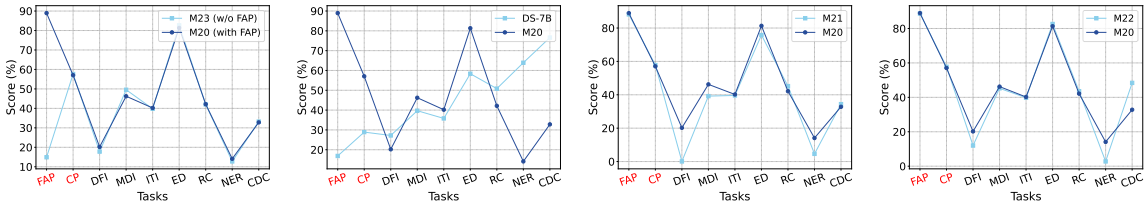


Figure 7: Data distribution comparison.



(a) (b) (c) (d)

Figure 8: Comparison between different versions of AA-LeLLM: (a) Training without FAP, (b) M_{20} vs. DS-7B, (c) M_{20} vs. M_{21}, (d) M_{20} vs. M_{22}. M_{20} is our final version.

ageable and worthwhile. Articles are relatively stable compared to legal cases. Meanwhile, SCR does not require full finetuning but utilizes LoRA finetuning. As detailed in Section 6.1, we employ rank $r = 8$, which updates less than 0.1% of the total parameters. Updating such few parameters allows us to achieve high accuracy surpassing traditional Dense Retrieval. And our APR supports direct updates. For the purpose of establishing the framework from 0 to 1 and verifying the methodology, this work has been completed. For future work, we are considering online finetuning. We can build an RL system or other systems under the current architecture to adapt to small scale finetuning.

A.2.9 Datasets Selection of Training Process

We experimented with three different training strategies, resulting in M_{20}, M_{21}, and M_{22}. After careful consideration, we selected M_{20} as our final AA-LeLLM into the AALaywer. The training datasets used for each model are shown in Table 12. The comparison results are displayed in Figure 8, where we compare the performance of all Group 1 classification tasks and those tasks in Group 2.

Change in Incremental Pretraining. We explored different incremental pretraining strategies to generate M_1. One strategy included only all law articles, criminal law articles and some case documents, resulting in M_{10}, while the other added extra case documents from criminal law to the training data, generating M_{11}. M_{10} was fine-tuned to produce M_{20}, and M_{11} was fine-tuned to produce M_{21}. The results showed that the model M_{21}, which was trained with more data for incremental pretraining, performed worse. This is because our new data consisted of pure criminal law cases, which affected the general legal classification tasks. Additionally, our focus tasks in criminal law, FAP and CP, did not show significant improvement on M_{21}. Therefore, we chose M_{10} for the subsequent SFT stage.

Change in SFT. In the SFT task, we excluded the amount prediction and sentence prediction tasks that influenced CDC (Criminal Damages Calculation) and then fine-tuned M_{10} to produce M_{22}. The results showed that CDC performance improved, confirming our later reasoning that these two tasks indeed affected the model's mathematical operations and logic. However, our AALawyer does not

Baseline	Method	Hallu↓	Prof	Info	Expa	Avg.
DS-7B (non-finetuned)	Vanilla	77.22 ± 0.63	56.13 ± 0.46	38.73 ± 0.12	45.63 ± 0.35	40.82 ± 0.33
	SCR	79.76 ± 0.41	45.93 ± 0.35	37.27 ± 0.40	43.57 ± 0.45	36.75 ± 0.36
	AALawyer	74.30 ± 0.18	53.12 ± 0.07	72.72 ± 0.27	66.31 ± 0.34	54.46 ± 0.17
Qwen3-8B (non-finetuned)	Vanilla	31.21 ± 0.68	72.00 ± 0.36	35.60 ± 0.10	57.50 ± 0.10	58.47 ± 0.28
	SCR	36.61 ± 0.50	77.23 ± 0.05	48.63 ± 0.23	75.63 ± 0.25	66.22 ± 0.20
	AALawyer	29.74 ± 0.14	92.87 ± 0.32	97.77 ± 0.05	96.93 ± 0.32	89.45 ± 0.03
AA-LeLLM (finetuned)	Vanilla	46.77 ± 0.43	64.07 ± 0.12	28.40 ± 0.10	48.53 ± 0.46	48.56 ± 0.21
	SCR	31.68 ± 0.06	74.97 ± 0.21	42.00 ± 0.10	73.63 ± 0.15	64.73 ± 0.07
	AALawyer	36.40 ± 0.44	70.73 ± 0.15	78.70 ± 0.17	84.70 ± 0.40	74.43 ± 0.05

Table 11: Comparison of models on HR-Benchmark on the latest criminal law (different setting with main part table).

require mathematical logic computation, and the focus identification ability of the DFI (Dispute Focus Identification) model in M_{22} declined, which is more important for our model. Therefore, we decided to continue including the amount prediction and sentence prediction training data.

