LRAS: Advanced Legal Reasoning with Agentic Search

Anonymous ACL submission

Abstract

While Large Reasoning Models (LRMs) have demonstrated exceptional logical capabilities in mathematical domains, their deployment in legal settings remains hindered by two fundamental failures: an *introspection deficit*, where models cannot recognize their own knowledge boundaries and produce confident yet incorrect conclusions, and *fragility under complexity*, where static retrieval cannot support multi-statute reasoning requiring adaptive, iterative search. We present **LRAS** (**L**egal **R**easoning with **A**gentic **S**earch), the first framework to internalize agentic search capabilities into legal LLMs through end-to-end training, shifting models from static "closed-loop thinking" to dynamic "Active Inquiry" without relying on external orchestration. **Introspective Imitation Learning** teaches models to recognize legal knowledge boundaries via a three-level progressive mechanism, while **Difficulty-aware Reinforcement Learning** enables autonomous multi-step search planning calibrated to legal task complexity. Experiments across four Chinese legal benchmarks and three model scales (4B/8B/14B) show consistent gains of 8.2–32% over strong baselines. Under identical retrieval infrastructure, LRAS outperforms RAG, prompt-based agentic search, and multi-agent frameworks, while introducing only 32% additional token overhead over the base model.

1 Introduction

Large Reasoning Models (LRMs) have achieved great success in expanding the boundaries of complex logical problem-solving in mathematical and symbolic domains (Yang et al., 2025a; Guo et al., 2025; Cao et al., 2025a; Chen et al., 2025). However, despite these advancements, existing approaches fall short in legal scenarios, which demand extreme precision and strict adherence to procedural logic. While pioneering works have attempted to solve complex legal problems through internalized logical deduction (Lab, 2024; Cai et al., 2025; Han et al., 2025; Cao et al., 2025b), they face a critical bottleneck: reliance on "closed-loop" reasoning without interaction with external knowledge.

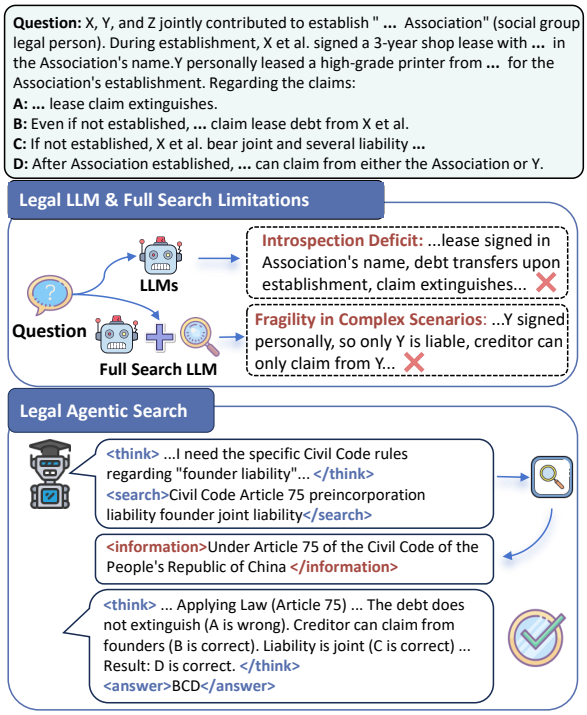

Figure 1: Comparison of legal LLMs, traditional search, and our proposed legal agentic search framework. Conventional legal LLMs suffer from limited introspection capabilities, failing to recognize their knowledge boundaries. In contrast, LRAS systematically identifies knowledge gaps, retrieves relevant statutes with precision, and generates well-grounded legal conclusions.

This isolation results in two primary limitations:

(i) The Introspection Deficit: LRMs suffer from a fundamental lack of self-awareness regarding their knowledge boundaries. LRMs frequently fail to distinguish between verified knowledge and information gaps, resulting in confident hallucinations rather than acknowledgments of uncertainty (Dahl et al., 2024; Blair-Stanek and

Van Durme, 2025). Our experiment shows that in over 70% of error cases, models fail to trigger necessary verification even when tools are available. This indicates that the primary failure mode is not merely a lack of knowledge, but a deficit in introspection, which is the inability to decide whether to search.

(ii) Fragility in Complex Legal Scenarios: In our preliminary experiments, we observe that for complex legal reasoning scenarios, relying solely on full search is completely insufficient. We find a substantial performance gap between full search and agentic search on deep reasoning, indicating that static methods struggle to bridge the reasoning deficit. This inherent inflexibility largely restricts the upper bound of the models' abilities. Consequently, models should learn to search autonomously rather than passively processing static search results. This enables them to adapt to diverse and complex legal scenarios.

To address these limitations, we propose **LRAS** (**L**egal **R**easoning with **A**gentic **S**earch), a framework that shifts legal reasoning from static, parametric "closed-loop thinking" to dynamic "Active Inquiry" (Figure 1). Unlike prior work that relies on external orchestration (**?**Yang et al., 2025b), LRAS is the first framework to *internalize* agentic search capabilities directly into the model through end-to-end training, enabling autonomous and adaptive retrieval without external coordination. The framework integrates two complementary learning mechanisms: **(i) Introspective Imitation Learning** addresses the introspection deficit by training the model to recognize its own knowledge boundaries through a three-level progressive mechanism. The model learns to proactively trigger external search when facing ambiguous or precision-critical legal content, grounding responses in verified statutes and case law. **(ii) Difficulty-aware Reinforcement Learning** resolves fragility in complex scenarios by enabling the model to autonomously plan and execute multi-step exploratory searches. This evolves the system from reactive boundary-triggered retrieval to autonomous handling of multi-statute reasoning, effectively bridging the reasoning gap that static methods cannot close.

Our main contributions are summarized as follows:

- We introduce LRAS, the first framework that shifts legal LLMs from static "closed-loop thinking" to dynamic "Active Inquiry," enabling models to identify knowledge boundaries and handle legal reasoning complexity.
- We design a dual-mechanism learning architecture, comprising Introspective Imitation Learning and Difficulty-aware Reinforcement Learning. This effectively resolves the introspection deficit regarding whether to search, and overcomes the fragility of static retrieval in complex legal scenarios by autonomously learning how to search.
- Extensive experiments demonstrate the effectiveness of LRAS, showing performance gains from 8.2% to 32%, particularly in tasks requiring deep reasoning with reliable knowledge.

2 Related Works

2.1 Legal Reasoning with LLMs

Early LLM applications in legal domains focused on general-purpose tasks such as judicial judgment prediction and contract generation (Hou et al., 2025; Dehghani et al., 2025). Subsequent benchmarks,(e.g.LegalBench (Guha et al., 2023), LawBench (Fei et al., 2024), and LexEval (Li et al., 2024)) revealed significant shortcomings in the legal domain, prompting a shift toward step-by-step reasoning enhancement. Recent efforts have proposed to formalize legal syllogism (LoT (Jiang and Yang, 2023) and LawChain (Xie et al., 2025)) or improve reliability via process supervision (LegalReasoner (Shi et al., 2025) and SyLeR (Zhang et al., 2025b)), while agentic frameworks like GLARE (Yang et al., 2025b) seek to deepen legal reasoning by invoking dynamic modules. However, most current approaches operate as closed-book systems, relying on static parameters rather than dynamic knowledge retrieval. This limits their ability to shift from "closed-loop thinking" within fixed weights to "active inquiry" that engages external sources. To address this, we propose the first unified framework integrating agentic search with legal reasoning, equipping LLMs with the capacity for dynamic interaction with external legal environments.

2.2 Legal-Aware Reinforcement Learning

Recent work has adapted reinforcement learning (RL) to enhance legal reasoning in LLMs. One line of research structures logical reasoning through legal-augmented or structure-aware rewards, as in SyLeR (Zhang et al., 2025b), LegalDelta (Dai et al.,

2025), and Unilaw-R1 (Cai et al., 2025), to enforce rigorous syllogistic reasoning. Another focuses on factual and procedural correctness, employing strategies such as citation-aware rewards (Akarajaradwong et al., 2025) and hard-sample mining (e.g., HIPO (Hu et al., 2025)) to reduce hallucinations and ensure precision. Further efforts like LexPam (Zhang et al., 2025a) extend RL to legal-mathematical reasoning via curriculum learning. Despite these advances, current legal RL methods remain limited by a static training environment hindering models from adaptively navigating complex legal scenarios, a gap which our proposed method seeks to bridge.