A.2.10 Single Task Training

When we use single-task training, we find that only the training task can work well, as shown in Table 13. So we choose the multi-task training (Caruana, 1997; Yue et al., 2023), combining multiple tasks' training data, as shown in line M_{20} of Table 12, to train our model. We can see that the performance of M_2-multi works well in all tasks, maintain the good performance from DeepSeek-7B and outperforms in our classification task. However, comparing with single-task training, the performance shows slight degradation in CP, but the impact is small, and other tasks perform excellently. For our multi-task training requirements in AALawyer and training efficiency, this approach is clearly better, with minimal impact on the valuable pretrained weights of DeepSeek-7B.

A.2.11 Evaluation on Full Domain Legal Tasks

We further calculate the average score of 11 metrics in LawBench (Fei et al., 2024), comparing with the base model and other models. As shown in the Figure 9, despite being tested on general-domain tasks, our model still maintained a strong level of performance. Since we focus on the classification and analysis of criminal law, and the training data primarily consists of single-label classification and legal analysis tasks, performance on some unseen tasks in some Full-domain legal evaluation is suboptimal. However, the model still maintains a good overall performance. Among all models with comparable parameter sizes, it achieved the SOTA level.

Our performance on some unseen tasks in the full-domain legal evaluation is suboptimal. However, the following analysis indicates that these tasks do not impact the primary objectives of AALawyer. The results are shown in Table 14.

Regarding RC (Reading Comprehension), the reference answers are short, but our model's answers include more analysis (long), leading to a lower score. However, the answers are mostly correct, just with some differences in similarity to the reference examples (the examples are short, and ours are longer), but longer answers are more suitable for our analysis task. The model's performance is not good in NER (Named-Entity Recognition) because our SFT training data only includes tasks related to recognizing criminals, and it doesn't involve other entity recognition tasks, which means it can only recognize criminals. However, as our main tasks are classification and analysis, primarily focusing on identifying the criminal and the relevant legal article numbers, this metric does not have a significant effect. OS (Opinion Summarization) metric is effective for case analysis to supports legal analysis tasks. The model's performance is well. CDC (Criminal Damages Calculation) evaluate the model's numbers extraction and mathematical ability. Due to the presence of tasks related to predicting criminal financial amounts in the SFT training data, this may affected normal calculations. After testing with standard mathematical summation tasks, the results were correct. So we deem that the math level maintains a enough level. And the target tasks do not involve financial amount calculation, so this metric has little impact. Co (Consultation) questions are general law issues, and since our model specializes in criminal law, it can't answer the question in other legal domains accurately, but the results are comparable to other legal domain-specific models.

Process	Source	Type	Num	Size	M_{20}	M_{21}	M_{22}	M_{23}
IPT	(1)	Full Domain Chinese Legal Articles	57k	21MB	✓	✓	✓	✓
	(1)	Articles of the Criminal Law	452	266kB	✓	✓	✓	✓
	(1)	Legal Case Documents	1k	16MB	✓	✓	✓	✓
	(4)	Case Documents in Criminal Law	172k	1052MB		✓		
SFT	(2)	Case - Charge(s)	155k	248MB	✓	✓	✓	✓
	(2)	Case - Article Number(s)	155k	251MB	✓	✓	✓	
	(2)	Case - Criminal	155k	250MB	✓	✓	✓	✓
	(2)	Case - Fine	155k	244MB	✓	✓		✓
	(2)	Case - Sentencing	155k	252MB	✓	✓		✓
	(3)	Criminal Law QA Events Analyses	13k	43MB	✓	✓	✓	✓
	(3)	Full Legal Domain QA Events Analyses	19k	69MB	✓	✓	✓	✓
	(3)	Full Legal Domain QA Tasks	205k	361MB	✓	✓	✓	✓
SCR	(1)	Criminal Article Number - Content	452	164kB	✓	✓	✓	✓
APR	(4)	Vector of Case in Criminal Law	172k	676MB	✓	✓	✓	✓

Table 12: Datasets of training: We use a number to represent the source. (**1**) fuzi.mingcha (Deng et al., 2023) (**2**) CAIL2018 (Xiao et al., 2018) (**3**) DISC (Yue et al., 2023) (**4**) law-lib.