3 Preliminary and Motivation

3.1 Preliminary

Problem Formulation. We formulate legal agentic search as a sequential decision-making process augmented by an external search environment. Let $x \in \mathcal{X}$ denote the input legal query and $\mathcal{E}_{\text{search}}$ represent the external search environment. Unlike standard LRMs that rely solely on parametric memory to map $x \to y$, LRAS operates as a policy π_θ. The agent actively interacts with $\mathcal{E}_{\text{search}}$ to acquire external legal information, constructing a reasoning trajectory τ to support its final answer y.

Introspective Action Space. The reasoning process unfolds sequentially over T steps. At each step t, the agent generates an action z_t conditioned on the input x and the history $h_t = (z_1, o_1, \ldots, z_{t-1}, o_{t-1})$. We define the action space $\mathcal{Z}$ as the following three types:

- **Introspective Reasoning (z_t^{think}):** The process initiates with internal thinking (encapsulated within `<think>...</think>`). The agent evaluates whether its parametric knowledge is sufficient. Based on this assessment, it transitions to one of two subsequent states.
- **Active Inquiry (z_t^{search}):** If internal knowledge is deemed insufficient, the agent formulates a search query (encapsulated within `<search>...</search>`). This action triggers the environment to return an observation $o_t = \mathcal{E}_{\text{search}}(z_t)$ (wrapped in `<information>...</information>`). The process then loops back to Introspective Reasoning to assess the adequacy of the new information.
- **Answer Generation (y):** Once the agent determines that the accumulated information in history h_t is sufficient, it terminates the search loop and generates the final legal response.

Category	Count (Percentage)
Total Base Model Failures	**705 (100%)**
Samples Where Search Was Triggered	202 (28.7%)
Samples Where Search Was *Not* Triggered	503 (71.3%)
Successfully Corrected by Search	**58 (8.2% of Total)**
Search Success Rate (Corrected / Triggered)	*28.7%*

Table 1: Impact of Search Augmentation on Failed Queries. Among 705 instances randomly sampled from JEC-QA where the base model failed, search was triggered in only 28.7% of cases (revealing a lack of introspection). When triggered, search successfully corrected the answer 28.7% of the time.

Thus, the complete trajectory can be defined as $\tau = (z_1^{\text{think}}, [z_1^{\text{search}}, o_1, z_2^{\text{think}}, \ldots], y)$, representing an introspective process of dynamic think and active inquiry that concludes upon reaching knowledge sufficiency.

Objective. The objective is to maximize the accuracy of legal answers conditioned on introspective reasoning:

$$\max_\theta \sum_{i=1}^{N} \mathbb{I}(y_i = y_i^*), \quad y_i = \pi_\theta(x_i, \tau_i), \quad (1)$$

where τ_i is the reasoning trajectory (think and search actions) for query x_i, and y_i is the answer generated after trajectory completion.

3.2 Motivation

To characterize the limitations of current LRMs in the legal domain, we conduct exploration experiments and identify a critical bottleneck: LRMs suffer from a fundamental lack of introspection, frequently resulting in confident hallucinations rather than acknowledgments of uncertainty. This deficit is exemplified in our case study (Figure 7), where the model fabricates a "one-year" deadline alongside a spurious rationale, failing to distinguish between verified knowledge and information gaps. Quantitative analysis of 705 incorrect responses (Table 1) confirms this pattern: while external search tools could have corrected 8.2% of errors, the model triggered the search mechanism in only 28.7% of failure instances. This indicates that in over 71% of cases, the model failed to recognize its own need for verification. Consequently, the primary failure mode is not a mere lack of knowledge, but a deficit in introspective capability—the

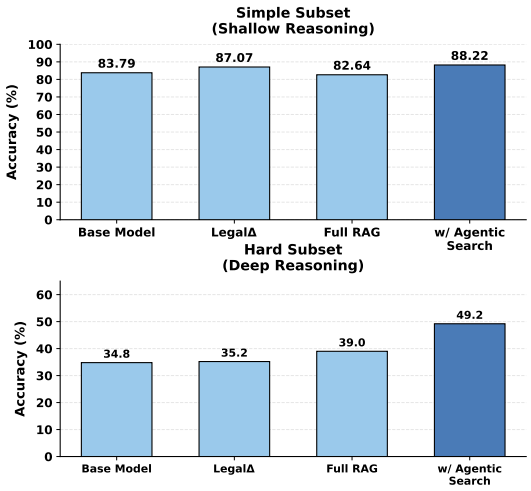

Figure 2: Impact of search strategies on Shallow and Deep Reasoning.

inability to identify knowledge boundaries prevents the model from utilizing available tools to remedy its errors. Consequently, this evidence directly motivates our approach of integrating search as a core mechanism to enhance the reliability and extend the functional boundaries of legal LLMs.

While external knowledge is critical for complex legal reasoning, standard retrieval methods like Full RAG, which incorporate an external knowledge database with the model for a static pipeline, prove insufficient. As illustrated in Figure 2, the necessity of search varies significantly by difficulty. On the Simple Subset (Shallow Reasoning), internal knowledge is largely adequate, with search yielding only marginal gains (improving the Base Model from 83.79% to 88.22%). In stark contrast, the Hard Subset (Deep Reasoning) reveals a severe capability gap, where models plateau around 35% accuracy. Crucially, passive retrieval strategies fall short here: standard "Full RAG" only lifts performance to 39.0%. However, Agentic Search achieves a breakthrough, reaching 49.2%. This substantial margin demonstrates that complex legal problems require more than just context; they demand an agentic capability to actively navigate and synthesize information.

These observations motivate our **LRAS** framework. To address the introspection deficit, we employ Introspective Imitation Learning, teaching the model to recognize its knowledge boundaries and decide "whether to search" to avoid precision cased by the Introspection Deficit. To further increase the fragility in Complex Legal Scenarios, the model's autonomous decision capacity across complex and underspecified scenarios, we introduce Difficulty-aware Reinforcement Learning, which can teach models "how to search" by shifting from passive context consumption to systematic, evidence-based exploration for complex queries.

4 Methodology

4.1 Overview

We propose a dual-mechanism learning architecture designed to address two key challenges prevalent in the legal domain: the Introspection Deficit and Fragility in Complex Legal Scenarios. Our framework integrates Introspective Imitation Learning and Difficulty-aware Reinforcement Learning to achieve precise, proactive, and trustworthy legal reasoning as shown in Figure 3.

Stage 1: Introspective Imitation Learning. To mitigate the Introspection Deficit, we establish a specialized data curation pipeline. By employing uncertainty-aware query sampling and introspective process synthesis, we construct a dataset that explicitly models knowledge boundaries. Through introspective learning, we train the model to emulate expert behaviors, proactively triggering external searches only when facing ambiguous or precision-critical legal content.

Stage 2: Difficulty-aware Reinforcement Learning. To counter the fragility in complex, underspecified scenarios, we advance beyond static supervision compared with stage 1. We construct a "hard" dataset based on SFT failure rates and employ Group Relative Policy Optimization (GRPO). This stage empowers the model to autonomously plan and execute multi-step exploratory searches, evolving the agent from reactive fact-checking to the deep reasoning required for navigating intricate legal problems.

4.2 Introspective Imitation Learning

4.2.1 Introspective Data Synthesis

To facilitate Introspective Imitation Learning, we curate a high-quality dataset that compels the model to learn structured reasoning and search invocation. This process involves a rigorous pipeline of sourcing, synthesis, and verification.