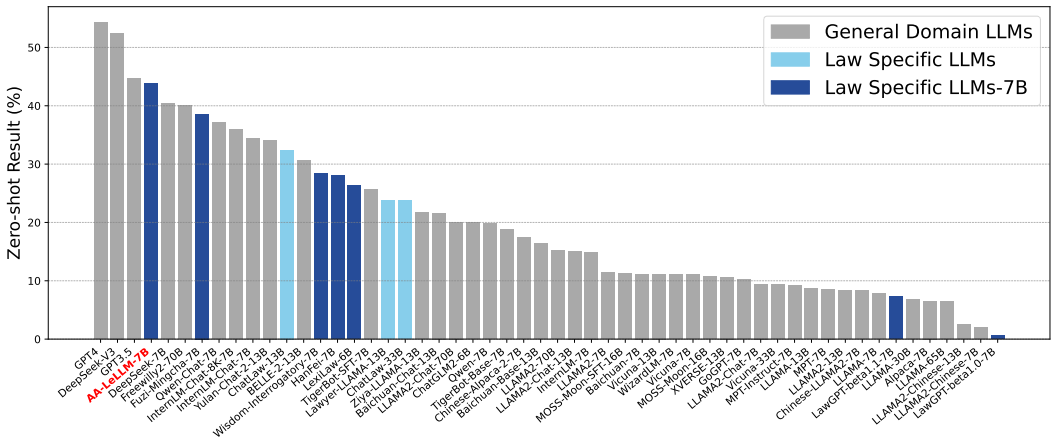


Figure 9: Full domain evaluation on LawBench.

A.2.12 Showcases

We randomly selected a case in the Chinese Court Website as input, and the showcase result is shown in the Table 17. In this real-world case analysis, our AALawyer accurately find the relevant legal articles, generates professional legal reasoning, and retrieves highly similar cases. This demonstrates its high accuracy in legal citation and strong explainability. And even when increasing APR cases k to 5 or varying the input x, the retrieved cases are still closely match to the input.

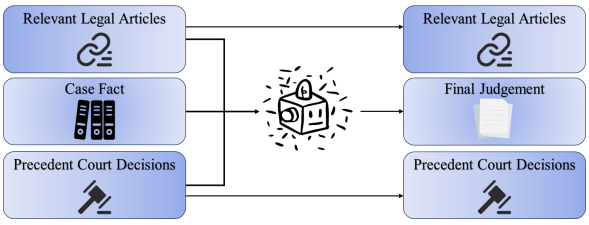


Figure 10: The system concept based on the legal syllogism and actual legal judgment.

A.3 Additional Details

A.3.1 Benchmark and Datasets

Benchmarks. We use 11 metrics in Lawbench, and a 4-dimension HR-Benchmark.

LawBench is a legal benchmark designed based on Bloom's cognitive model, covering Knowledge Memorization, Knowledge Understanding, and Knowledge Applying. We chose LawBench as the benchmark for model ability testing. The evaluation results are in the main paper under the Main Results section. Before evaluation, we analyzed the principles and objectives of the related metrics, removed metrics that were ineffective for all models and excluded all multiple-choice questions. In the end, we leave 11 relevant metrics and created an automated evaluation script to assess each model. The selected metrics are divided into two groups:

The 1st group is based on classification-related tasks in LawBench, which are consistent with our project model's training goals, as shown in Main

Result. It includes 2 criminal law domain MLC (Multi-label Classification) tasks: FAP and CP, 3 full legal domain MLC tasks: DFI, MDI, and ED, and 1 SLC (Single-label Classification) task: ITI. The evaluation details and metrics for each task are as follows:

FAP (Fact-based Article Prediction) predicts the relevant criminal law article numbers for a given event. **CP** (Charge Prediction) predicts the criminal charge related to a given event. **DFI** (Dispute Focus Identification) provides several focus categories for disputes and gives an event sentence to predicts the focus category. **MDI** (Marital Disputes Identification) assigns a classification label to the sentence for marital events and predicts the sentence category. **ED** (Event Detection) provides several event types and makes the model determine which event type is involved. The evaluation metric is F1-Score, shown in Formulas 13, 14, and 15.