We derive our initial corpus from three primary legal benchmarks: JEC-QA (Zhong et al., 2020), CAIL2018 (Xiao et al., 2018), and CAIL2021 (CAIL, 2021), which provide a mix of objective multiple-choice and complex QA entries. To ensure the training data provides a suffi-

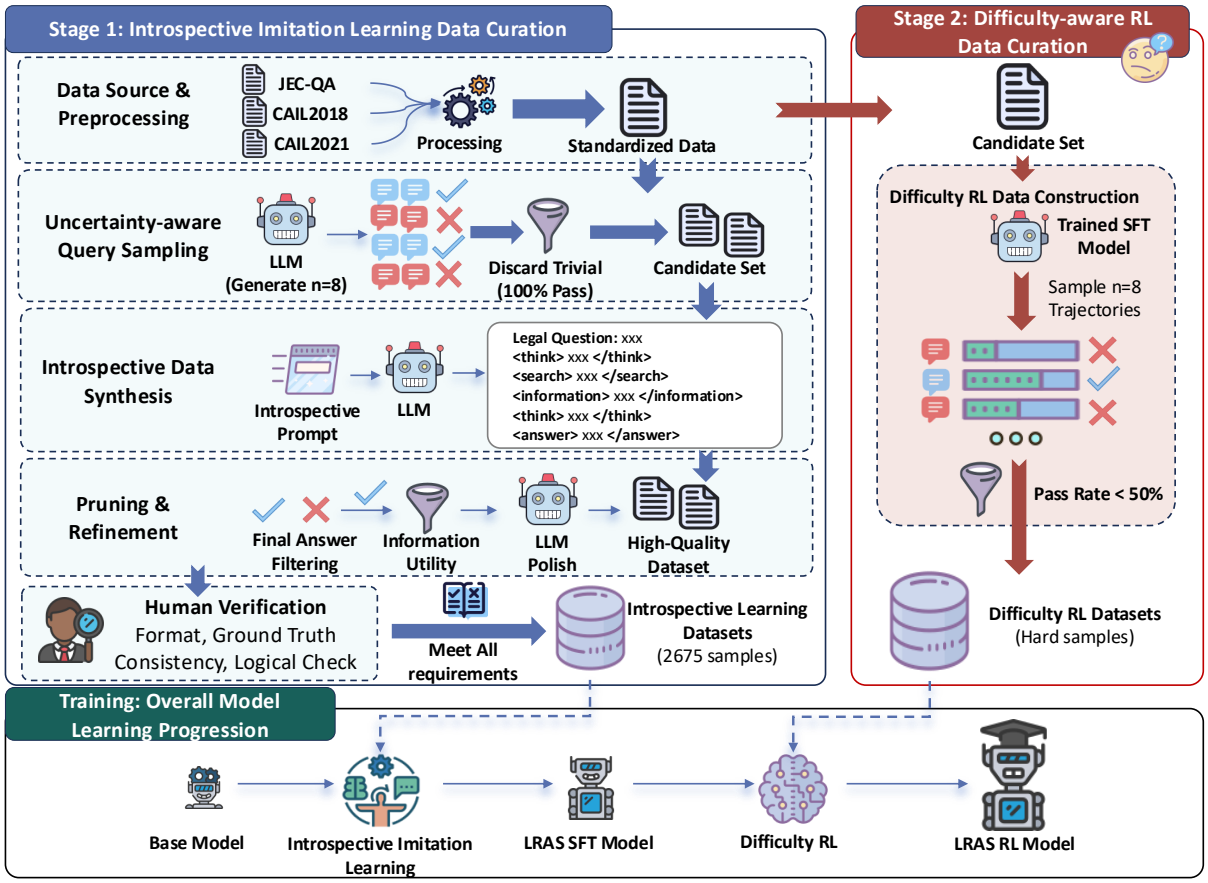

Figure 3: The overall framework of our proposed method. The workflow consists of two main phases: Stage 1, Introspective Imitation Learning Data Curation, where raw legal data is processed, filtered for uncertainty, and synthesized into reasoning trajectories with search actions; and Stage 2, Difficulty-aware RL Data Curation, where hard samples are selected based on pass rates to train the model via reinforcement learning. The bottom panel illustrates the progressive training pipeline from the Base Model to the final LRAS RL Model.

cient learning signal, we implement an uncertainty-aware query sampling strategy. Using Qwen3-235B-Instruct, we generate $n = 8$ candidate responses for each source question. We discard trivial queries where the model achieves a perfect pass rate, retaining only those exhibiting inherent difficulty or reasoning ambiguity. This filtering ensures the dataset focuses on scenarios where the model's internal parametric knowledge is insufficient, necessitating external retrieval.

For the retained queries, we synthesize high-quality reasoning trajectories via prompt-based generation. We enforce a structured format using specialized tags: `<think>` for internal reasoning, `<search>` for query generation, `<information>` for retrieval, and `<answer>` for the final response. To eliminate hallucinations and redundant operations, we apply a two-stage trajectory pruning and refinement process. First, trajectories yielding incorrect answers are discarded. Second, we employ Claude-4.5-Sonnet to evaluate the utility of retrieved information, removing trajectories where search actions fail to support the query (i.e., $\exists I_i \in T, \text{Support}(I_i, Q) = \text{False}$). Remaining high-quality trajectories are rewritten for fluency and logical consistency.

The final stage involves rigorous human verification. A professional annotator reviews the synthesized data for format correctness, ground-truth consistency, and the logical necessity of search operations. This process yields 2,675 high-quality samples, as detailed in Appendix E.

4.2.2 Training Objective

We perform full-parameter Supervised Fine-Tuning (SFT) on the Qwen3 series models. Each sample s in the training dataset $\mathcal{S}$ consists of a question x and a structured trajectory $t = (t_1, t_2, \ldots, t_n)$. The model optimizes the standard autoregressive language modeling objective:

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{n} \log P(t_i | x, t_{<i}) \qquad (2)$$

Model	Parameters	LexEval	LawBench	Unilaw-R1-Eval	DiscLaw	Avg.(%)
			Legal LLMs			
Fuzi-7B	7B	13.56	41.22	6.50	17.55	19.71
LawLLM-7B	7B	49.72	77.87	29.12	48.25	51.24
DiscLaw-13B	13B	35.07	59.49	22.38	33.11	37.51
LegalΔ-3B	3B	33.58	71.39	26.50	51.56	45.76
LegalΔ-7B	7B	48.02	75.78	37.88	45.75	51.86
LegalΔ-14B	14B	**49.74**	**80.42**	**45.88**	**52.14**	**57.05**
			Qwen3-4B			
Base	4B	38.96	61.67	30.50	49.47	45.15
LRAS-SFT	4B	51.37	72.38	38.25	61.64	55.91↑23.8%
LRAS-RL	4B	**51.40**	**76.83**	**44.38**	**65.72**	**59.58**↑32.0%
			Qwen3-8B			
Base	8B	49.22	71.96	39.50	56.93	54.40
LRAS-SFT	8B	55.29	75.94	49.38	67.18	61.95↑13.9%
LRAS-RL	8B	**57.44**	**78.76**	**51.00**	**73.23**	**65.11**↑19.7%
			Qwen3-14B			
Base	14B	52.28	73.88	43.00	66.59	58.94
LRAS-SFT	14B	57.07	78.76	49.38	71.20	63.80↑8.2%
LRAS-RL	14B	**59.84**	**80.10**	**54.37**	**75.66**	**67.49**↑14.5%

Table 2: Evaluation results across different benchmarks and model sizes. Highlighted rows indicate our models. Red arrows denote relative percentage improvement over the corresponding base model.

By minimizing this loss, the model learns to generate complete trajectories with proper tag usage. Crucially, this phase instills the "introspective" capability: the model learns to internalize the expert behavior of recognizing knowledge boundaries and initiating searches only when necessary, rather than hallucinating answers to complex legal queries.

4.3 Difficulty-aware Reinforcement Learning

4.3.1 Difficulty-aware Data Synthesis

To maximize training efficiency, we construct a dataset specifically focused on hard samples where the supervised model struggles. We utilize the trained SFT model to sample $n = 8$ trajectories for each query in the original dataset. We filter these queries based on performance, retaining only those with a pass rate below 50%. This selection strategy targets complex scenarios where the SFT model, despite invoking the search tool, failed to derive the correct answer. Consequently, the RL phase concentrates on correcting logic gaps and refining search strategies for long-tail knowledge.

4.3.2 Training Objective

We employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024). In each training iteration, for a question x, we sample K candidate trajectories $\{y_k\}_{k=1}^{K}$ from the current policy π_ϕ. Each trajectory receives a reward R_k. We compute the group-normalized advantage $A_k = (R_k - \bar{R})/\sigma_R$, where $\bar{R}$ and σ_R are the mean and standard deviation of rewards within the group. The policy is updated to maximize:

$$\mathcal{L}_{\text{GRPO}}(\phi) = \mathbb{E}_{x\sim\mathcal{D},\{y_k\}_{k=1}^{K}\sim\pi_\phi}\Bigg[\frac{1}{K}\sum_{k=1}^{K}\min\Big(\rho_k A_k, \text{clip}(\rho_k, 1-\epsilon, 1+\epsilon)A_k\Big)\Bigg] - \lambda \text{KL}(\pi_\phi||\pi_{\text{ref}}) \tag{3}$$

where ρ_k is the probability ratio between the new and old policies, and λ controls the KL divergence penalty.

To guide the model toward both accuracy and structural integrity, we employ a composite reward function $R(x, y)$ that evaluates the final answer and format compliance:

$$R(x,y) = \begin{cases} 1 & \text{if } a_{\text{pred}} = a_{\text{gold}} \wedge f_{\text{fmt}}(y), \\ 1-\lambda_f & \text{if } a_{\text{pred}} = a_{\text{gold}} \wedge \neg f_{\text{fmt}}(y), \\ \lambda_f & \text{if } a_{\text{pred}} \neq a_{\text{gold}} \wedge f_{\text{fmt}}(y), \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

where $f_{\text{fmt}}(y)$ validates the correct usage of tags (e.g., <think>). This reward structure incentivizes the model to produce correct answers while strictly adhering to the required output format.

5 Experiments

5.1 Experimental Setup

Benchmarks and Metrics. We conduct evaluations on three in-distribution Chinese legal domain

multi-task benchmarks: LexEval (Li et al., 2024), LawBench (Fei et al., 2024), and UniLaw (Cai et al., 2025), along with one out-of-distribution (OOD) benchmark, DiscLaw (Yue et al., 2024). The detailed task descriptions are provided in Appendix A. For evaluation metrics, we follow the original metrics specified by each benchmark for their respective tasks.

Baselines. We evaluate LRAS-SFT and LRAS-RL against representative baselines from two categories. First, we compare with existing legal language models: Fuzi-7B (Wu et al., 2023), LawLLM-7B (Yue et al., 2024), and DiscLaw-13B (Yue et al., 2023). Second, we include specialized legal reasoning models: LegalΔ-3B/7B/14B (Dai et al., 2025). We also report results for the base model to demonstrate the effectiveness of our legal domain adaptation.

Implementation Details. We adopt the Qwen3 series (4B/8B/14B) as backbone models to ensure comprehensive evaluation across different model scales. For legal-aware imitation learning, we perform full-parameter fine-tuning with a learning rate of 1×10^{-5}, using cosine learning rate scheduling with 10% warmup ratio over 3 epochs. For difficulty-aware reinforcement learning, we employ the GRPO advantage estimator without KL penalty in the reward function. The actor model is optimized with a learning rate of 1×10^{-6}, while the critic uses 1×10^{-5}. We set the rollout sample size to 5 per prompt and train for 15 epochs. Additional hyperparameters are provided in Appendix E.1.

5.2 Main Results and Findings

LRAS achieves state-of-the-art performance across all benchmarks, substantially outperforming both pure reasoning models and existing legal domain models. Table 2 presents the evaluation results of LRAS across multiple benchmarks. LRAS-RL (14B) attains an average score of 67.49%, surpassing the previous best model LegalΔ-14B by 18.3% relatively. The progression from base models to LRAS-SFT to LRAS-RL demonstrates the cumulative effectiveness of both training stages. Specifically, Introspective imitation learning teaches models to identify knowledge boundaries and initiate searches appropriately, yielding substantial improvements over base models (e.g., 8.2% relative improvement for 14B). Subsequently, difficulty-aware RL enables models to conduct autonomous multi-round

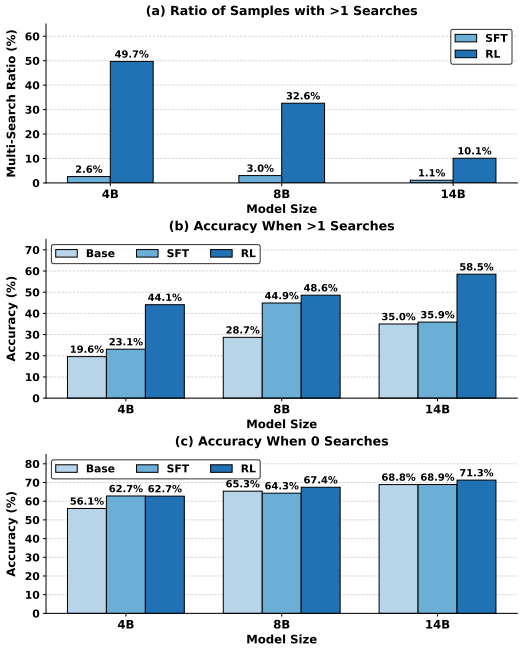

Figure 4: Analysis of search behaviors across model sizes. (a) Proportion of samples with > 1 search actions. (b) Accuracy on samples requiring multi-step search. (c) Accuracy on samples with 0 searches.

exploratory search for complex scenarios which need deep reasoning, providing additional relative gains of 5.8% on average. Notably, even our smallest variant LRAS-RL (4B) achieves 59.58%, outperforming the larger LegalΔ-14B (57.05%) by 4.4% relatively, demonstrating that our agentic search paradigm—shifting from closed-loop thinking to active inquiry—is more effective than scaling model parameters alone. Furthermore, consistent improvements across all model sizes (4B, 8B, 14B) validate the robustness of our approach.

LARS demonstrates strong generalization capability, maintaining robust performance on OOD legal scenarios. LRAS-RL (14B) achieves 75.66% on the DiscLaw benchmark, representing a 13.6% relative improvement over the base model and outperforming all baseline legal models. Importantly, both training stages contribute to enhanced OOD performance: introspective imitation learning improves DiscLaw scores by 6.9% relatively, while difficulty-aware RL adds another 6.3%. This indicates that learning to actively inquire and verify information—rather than relying solely on parametric memory—fundamentally strengthens the model's ability to handle diverse legal scenarios. The consistent gains across both in-distribution benchmarks (LexEval, LawBench, and Unilaw-R1-Eval) and OOD DiscLaw suggest that our approach develops generalizable legal reasoning capabilities.

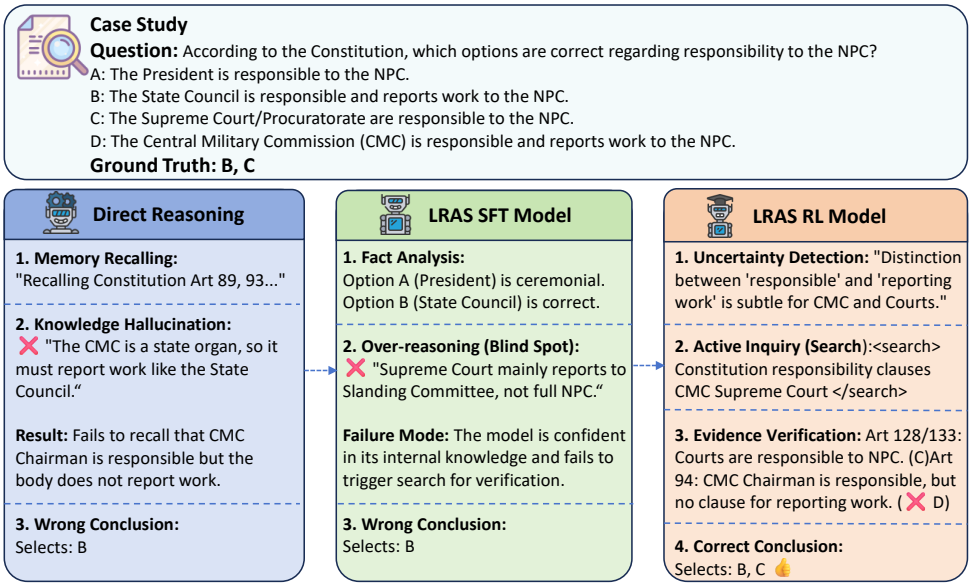

Figure 5: A case study comparing reasoning processes.