ITI (Issue Topic Identification) provides a set of consultation category labels and legal questions, then predicts the consultation category of the given sentence. The evaluation metric is Accuracy, shown in Formula 20.

The 2nd group focuses on the model's performance on other types of tasks in the full legal domain, including 2 extraction tasks, RC and NER, 2 generation tasks, OS and Co, and 1 regression task, CDC. Results are shown in Table 14. The evaluation details and metrics for each task are as follows:

RC (Reading Comprehension) gives an event and a related question, and the task is to answer the question. Most of the questions are focused on information extraction. The evaluation metric is rc-F1, as shown in Formulas 13, 16 and 17.

NER (Named-Entity Recognition) focuses on identifying multiple named entities, such as criminals, victims, stolen currency, time, and location. The evaluation metric is soft-F1, as shown in Formulas 13, 18 and 19.

OS (Opinion Summarization) extracts key elements from input to generate event summaries. **Co** (Consultation) involves answering questions and providing reasons. The evaluation metric is ROUGE-L, as shown in Formulas 21, 22 and 23.

CDC (Criminal Damages Calculation) involves summing the criminal financial amount in the examples to evaluate the model's numbers extraction and mathematical ability. The evaluation metric is Accuracy, as shown in Formula 20.

Our four dimensions proposed by **HR-Benchmark** are based on the evaluation process of real legal cases and the legal syllogism. Among them, the Hallucination of citing legal articles is associated with the Major Premise of the syllogism, and we used a quantification formula similar to Definition 1 for guidance: **Accuracy** of legal article citation = Accuracy of article number prediction (0-5) × Authenticity of the description of the corresponding article content (0-5) ÷ 5. **Informativeness** measures the comprehensiveness of the given information, providing rich material support for reasoning. **Explainability** is to provide accurate and favorable material support for reasoning. **Professionalism** is to make our analysis conform to the rigorous requirements of the legal domain, including word and logic. The detailed rules are provided in Appendix A.3.3. Our metrics can be verified by other legal LLM papers. DISC-LawLLM (Yue et al., 2023) used the same Accuracy (Hallu), Completeness (Info), and Clarity (Expa) in its Subjective Evaluation. And in ChatLaw (Cui et al., 2023), they used Completeness (Info), Logic, Language Quality(Prof), Correctness (Hallu), Guidance, and Authority (Expa) as their criteria to evaluate in their legal consultations.

Baselines. We selected six models for performance comparison, including the base model DeepSeek-R1-Distill-Qwen-7B (DeepSeek-7B) and five legal models of similar parameter scale: LaWGPT-7B-beta1.1, LexiLaw, HanFei, Wisdom-Interrogatory and Fuzi-Mingcha. In the LawBench paper, the Fuzi-Mingcha model was the best-performing 7B legal domain model in their evaluation. As our research progressed, we also conducted supplementary tests on newly emerged SOTA models, such as Qwen3-8B and Internlm3-Instruct-8B.

After checking the output, we found that the DeepSeek model outputs the $< think >$ tag along with the regular output, affecting evaluations. Therefore, we designed a script to process the output, removing the content within the $< think >< think/ >$ tag to only maintain the valid output, which ensure the normal operation of the evaluations.

Datasets. The metrics that we used in benchmarks use CAIL2018, CAIL2019, CAIL2021, CAIL2022, LAIC2021, LEVEN, CrimeKgAssitant, AIStudio, hualv.com. And we use DISC-Law-SFT, fuziminghca to train our model, use lawlib in

Model	Criminal		Full Domain					
	FAP	CP	DFI	MDI	ITI	RC	OS	Co
DeepSeek-7B	16.86	28.89	**27.20**	39.69	35.80	**50.83**	31.66	15.10
M_2-single	0.00	**59.03**	0.00	3.60	0.00	0.00	0.01	0.05
M_2-multi	**88.94**	57.05	20.20	**46.24**	**40.20**	42.14	**43.10**	15.29

Table 13: Comparison of Single and Multiple Task Training under Different Tasks.

Model	Criminal		Full Domain				
	FAP	CP	RC	NER	OS	CDC	Co
DeepSeek-7B	16.86	28.89	50.83	**63.90**	31.66	**76.60**	15.10
LawGPT-beta1.1-7B	0.15	15.68	2.27	2.00	8.61	15.40	7.62
LexiLaw-6B	13.15	39.99	45.39	48.74	33.12	35.80	15.82
Fuzi-Mingcha-7B	25.19	55.93	**97.59**	44.07	**54.32**	47.20	**16.64**
AA-LeLLM-7B	**88.94**	**57.05**	42.14	14.10	43.10	32.80	15.29

Table 14: Scores for criminal law and full domain law tasks on Lawbench.

our APR.

A.3.2 Evaluation Metrics

F1-score FAP (Fact-based Article Prediction), CP (Charge Prediction), DFI (Dispute Focus Identification), MDI (Marital Disputes Identification) and ED (Event Detection) use this metric, as shown in Formulas 13, 14, and 15. ϵ is a small constant (e.g., 10^{-10}) to prevent division by zero. TP is true positive. FP is false positive. FN is false negative.

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall} + \epsilon} \tag{13}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

rc-F1 RC (Reading Comprehension) use this metric, as shown in Formulas 13, 16 and 17. $|P|$ is the total number of tokens in the predicted answer. $|R|$ is the total number of tokens in the reference(answer). $|S|$ is the number of overlapping tokens between the predicted answer and the reference.

$$\text{Precision} = \frac{|S|}{|P|} \tag{16}$$

$$\text{Recall} = \frac{|S|}{|R|} \tag{17}$$

soft-F1 ER (Named-Entity Recognition) use this metric, as shown in Formulas 13, 18 and 19. P represents the set of predicted legal entities by the model. R represents the set of actual legal entities in the reference (answer).

$$\text{Precision} = \frac{\sum_{\text{i} \in P \cap R} F1_{\text{i}}}{|P|} \tag{18}$$

$$\text{Recall} = \frac{\sum_{\text{i} \in P \cap R} F1_{\text{i}}}{|R|} \tag{19}$$

Accuracy ITI (Issue Topic Identification) use this metric, as shown in Formula 20. TP is true positive. FP is false positive. FN is false negative. FN is false negative.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{20}$$

ROUGE-L OS (Opinion Summarization), Co (Consultation) use this metric, as shown in Formulas 13, 18 and 19. LCS (Longest Common Subsequence) is a sequence that appears in the same relative order in both input sequences, but not necessarily consecutively. We choose $\beta = 1$ because there is no specific need to favor either Precision or Recall over the other.

$$F_{\text{lcs}} = \frac{(1 + \beta^2) \cdot R_{\text{lcs}} \cdot P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 \cdot P_{\text{lcs}}} \tag{21}$$

$$P_{\text{lcs}} = \frac{|LCS(P, R)|}{|P|} \tag{22}$$

$$R_{\text{lcs}} = \frac{|LCS(P, R)|}{|R|} \tag{23}$$

A.3.3 Disclosure of Prompt Templates

Two prompts of AALawyer are shown in Table 15. Detailed standards for four-dimensional metrics in HR-Benchmark are shown in Table 16, the complete evaluation prompts will be released in our codebase.

A.4 More details in Execution Process

A.4.1 Data Collection and Data Preprocessing

The data for the Incremental Pretraining (IPT) consists of 57k full domain Chinese legal articles, 452 articles of the criminal law and 1k legal case documents. The data is processed into a JSON format containing only the "text" attribute, which is suitable for model pretraining.

Considering that these model pretraining data is insufficient, may resulting in poor performance. We have used a web scraper to collect legal case documents from the website, collecting 172k documents.

In the Supervised Fine-Tuning (SFT), we applied part of the CAIL2018 training datasets and DISC training dataset. The data was divided into nine parts. The first five tasks used CAIL, where key information in the documents was masked to predict the charges, relevant legal article numbers, criminals, fines, and sentencing. The remaining four tasks were processed using DISC, generated 13k criminal law QA events analyses, 19k full legal domain QA events analyses, and 66k and 139k full legal domain QA tasks.

Although we only trained the model with the training set, considering that DISC and 2 evaluation tasks use the same dataset. To ensure the final accuracy, we converted the text into TF-IDF (Term Frequency-Inverse Document Frequency) vectors and computed the cosine similarities

$$\cos(\vec{\alpha}, \vec{\beta}) = \frac{\vec{\alpha} \cdot \vec{\beta}}{\|\vec{\alpha}\| \|\vec{\beta}\|} \tag{24}$$

for the overlapping portions of the training set and the test set, with a threshold set at 0.5. We find that no segments in the test sets exceeded this threshold.