Difficulty-aware RL fundamentally stimulates LRAS's active inquiry capability, transforming it from a passive responder into an adaptive investigator. As shown in Figure 4(a), SFT models rarely engage in deep exploration (<3% multi-search rate), whereas RL drastically increases this behavior (e.g., 49.7% for 4B); At the same time, Figure 4(b) confirms that this active inquiry directly translates to performance gains on complex queries, with the RL-14B model outperforming SFT by over 22% (58.5% vs. 35.9%). Furthermore, Figure 4(c) demonstrates that this aggressive search strategy does not compromise performance on simple, knowledge-internal queries (0 searches), where the RL model maintains or slightly exceeds the high accuracy of the SFT baseline.

Ablation Studies of the Introspection Levels. To investigate the effectiveness of our hierarchical reasoning design, we conduct an ablation study across three progressive levels of introspection. As shown in Table 3, increasing the depth of self-reflection consistently yields performance gains across all benchmarks. Specifically, the full **Level-3** configuration achieves the highest average score of **60.20%**, significantly outperforming the baseline Level-1 by a margin of **+4.16%**. This trend is particularly pronounced on the more complex UniLaw dataset (**+7.88%**), validating that deeper introspection is crucial for handling intricate legal scenarios.

5.3 Case Study

Figure 5 illustrates a case study comparing model performance on a constitutional law query. Both baseline models failed: the Direct Reasoning model hallucinated non-existent reporting duties for the Central Military Commission, while LRAS-SFT suffered from overconfidence, failing to verify its internal knowledge. Conversely, LRAS-RL succeeded by recognizing the query's complexity. It utilized Active Inquiry to retrieve specific legal articles, allowing it to correctly distinguish between the concepts of "responsibility" and "reporting work," thereby selecting the correct answer.

6 Conclusion

In this work, we introduce LRAS, a framework designed to shift the paradigm of legal reasoning from static "closed-loop thinking" to dynamic "Active Inquiry." By integrating Introspective Imitation Learning with Difficulty-aware Reinforcement Learning, LRAS effectively addresses the fundamental introspection deficit by teaching models whether to search, while simultaneously overcoming the fragility in complex scenarios by enabling them to learn how to search autonomously. Extensive experiments demonstrate that our approach significantly outperforms static retrieval methods, successfully bridging the reasoning deficit in deep legal reasoning tasks. Ultimately, we believe LRAS establishes a robust paradigm for autonomous legal agents, contributing meaningfully to the advancement of trustworthy and precise Legal AI.

7 Limitations

Despite notable advancements, LRAS faces several limitations. Due to computational resource

constraints, we limited the models size and types of tools. Consequently, the model may not fully explore extremely complex scenarios requiring deep, multi-layered inquiry chains. Future work will explore extended interaction horizons to handle broader long-tail complexities. Additionally, LARS's active inquiry relies heavily on the precision of external tools. If the retrieval system returns incomplete or irrelevant statutes, the model's subsequent reasoning may be compromised. We will integrate robustness mechanisms to better detect and correct tool-induced errors in our future work.

8 Ethical considerations

The dataset construction relied exclusively on publicly available or appropriately licensed resources, ensuring full compliance with licensing requirements. No personally identifiable or sensitive data were collected during the study. Human annotators were only involved in the verification of the dataset quality, and all annotators were recruited from a pool of qualified experts with backgrounds in law or computer science.

References

Pawitsapak Akarajaradwong, Chompakorn Chaksangchaichot, Pirat Pothavorn, Attapol Thamrongrattanarit-Rutherford, Ekapol Chuangsuwanich, and Sarana Nutanong. 2025. Can group relative policy optimization improve thai legal reasoning and question answering? *arXiv preprint arXiv:2507.09638*.

Andrew Blair-Stanek and Benjamin Van Durme. 2025. Llms provide unstable answers to legal questions. *arXiv preprint arXiv:2502.05196*.

Hua Cai, Shuang Zhao, Liang Zhang, Xuli Shen, Qing Xu, Weilin Shen, Zihao Wen, and Tianke Ban. 2025. Unilaw-r1: A large language model for legal reasoning with reinforcement learning and iterative inference. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18128–18142.

CAIL. 2021. Cail2021. https://github.com/china-ai-law-challenge/CAIL2021.

Chuxue Cao, Mengze Li, Juntao Dai, Jinluan Yang, Zijian Zhao, Shengyu Zhang, Weijie Shi, Chengzhong Liu, Sirui Han, and Yike Guo. 2025a. Towards advanced mathematical reasoning for llms via first-order logic theorem proving. *arXiv preprint arXiv:2506.17104*.

Chuxue Cao, Han Zhu, Jiaming Ji, Qichao Sun, Zhenghao Zhu, Wu Yinyu, Josef Dai, Yaodong Yang, Sirui Han, and Yike Guo. 2025b. Safelawbench: Towards safe alignment of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14015–14048.

Zhengyu Chen, Jinluan Yang, Teng Xiao, Ruochen Zhou, Luan Zhang, Xiangyu Xi, Xiaowei Shi, Wei Wang, and Jinggang Wang. 2025. Can tool-integrated reinforcement learning generalize across diverse domains? *arXiv preprint arXiv:2510.11184*.

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

Xin Dai, Buqiang Xu, Zhenghao Liu, Yukun Yan, Huiyuan Xie, Xiaoyuan Yi, Shuo Wang, and Ge Yu. 2025. Legal *delta*: Enhancing legal reasoning in llms via reinforcement learning with chain-of-thought guided information gain. *arXiv preprint arXiv:2508.12281*.

Fatemeh Dehghani, Roya Dehghani, Yazdan Naderzadeh Ardebili, and Shahryar Rahnamayan. 2025. Large language models in legal systems: A survey. *Humanities and Social Sciences Communications*, 12(1):1977.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, and 1 others. 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.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Sophia Simeng Han, Yoshiki Takashima, Shannon Zejiang Shen, Chen Liu, Yixin Liu, Roque K Thuo, Sonia Knowlton, Ruzica Piskac, Scott J Shapiro, and Arman Cohan. 2025. Courtreasoner: Can llm agents reason like judges? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35279–35294.

Zhitian Hou, Zihan Ye, Nanli Zeng, Tianyong Hao, and Kun Zeng. 2025. Large language models meet legal artificial intelligence: A survey. *arXiv preprint arXiv:2509.09969*.

Yinghao Hu, Leilei Gan, Wenyi Xiao, Kun Kuang, and Fei Wu. 2025. Fine-tuning large language models for improving factuality in legal question answering. *arXiv preprint arXiv:2501.06521*.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the nineteenth international conference on artificial intelligence and law*, pages 417–421.

HKAIR Lab. 2024. Hk-o1aw models: Leveraging o1 slow thinking in the development of hong kong legal large language models. https://github.com/HKAIR-Lab/HK-O1aw.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:25061–25094.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Weijie Shi, Han Zhu, Jiaming Ji, Mengze Li, Jipeng Zhang, Ruiyuan Zhang, Jia Zhu, Jiajie Xu, Sirui Han, and Yike Guo. 2025. Legalreasoner: Step-wised verification-correction for legal judgment reasoning. *arXiv preprint arXiv:2506.07443*.

Shiguang Wu, Zhongkun Liu, Zhen Zhang, Zheng Chen, Wentao Deng, Wenhao Zhang, Jiyuan Yang, Zhitao Yao, Yougang Lyu, Xin Xin, Shen Gao, Pengjie Ren, Zhaochun Ren, and Zhumin Chen. 2023. fuzi.mingcha.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and 1 others. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Huiyuan Xie, Chenyang Li, Huining Zhu, Chubin Zhang, Yuxiao Ye, Zhenghao Liu, and Zhiyuan Liu. 2025. Lawchain: Modeling legal reasoning chains for chinese tort case analysis. *arXiv preprint arXiv:2510.17602*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xinyu Yang, Chenlong Deng, and Zhicheng Dou. 2025b. Glare: Agentic reasoning for legal judgment prediction. *arXiv preprint arXiv:2508.16383*.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and 1 others. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, and 1 others. 2024. Lawllm: Intelligent legal system with legal reasoning and verifiable retrieval. In *International Conference on Database Systems for Advanced Applications*, pages 304–321. Springer.