Table 12 provides detailed descriptions of each dataset used in our project and specifies which data were utilized for training each weight of the model.

A.4.2 Training Process

DeepSeek has demonstrated strong general-domain capabilities and shows promising performance on legal-domain tasks even before fine-tuning, as shown in Table 12. To further enhance its effectiveness on our AALawyer-related legal tasks, we incremental pretrained and finetuned it using domain-specific legal datasets via LlamaFactory (Zheng et al., 2024), enabling the model to better adapt to our target objectives. Followed by the normal pipeline described in Section 3.2, our entire process is shown in Figure 11.

Figure 11: Model Training Flowchart

Stage 1 Incremental Pretraining (IPT). Our goal was to adapt the base DeepSeek-R1-Distill-Qwen-7B model (Guo et al., 2025) to the legal domain's language and style. We used the datasets described in Section A.4.1 for unsupervised Causal Language Modeling (CLM), with the training objective

$$L_{\text{IP}} = -\sum_{t=1}^{T} \log P_\theta(y_t \mid y_{<t}). \tag{25}$$

This process, which resulted in model M_1, was conducted using LoRA (Hu et al., 2022).

For the LoRA adaptation, we set the rank r to 8, the scaling factor α to 16, and the dropout rate to 0. We used the AdamW optimizer with a learning rate of 5×10^{-5} and a cosine learning rate scheduler. The training was conducted with a per-device batch size of 2, 8 gradient accumulation steps, and a sequence length of 2048.

Stage 2 Supervised Finetuning (SFT). We then finetuned model M_1 to create our final model, AALeLLM, improving its performance on the legal tasks we are focusing on. Recognizing that single-task training was ineffective for other tasks, we employed multi-task training (Caruana, 1997; Yue et al., 2023) by mixing all SFT datasets. The SFT and multi-task learning objectives are

$$L_{\text{SFT}} = -\sum_{t=1}^{T} \log P_\theta(y_t \mid x, y_{<t}) \tag{26}$$

and

$$L_{\text{MTL}} = \sum_{i=1}^{N} \lambda_i \cdot L_{\text{SFT}}^{(i)}. \tag{27}$$

The per-device batch size was set to 1, while other settings remained consistent with the IPT stage.

B Clarify of LLMs Usage

We used ChatGPT and Gemini to improve the phrasing and grammar of our paper during the writing process. In the experimental phase, these models were also used to debug and modify our code. All content is authentic, and the final wording and the effectiveness of the code and results have been checked by us.

Format 1: (Xiao et al., 2018) "Based on the following facts and charges, provide the relevant Criminal Law articles. Only provide the article numbers, and place your answer between [Article] and <eoa>. For example: [Article] Criminal Law Article 128, Criminal Law Article 341 <eoa> Facts: {USER_INPUT}"
Format 2: "Case Details: {USER_INPUT} Relevant Articles Content: {ARTICLES_CONTENT} (Note: The relevant articles might not be fully utilized; it could be a specific paragraph of the article.) Please comprehensively analyze the case based on the case details, the relevant articles, and the similar cases provided below. Explain who committed the crime, why they are considered guilty, the applicable laws, and the specific charges. Keep the analysis to around 100 words and refer to the formatting of the similar cases. [Similar Cases]: {SIMILAR_CASES}"

Table 15: The prompts of AALawyer.

"1. Professionalism: Whether the [analysis content] conforms to standard legal terminology and the logic of criminal law analysis, and can provide correct analytical reference for legal practitioners.
2. Accuracy: Accuracy of legal article citation = Accuracy of article number prediction (0-5) × Authenticity of the description of the corresponding article content (0-5) ÷ 5. Scoring criteria for Accuracy of article number prediction (0-5 points): This item only assesses the matching degree of the "article" number, and does not assess the "paragraph". If the legal article numbers (articles) cited in the [analysis content] completely cover the [reference legal article numbers], then this item scores 5 points. If the [analysis content] cites articles other than the [reference legal article numbers], or fails to cite any article from the [reference legal article numbers], then the calculation is performed with reference to the F1 score (the harmonic mean of precision and recall). The [reference legal article numbers] are the most important; if any one of them is not covered, the score for this item will be significantly reduced; if all [reference legal article numbers] are missing, then this item scores 0 points. Due to data annotation reasons, legal articles other than the [reference legal article numbers] may also be correct legal articles, you can judge by yourself. Special note: If the article numbers cited in the [analysis content] are consistent with the [reference legal article numbers], even if other paragraphs under that article or definitions of other crimes are additionally cited, no points will be deducted. For example, the [reference legal article number] is "Article 343", and the [analysis content] cites both the crime of illegal mining and the crime of destructive mining under "Article 343", it is still considered that the article number prediction is completely accurate, and this item scores 5 points. Scoring criteria for Authenticity of the description of the corresponding article content (0-5 points): Assess the authenticity of the cited legal article content. As long as the [analysis content] provides any real, non-fabricated legal article content (whether it is the entire article, part of a paragraph, or a definition of a crime) for the cited legal article number, this item scores 5 points. The completeness of the content and the citation format are not assessed. This indicator is only related to the above two factors and has nothing to do with the quality of the analysis content. If the [analysis content] provides any real legal article content, it means the authenticity score is 5. If the [analysis content] contains complete relevant legal articles, it means the authenticity score is 5, and only the accuracy of the article number prediction affects the accuracy score.
3. Informativeness: Whether the case analysis in the [analysis content] has a clear structure and complete content, whether it combines the specific facts of the case with the legal provisions for in-depth explanation, and whether it contains supplementary information or reasonable reasoning. Whether it can provide rich reference for legal practitioners.
4. Explainability: The transparency of the [analysis content]. That is, whether the analysis content contains supporting reliable materials, such as real legal provisions and relevant judgment cases, etc. Whether legal practitioners can feel that the analysis is very reliable."

Table 16: Detailed standards for four-dimensional metrics in HR-Benchmark.

Input(Case): Concert tickets turn out to be a blank cheque “Scalpers" buy concert tickets on behalf of the “mystery". In October 2023, the defendant Gao Moqi published a false advertisement for the sale of concert tickets of a star’s fans in the wechat group, and the victim Xiao Ji (a pseudonym) contacted him to buy concert tickets after seeing it. Gao Moqi defrauded Xiao Ji of a total of 4,500 yuan by issuing tickets. In November of the same year, the defendant Gao Mouqi was seized by the public security organs and was released on bail on the same day. In July 2024, the defendant Gao Mou Qi resumed his old business, published false advertisements for the sale of concert tickets under a star’s microblog dynamics, and defrauded multiple victims of a total of 19,723 yuan by issuing tickets and unfreezing accounts.
Related Articles: Article 266 Whoever defrauds public or private property, if the amount involved is relatively large, shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance and shall also, or shall only, be fined; If the amount involved is huge, or if there are other serious circumstances, he shall be sentenced to fixed-term imprisonment of not less than three years but not more than 10 years and shall also be fined; If the amount involved is especially huge or if there are other especially serious circumstances, he shall be sentenced to fixed-term imprisonment of not less than 10 years or life imprisonment and shall also be fined or have property confiscated. Where this Law provides otherwise, such provisions shall prevail.
Case Analysis: The defendant Gao Mouqi is guilty of fraud. According to Article 266 of the Criminal Law, whoever, with the purpose of illegal possession, defrauds another person of money or property, if the amount involved is relatively large, shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance and shall also, or shall only, be fined. Gao Mou Qi defrauded Xiao Ji with a total of 4,500 yuan on the grounds of ticket issuance, and defrauded a number of victims with a total of 19,723 yuan on the grounds of ticket issuance and unfreezing accounts, which is a large amount, which is in line with the constitutive requirements of the crime of fraud. Therefore, Gao should be sentenced to less than three years in prison and fined.
Similar Cases: (Too long, shown in Table 18)

Table 17: Output of AALawyer.