Kepu Zhang, Guofu Xie, Weijie Yu, Mingyue Xu, Xu Tang, Yaxin Li, and Jun Xu. 2025a. Legal mathematical reasoning with llms: Procedural alignment through two-stage reinforcement learning. *arXiv preprint arXiv:2504.02590*.

Kepu Zhang, Weijie Yu, Zhongxiang Sun, and Jun Xu. 2025b. Syler: A framework for explicit syllogistic legal reasoning in large language models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4117–4127.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jecqa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

A Datasets

A.1 Lexeval

LexEval is a comprehensive Chinese legal benchmark comprising 23 tasks and approximately 14,150 questions. It introduces the Legal Cognitive Ability Taxonomy (LexCog) to organize tasks systematically. The data is derived from existing legal datasets, real-world examinations, and expert annotations. LexEval evaluates models not only on fundamental legal knowledge application but also on their ability to handle ethical issues within legal contexts.

A.2 Lawbench

LawBench is designed to evaluate legal capabilities across three specific cognitive levels: (1) Legal Knowledge Memorization (recall of concepts, articles, and facts); (2) Legal Knowledge Understanding (comprehension of entities and relationships); and (3) Legal Knowledge Application (reasoning to solve realistic tasks). It contains 20 diverse tasks categorized into single-label classification, multi-label classification, regression, extraction, and generation.

A.3 UniLaw-Eval

UniLaw-R1-Evalconsists of 800 items, split into 426 single-choice and 374 multi-choice questions. It is designed to rigorously test the model's ability to perform logical inference within legal scenarios.

A.4 Disc-Law

DISC-Law (also referred to as Law-Eval) assesses legal LLMs from both objective and subjective dimensions: Objective evaluation consists of multiple-choice questions sourced from standardized examinations (e.g., National Judicial Examination, CPA). Questions are categorized into Easy, Normal, and Hard levels to test knowledge retrieval and deduction. Subjective evaluation comprises 300 manually constructed examples covering legal consultation and judgment prediction. Evaluation utilizes a referee model (e.g., GPT-3.5) to score responses based on accuracy, completeness, and clarity.

A.5 Selected Tasks

We focus on knowledge-intensive tasks that require legal knowledge retrieval and reasoning. Specifically, we evaluate on:

Method	UniLaw	LexEval	LawBench	Avg.
Level-1	41.50	52.03	74.60	56.04
Level-2	43.00	53.40	75.08	57.16
Level-3	**49.38**	**55.29**	**75.94**	**60.20**
Δ (L3 - L1)	*+7.88*	*+3.26*	*+1.34*	*+4.16*

Table 3: Ablation on introspection levels.

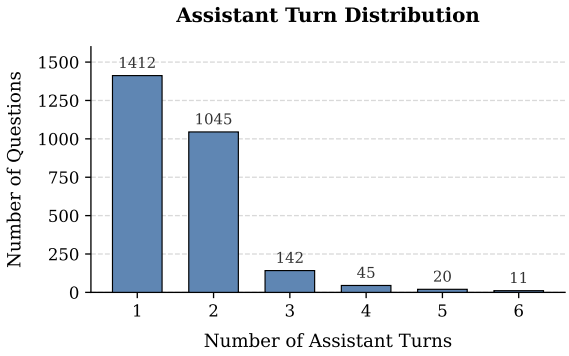

Figure 6: Distribution of assistant turns in the SFT dataset, illustrating the frequency of conversation lengths.

- **LexEval**: Tasks 1-1, 1-3, 3-1, 3-2, 3-3, 3-4, 3-5, 3-6, 4-1, and 4-2
- **LawBench**: Tasks 1-2, 3-1, 3-3, 3-4, 3-5, and 3-6
- **UniLaw-Eval**: Complete benchmark (all 800 questions)
- **DISC-Law**: Complete benchmark (both objective and subjective evaluation sets)

B Dataset Licenses

Several foundational Chinese legal datasets are publicly available under open-source licenses, including JEC-QA, CAIL2018, and CAIL2021. Additionally, LexEval is released under the MIT license, while both LawBench and DISC-LawLLM are available under the Apache-2.0 license.

C Ablation Results

The ablation results on introspection levels are presented in Table 3.

D Details of the Training Data

This section presents the statistical characteristics of the Legal Search-R1 SFT dataset, which contains 2,675 samples. As illustrated in Figure **??**,

Metric	Mean	Max
Interaction Dynamics		
Conversation Turns	1.60	6
Search Calls (per question)	1.27	5
Length Characteristics (Chars)		
Total Response	615	1,512
Reasoning Process (`<think>`)	703	1,537
Search Query	19.6	76

Table 4: Statistics of the Legal Search-R1 SFT dataset. Search statistics are computed on search-based samples.

the dataset exhibits a diverse distribution of conversation lengths, ranging from single-turn queries to complex interactions of up to 6 turns, with a mean of 1.60 turns. Table 4 further details these metrics: the reasoning process (encapsulated in `<think>` tags) averages 703 characters, reaching a maximum of 1,537, indicating substantial cognitive depth. Additionally, the inclusion of search calls (mean=1.27, max=5) in search-based samples supports the robust training of agentic behaviors, while the total response length averages 615 characters.

E Details of Training Setup

In this section, we provide the detailed hyperparameters and configurations used for training Legal Search-R1. We utilize the Qwen3 series (4B/8B/14B) as our backbone models. All experiments are conducted on a cluster of 8 NVIDIA GPUs.

E.1 Hyperparameters

Table 5 lists the specific hyperparameters used for both the Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) stages.

For the SFT stage, we employ DeepSpeed ZeRO-3 optimization to handle the full-parameter fine-tuning efficiently. We set the maximum sequence length to 32,768 tokens to accommodate long legal contexts.

For the RL stage, we utilize the GRPO algorithm. A key distinction in our setup is the separate learning rates for the actor (1×10^{-6}) and the critic (1×10^{-5}) to stabilize training. The reward function is a weighted sum of accuracy and formatting constraints: $R = 1.0 \times \text{Score}_{\text{acc}} + 0.2 \times \text{Score}_{\text{struct}} + 0.1 \times \text{Score}_{\text{format}}$.

Hyperparameter	Value
Supervised Fine-Tuning (SFT)	
Optimizer	AdamW
Learning Rate	1×10^{-5}
LR Scheduler	Cosine
Warmup Ratio	0.1
Batch Size (Global)	128
Epochs	3
Max Sequence Length	32,768
Precision	bf16
DeepSpeed Stage	ZeRO-3
Reinforcement Learning (RL)	
Algorithm	GRPO
Actor Learning Rate	1×10^{-6}
Critic Learning Rate	1×10^{-5}
KL Penalty in Reward	False
Actor KL Loss Coefficient	0.001
Rollout Sample Size (N)	5
Max Prompt Length	8,192
Max Response Length	8,192
Total Epochs	15
Global Train Batch Size	256

Table 5: Hyperparameters for SFT and RL training stages.

E.2 Tool Configuration

The agent is equipped with a legal search tool powered by an external retrieval service (SerpAPI combined with Jina Reader). The configuration for the multi-turn interaction is as follows:

- **Max Assistant Turns:** 6 (aligned with the maximum turns in the SFT dataset).
- **Retrieval Top-K:** 5 results per query.
- **Timeout:** 600 seconds per request to ensure stability during extensive legal database queries.

F Implementation Details

F.1 Search and Summarization Pipeline

To equip the model with up-to-date legal knowledge, we implement a robust retrieval pipeline consisting of three stages: search, extraction, and summarization.

1. **Search and Extraction:** We utilize **SerpAPI** to query the internet, retrieving the top-5 most relevant search results for each query generated by the agent. To process the raw URLs, we employ **Jina Reader**, which effectively extracts clean text content from the webpages, filtering out HTML boilerplate and navigation elements.

2. **Summarization:** Since raw web content often contains redundancy and noise, we introduce a summarization step before feeding information back to the agent. Following previous works, we deploy an external **Qwen3-32B** model as a dedicated summarizer. This model condenses the extracted content into concise, query-focused legal abstracts.

3. **Summarizer Configuration:** For the Qwen3-32B summarizer, we adhere to the official best practices to balance diversity and coherence. We disable the "thinking" mode (i.e., `enable_thinking=False`) and use the following generation hyperparameters: temperature: 0.7, top-p: 0.8, top-k: 20, min-p: 0, and a maximum of $8{,}192$ new tokens.

F.2 Inference Settings for Main Models

For the evaluation of our trained Legal Search-R1 models (including the backbone Qwen3 series), we prioritize determinism and reproducibility. Therefore, during the inference stage:

- We use **Greedy Decoding** by setting the Temperature to 0.
- The maximum generation length is set to $8,192$ tokens to allow for comprehensive legal reasoning and tool usage.

G Additional Baselines with Identical Retrieval Infrastructure

To ensure fair comparison, we evaluate LRAS against retrieval-augmented and agentic baselines that use the same retrieval infrastructure (SerpAPI + Jina Reader).

- **Full-RAG:** A retrieval-augmented generation approach that performs external search for every query, regardless of whether the model's parametric knowledge is sufficient.
- **Prompt-based Agentic Search:** A prompting strategy that instructs the base model to perform agentic search (deciding when and what to retrieve) without any additional training.
- **L-MARS (?):** A multi-agent orchestration framework that decomposes legal queries into subproblems and coordinates multiple specialized agents for targeted retrieval and verification before answer synthesis.

LRAS-RL outperforms all baselines despite using the same retrieval infrastructure, demonstrating that performance gains stem from our training paradigm rather than retrieval advantages. Notably, prompt-based agentic search (46.03%) substantially underperforms trained models, confirming that legal introspective capabilities require explicit training rather than prompt engineering alone.

H Computational Efficiency Analysis

LRAS is designed to invoke retrieval only when necessary, avoiding the overhead of unconditional retrieval. Table 7 reports average output token counts as a proxy for computational cost.

Training Cost.

- **SFT Stage:** $4\times$ H800 GPUs, approximately 2–3 hours per full training run.
- **RL Stage:** $8\times$ H800 GPUs, approximately 48 hours per full training run.

Inference Cost. LRAS (Qwen3-8B) can be deployed using vLLM with a minimum of $1\times$ H800 GPU (80 GB VRAM), comparable to LegalDelta-7B. The selective retrieval design (47.8% retrieval rate) further reduces average latency relative to full-RAG approaches.

I Data Contamination Analysis

To verify the integrity of our evaluation, we conducted a rigorous two-stage contamination analysis. Our SFT data synthesis exclusively samples from the **training splits** of CAIL2018, CAIL2021, and JEC-QA. We then performed post-hoc decontamination using both MinHash and N-gram matching (similarity threshold > 0.85) at two granularities: (1) question-only, and (2) question + options + answer.

The overall contamination rate of 0.11% is negligible. LRAS achieves consistent improvements across all benchmarks, including those with zero contamination (LexEval, UniLaw-R1-Eval), confirming that performance gains reflect genuine reasoning improvement rather than data memorization.

J Detailed Data Construction Pipeline

The Introspective Imitation Learning stage relies on a 6-stage data construction pipeline designed to ensure legal precision and annotation reliability.

Method	Retrieval	UniLaw	DiscLaw	Avg.
RAG (Qwen3-8B)	✓	47.75	71.57	59.66
Prompt-based Agentic Search (Qwen3-8B)	✓	34.38	57.67	46.03
L-MARS (Qwen3-8B)	✓	31.74	62.45	47.10
LRAS-SFT (Qwen3-8B)	✓	49.38	67.18	58.28
LRAS-RL (Qwen3-8B)	✓	**51.00**	**73.23**	**62.12**

Table 6: Comparison with agentic and RAG-based baselines using identical retrieval infrastructure on UniLaw-R1-Eval and DISC-Law benchmarks.

Model	Avg. Output Tokens
Qwen3-8B (Base)	1,202
LRAS-RL-8B (Ours)	1,584
LegalDelta-7B (Prev. SOTA)	2,499
Qwen3-235B	3,911

Table 7: Average output token length comparison. LRAS introduces only 32% additional tokens over the base model while generating 36% fewer tokens than the previous SOTA (LegalDelta-7B).

1. **Source Sampling.** We sample from the training splits of JEC-QA, CAIL2018, and CAIL2021, covering diverse legal task categories (legal knowledge QA, judgment prediction, statute retrieval). We apply uncertainty-aware filtering: for each question, we generate n=8 candidate responses using Qwen3-235B-Instruct and retain only questions with a pass rate strictly below 100%, identifying genuine knowledge boundary cases.

2. **Trajectory Synthesis.** We use Qwen3-235B-Instruct with structured prompts (detailed in Appendix K) to generate reasoning trajectories with explicit format, incorporating the three-level introspection mechanism for legal knowledge boundary recognition.

3. **Quality Control.** We apply two-stage filtering: (a) discard trajectories with incorrect final answers; (b) use Claude-3.5-Sonnet to evaluate retrieval utility, removing trajectories where search results do not logically support the legal reasoning chain.

4. **Model Rewriting.** We use Claude-3.5-Sonnet to rewrite trajectories, ensuring format correctness, legal terminology accuracy, and reasoning coherence that meet domain standards.

5. **Human Verification.** Professional legal annotators verify ground-truth consistency and the logical necessity of search operations. To assess annotation reliability, two independent annotators reviewed a 30% random sample of the dataset. Inter-rater agreement measured by Cohen's Kappa achieved $\kappa = 0.77$, indicating strong agreement.

6. **Balanced Data Curation.** The final dataset carefully balances non-retrieval (52.2%) and retrieval (47.8%) responses, ensuring models learn adaptive tool usage based on legal reasoning requirements rather than arbitrary patterns.

Benchmark	Total Eval Samples	Contaminated Samples	Contamination Rate
DiscLaw	2,563	1	0.04%
LawBench	3,000	13	0.43%
LexEval	6,000	0	0.00%
UniLaw-R1-Eval	800	0	0.00%
Overall	**12,363**	**14**	**0.11%**

Table 8: Data contamination analysis across evaluation benchmarks. The overall contamination rate of 0.11% is negligible.

K Prompts

This section details the specific prompts designed for our pipeline, categorized into data construction, data rewriting, and reinforcement learning. For the data construction phase, we implemented a three-level introspection mechanism to progressively enhance the model's capability to recognize its knowledge boundaries.

K.1 Data Construction Prompts

Our introspection mechanism is divided into three levels, ranging from implicit tool usage to deep metacognitive evaluation.

K.2 Data Rewrite Prompts

This section includes prompts for assessing information relevance and refining the reasoning trajectories.

K.3 RL Prompts

This section details the prompts used during the Reinforcement Learning stage, including the RL prompt and the summarization prompt.

L Case Study

Level 1

```
UNIFIED_SYSTEM = """You are a professional legal assistant. Your task is to accurately answer legal questions.

You possess a search engine tool. When you encounter legal articles, cases, or facts that need verification, you can use
      the search function.
Your Action Guide:
1. If you believe external information is needed to assist your answer, please directly output <search>Search Query</search
      >.
2. The system will return <information>Search Results</information>.
3. Combine the search results or your existing knowledge to output <answer>Your Answer</answer>.

Please begin answering directly.
"""
```

Level 2

```
UNIFIED_SYSTEM = """You are a professional legal assistant. Your task is to accurately answer legal questions.

You must first write your reasoning process within the <think>Reasoning Process</think> tag. After reasoning, judge whether
       you lack legal knowledge, and then choose one of the following actions:

**Action 1: Use search tool to find legal information**
If you need to find legal articles, cases, or professional materials to support your answer, use the search tool after
      thinking. Search results will be returned between <information> Search Summary Content </information>; you may search
       multiple times as needed.

**Action 2: Provide answer directly**
If external knowledge is no longer required, provide the answer directly after thinking.

### Output Format

**Search Action Format:**
<think>Reasoning Process</think>
<search>Search Query</search>

**Provide Answer Format:**
<think>Reasoning Process</think>
<answer>Your Answer</answer>

### Examples

**Search Action Example:**
<think>
This question involves regulations regarding changes in insurance company shareholders. Option D mentions that "changing
      contributors or shareholders holding more than 5\% of the company's shares" requires approval. This should be
      explicitly stipulated in the "Administrative Measures for Equity of Insurance Companies" or relevant laws. I need to
      find specific articles to confirm.
</think>
<search>Approval for change of shareholders holding more than 5% in insurance companies</search>

**Provide Answer Example:**
<think>
According to the search results, Article 16 of the "Administrative Measures for Equity of Insurance Companies" explicitly
      stipulates: Changing shareholders who hold more than 5\% of the company's shares shall be approved by the insurance
      regulatory institution. The statement in Option D is consistent with the legal provision.
</think>
<answer>D</answer>

### Key Requirements
1. **The <think> tag must contain substantive reasoning content** and cannot be empty.
2. **Output only one <search> or one <answer> at a time**; do not output both simultaneously.
3. **Do not fabricate the <information> tag yourself**; search results will be automatically provided by the system.
"""
```