(2020) Shanghai No. 0101 Criminal Case 621
The public prosecutor is the People's Procuratorate of Huangpu District, Shanghai.
Defendant Wang 2, female, born on March 9, 1997, of the Yi ethnic group, has her registered residence in Yunnan Province.
The defense attorney is Sun Peng, a lawyer from Shanghai Hushi Law Firm.
The People's Procuratorate of Huangpu District, Shanghai Municipality, with the indictment No. 80 of criminal prosecution of Huangpu District, Shanghai Municipality [2020], charged the defendant Wang 2 with fraud and filed a public prosecution with this court on September 3, 2020. This court lawfully applied the simplified procedure, conducted a trial by a single judge, and held a public hearing on this case. The People's Procuratorate of Huangpu District, Shanghai Municipality, assigned prosecutor Chen Mou 2 to support the public prosecution in court. The defendant Wang 2 and his defense lawyer Sun Peng attended the trial. The case has now been concluded.
The People's Procuratorate of Huangpu District, Shanghai Municipality, has charged that between June 2018 and September 2019, the defendant Wang 2 repeatedly posted false information online claiming to have concert tickets for sale and fabricated a counterfeit "Zhuanzhuan" second-hand trading website page to deceive multiple victims who were seeking to purchase tickets. Subsequently, Wang 2 sent false payment links from the "Zhuanzhuan" website to the victims, defrauding them of a total of 24,125 yuan (the currency used hereinafter is the same). After obtaining the money, Wang 2 cut off contact with the victims and squandered all the ill-gotten gains. The specific facts are as follows:
1. On June 14, 2018, the defendant Wang 2 falsely claimed on Weibo that he had EXO concert tickets for sale, defrauding the victim Zhong Moumou of 4,022 yuan for ticket purchases. ..... On September 30, 2019, the defendant Wang 2 fabricated on Weibo that he had tickets for Troye Sivan's concert in Chengdu and defrauded the victim Zhou 3 of 1,498 yuan for the ticket purchase. On April 28, 2020, the public security authorities arrested the defendant Wang 2 in Kunming City, Yunnan Province. After being apprehended, with the assistance of his family, Wang 2 has returned all the ill-gotten gains.
The above facts are confirmed by the statements of the victims Zhong Moumou, Wan Moumou, Li Moumou, Xu Moumou, Dai Moumou, Man Moumou, Chen Mou1, Zhang Moumou, Shen Moumou, Zhou Mou1, Xue Mou, Yuan Moumou, Zhou Mou2, and Zhou Mou3; the testimony of witness Wang 1; screenshots of relevant WeChat chat records; screenshots of WeChat transfer records; the seizure decision and list issued by the Cultural Security Division of the Shanghai Municipal Public Security Bureau; the working situation issued by the public security authorities; the seizure decision, list of seized property and documents issued by the People's Procuratorate of Huangpu District, Shanghai; the letter of forgiveness; and the multiple confessions of the defendant Wang 2.
The public prosecutor believes that the defendant Wang 2, with the intent of illegal possession, fabricated facts and concealed the truth to defraud others of their property in a relatively large amount. His actions have violated Article 266 of the Criminal Law of the People's Republic of China and he should be held criminally responsible for fraud. The defendant Wang 2 has confessed to the crime and accepted the punishment. According to Article 15 of the Criminal Procedure Law of the People's Republic of China, he may be given leniency. The defendant Wang 2 has truthfully confessed to the criminal facts. According to Article 67, Paragraph 3 of the Criminal Law of the People's Republic of China, he may be given a lighter punishment. It is suggested that Wang 2 be sentenced to 11 months in prison and fined 2,000 yuan.
The defendant Wang 2 has no objection to the facts, evidence, charges and sentencing suggestions made by the public prosecutor regarding his crime of fraud and has signed and confirmed them. During the court hearing, he also expressed no objection. The defense lawyer has no objection to the facts, evidence, charges and sentencing suggestions made by the public prosecutor. The lawyer believes that the defendant Wang 2 can truthfully confess his criminal facts, has returned all the ill-gotten gains and has obtained the forgiveness of the victims Li Moumou and Zhang Moumou, and thus should be given a lighter punishment.
This court holds that the defendant Wang 2, with the intent of illegal possession, repeatedly posted false information on the internet and concealed the truth, defrauding multiple victims of a considerable amount of property. His actions have violated the criminal law and constitute the crime of fraud, for which he should bear criminal responsibility. The facts charged by the public prosecutor against the defendant Wang 2 for fraud are clear, the evidence is solid and sufficient, and the charge is valid. This court supports it. The defendant Wang 2 can truthfully confess his crime after being arrested, and thus can be given a lighter punishment; he can also accept the punishment and thus can be given leniency. The defense lawyer's opinion that the defendant Wang 2 should be given a lighter punishment is adopted. Therefore, in accordance with Article 266 and Article 67, Paragraph 3 of the Criminal Law of the People's Republic of China,

Table 18: Similar cases.