Level 3

```
UNIFIED_SYSTEM = """You are a professional legal assistant capable of efficiently solving knowledge-intensive tasks. You
      will answer legal questions.

Note: These questions are of a certain difficulty and require external knowledge to assist reasoning.

**Core Principle - Search Strategy**:
You can utilize your own knowledge or call an external search engine to collect additional information, but searching
      should only be performed when necessary. Specifically, you should only search when you encounter obvious knowledge
      gaps or uncertainties that prevent you from answering the question confidently.

**Judgment Criteria**:
Please judge whether to search based on your understanding of the question and your own knowledge reserve. The key is to
      evaluate:
- Do you have enough certainty and confidence in the answer?
- Will the lack of search affect the accuracy or completeness of the answer?
- Can searching bring substantial information gain?

Prioritize relying on your internal knowledge to answer. Only conduct a search when you clearly realize there is a
      knowledge gap and this gap will significantly affect the quality of the answer.

To derive the answer, you will proceed step-by-step according to a structured loop.

**Important: You must strictly use the following tag format for output, do not use any other format:**
- `<think>Thinking Content</think>` - Used for thinking and analysis
- `<search>Search Query</search>` - Used for searching (optional)
- `<answer>Final Answer</answer>` - Used to provide the answer

**Thinking Phase (<think>)**: You must first perform the following within the `<think>` tag:
1. Analyze the core requirements of the question.
2. Recall relevant legal knowledge you have mastered.
3. Evaluate whether your knowledge is sufficient to answer confidently.
4. Clearly judge whether a search is needed and explain the reason.
5. If knowledge is sufficient, proceed directly to the answering phase.

**Search Phase (<search>)**: Only construct a precise search query within the `<search>` tag after confirming the existence
       of a knowledge gap or uncertainty during thinking. If not needed, skip this phase.

**Information Phase (<information>)**: If you performed a search, the results will be returned within the `<information>`
      tag.

**Answering Phase (<answer>)**: When you have sufficient knowledge, provide a concise and accurate answer within the `<
      answer>` tag.

Here are a few examples:

---
**Example 1: Search Required**
<think>[Analysis and Judgment]</think>
<search>[Search Query]</search>
<information>[Search Results]</information>
<think>[Continue Analysis]</think>
<answer>C</answer>

---
**Example 2: No Search Required**
<think>[Analysis and Judgment]</think>
<answer>B</answer>

---
**Example 3: Multiple Searches**
<think>[Analysis and Judgment]</think>
<search>[First Search]</search>
<information>[Search Results]</information>
<think>[Continue Analysis]</think>
<search>[Second Search]</search>
<information>[Search Results]</information>
<think>[Final Analysis]</think>
<answer>ACD</answer>

---
**Key Tips**:
- You can search from 0 to N times, deciding autonomously based on actual needs.
- Each search should focus on a clear knowledge gap.
- Final Answer Format: <answer>Answer</answer>
"""
```

Relevance Assessment Prompt

```
prompt = f"""You are a legal QA quality assessment expert. Please judge whether the following retrieved information can
      help answer the legal question.

[Judgment Standards]
The information must contain clear legal articles, judicial interpretations, or authoritative definitions that can directly
       support the key arguments of the answer.

[Specific Requirements]
Must have a clear legal basis.

[Assessment Principles]
- If the information is related to the question topic and provides valuable legal knowledge or basis, judge as YES.
- Only judge as NO if the information is completely off-topic, vague and useless, or unrelated to the question.
- If the information is not perfect but helpful for answering, judge as YES.

Please answer only YES or NO (do not add any explanation):

[Question]
{question}

[Reference Answer]
{final_answer}

[Retrieved Information]
{info_text}

Judgment Result (YES/NO):"""
```

Refinement Prompt (Single Turn)

```
rewrite_prompt = f"""You are a professional legal AI assistant training data optimization expert. Please rewrite the
      analysis process (within the <think> tag) for the following legal question.

Requirements:
- Remove question type priors (do not mention "single choice/multiple choice", etc.).
- Reasoning should be natural, professional, and avoid repetitive legal articles.
- Retain <think></think> and <answer></answer>.
- Do not output <search> or <information>.
- The final answer must be exactly consistent: {original_answer}

Question:
{question}

Original Response:
{assistant_content}

Please output the optimized complete assistant response (containing only think and answer):"""
```

Refinement Prompt (Multi-Turn)

```
rewrite_prompt = f"""You are a professional legal AI assistant training data optimization expert. Please rewrite the
      analysis trajectory for the following legal question.

Hard Requirements:
1) Remove question type priors; do not mention "single choice/multiple choice", etc.
2) Output format allows only: <think>, <search>, <answer>.
3) Do not output <information>.
4) **Must output exactly {expected_search_count} times <search>, writing only the query text each time.**
5) The final answer must be exactly consistent with the original answer: {original_answer}

Question:
{question}

Available Retrieval Evidence (Total {expected_search_count} searches):
{evidence_text}

Please output the optimized complete assistant trajectory (must include {expected_search_count} times <search>,
      corresponding sequentially to the evidence above):"""
```

RL Prompt

```
USER_CONTENT_PREFIX = """You are a professional legal assistant, skilled in answering legal questions accurately through
      deep search.

Work Style:
1. First reason and analyze within <think></think>. If you judge that existing knowledge is insufficient to answer
      accurately, proceed to search.
2. If you need to find legal basis, use <search>Query</search> to call the search. Results are returned in <information></
      information>.
3. You may search multiple times until information is sufficient.
4. Finally, provide a concise answer within <answer></answer>.

Question:"""
```

Summarization Prompt

```
summarization_prompt = """You are a legal information extraction expert. Please extract legal information points relevant
      to the user's question from the following webpage content.

User Question:
{original_question}

Requirements:
1. Only extract legal information directly related to the question (articles, regulations, cases, interpretations, etc.).
2. Include:
   - Relevant legal articles and regulation names.
   - Specific legal provisions and requirements.
   - Scope of application, conditions, and exceptions.
   - Relevant legal consequences or liabilities.
3. Do not analyze the question, do not give conclusions, and do not add your own reasoning.
4. Remain objective and neutral; directly extract legal facts.
5. Remove all webpage URL links.
6. Summarize concisely in 6-10 sentences.
7. If the content is unrelated to the question, directly state "The webpage content is unrelated to the question".

Webpage Content:
{content}
"""
```

Case Study

Question: When a foreign investor establishes a foreign-funded enterprise, the capital contribution may be paid in installments, but the final installment shall be paid in full within (_____) from the date of issuance of the business license.
A: One year B: Two years C: Three years D: Five years

Ground Truth: C (Three years)

Base Model (Incorrect):

<*think*> According to Article 20 of the "Law of the People's Republic of China on Foreign-funded Enterprises," when establishing a foreign-funded enterprise, the foreign investor shall pay the capital contribution in full within a certain period from the date of issuance of the business license. Installment payment may be allowed, but the deadline for the final installment may be **one year**. <*/think*>

Prediction: A (One year)

Agentic Search (Correct):

<*think*> This question examines the regulations on capital contribution deadlines when foreign investors establish foreign-funded enterprises. I need to search for the relevant legal provisions to determine the accurate deadline. <*/think*>

<*search*> Foreign investor foreign-funded enterprise installment capital contribution deadline <*/search*>

<*information*> ... Article 30 of the "Detailed Rules for the Implementation of the Law of the People's Republic of China on Foreign-funded Enterprises" stipulates that foreign investors may pay capital contributions in installments, but the final installment shall be paid in full within **three years** from the date of issuance of the business license ... <*/information*>

<*answer*> C <*/answer*>

Prediction: C (Three years)

Figure 7: Case study of inherent imprecision