Who Should Answer? CoT-Guided Two-Stage Routing for Legal QA

Rujing Yao[1], Yang Wu[1], Jinhong Yu[1], Tong Zhang[2],
Zhuoren Jiang[3], Xiaozhong Liu[1]

[1]Worcester Polytechnic Institute, USA
[2]Nankai University, China; [3]Zhejiang University, China

ABSTRACT

Legal question-answering systems powered by LLMs can significantly enhance the efficiency and accessibility of legal services. However, their practical deployment is hindered by prohibitive computational costs and the risk of generating unreliable advice, leading to resource misallocation and safety concerns. To address this, model routing is essential, but generic routing solutions fail to meet the stringent demands of the legal domain. In the paper, we propose a Chain-of-Thought (CoT)-Guided Two-Stage Routing Framework to optimize resource allocation in legal QA. Our framework consists of three modules: (1) an LLM fine-tuned with Group Relative Policy Optimization (GRPO) to generate high-quality CoTs as routing features; (2) a human–machine gate that decides whether to defer a query to a human expert or answer automatically; and (3) a contextual-bandit selector that maximizes expected net utility, trading off predicted answer quality against inference cost. Experimental results demonstrate the effectiveness of our proposed framework.

INTRODUCTION

LLMs have rapidly advanced the field of Legal QA and are capable of functioning as agentic components within enterprise legal workflows. Compared to smaller or simpler models, large LLMs are capable of delivering more nuanced and accurate responses within legal automated QA systems. However, solutions that route every user query to these extremely large LLMs inherently incur prohibitive operational costs and prolonged response times. Furthermore, even the outputs of the most state-of-the-art LLMs may still be unreliable, unsubstantiated, or non-compliant with specific legal standards. Therefore, a workable and practical system must be designed to evaluate each query, select the most appropriate processing path, and provide an auditable, seamless handoff to human experts when necessary.

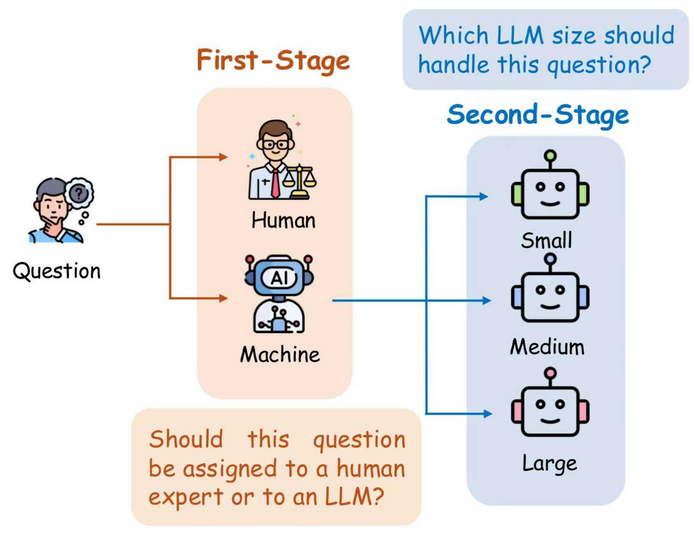

Figure 1. An illustration of our framework.

In this paper, we propose a novel Chain-of-Thought (CoT)-Guided Two-Stage Routing Framework to address these critical gaps. As illustrated in Figure 1, for a given legal query from a user, our framework first determines whether the question should be handled by a human expert or processed automatically by a machine. If an automated response is deemed appropriate, the framework then dynamically selects the most cost-effective LLM from a model pool to generate the answer.

The contributions are summarized as follows:

- We propose a novel CoT-guided two-stage routing framework specifically designed for the legal domain. This framework first decides on human–machine delegation and then performs model selection.
- We introduce CoT as enriched routing features. This provides our router with deeper, legal-aware reasoning signals.
- We collected a real-world legal dataset and conducted experiments on it to validate the effectiveness of our proposed framework. The dataset will be released to foster future research and development in the legal QA domain.

METHODOLOG

We cast resource allocation in a legal QA system as a two-stage routing problem driven by CoT. Figure 2 shows the framework, which comprises (i) GRPO-based CoT generation, (ii) human–machine deferral gate, and (iii) contextual-bandit model selector.

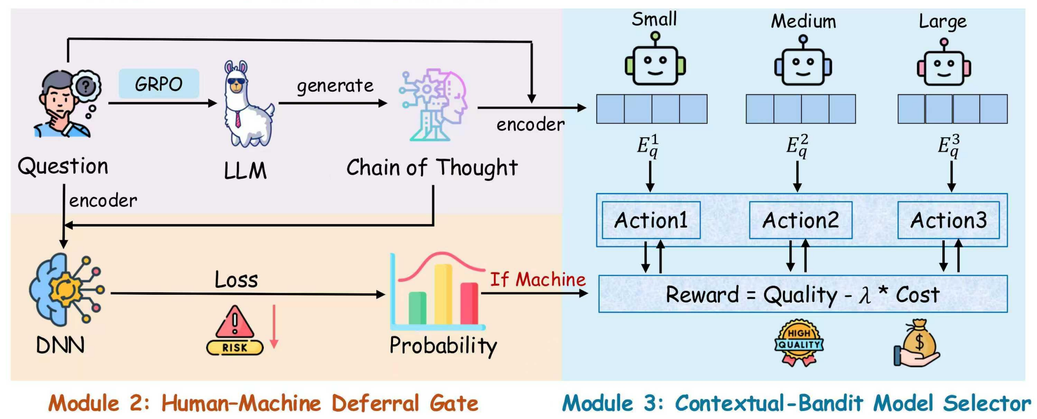

Figure 2. The overall framework.

1) GRPO-based CoT Generation

CoT serves as an explicit reasoning trace for each query. We train a CoT generator with GRPO by sampling multiple candidate rationales per query. A stronger LLM judge scores them with a fixed rubric, and GRPO optimizes the generator toward batch-relative high scorers, producing more reliable CoTs for routing.

2) Human–Machine Deferral Gate

A risk-aware gate uses both the query and the CoT representation to decide auto-answer vs. defer to human. It also enforces a target auto-coverage to avoid over-automation (risky answers) and under-automation (wasted human effort).

3) Contextual-Bandit Model Selector

For queries approved for automation, we choose an LLM from a costed model pool using a context-aware bandit selector. The selector optimizes a quality–cost reward, reserving expensive models only when necessary.

EXPERIMENTS

In the experiments, we employ Qwen2.5-7B, Qwen2.5-32B, and Qwen2.5-72B as LLMs at progressively larger scales for the second-stage selector.

Table 1 reports the results on the JUSTIA marriage-law dataset. Our proposed framework delivers uniformly higher quality across all metrics while maintaining superior cost-efficiency.

Method	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore	Cost
Random Router	13.24	2.04	12.01	18.20	82.48	8.90
LLM Router	13.72	2.24	12.74	19.20	82.80	12.20
ICL-Router	13.95	2.31	12.88	19.55	82.88	10.80
RouterDC	14.08	2.36	13.02	19.82	82.94	9.70
EmbedLLM	14.15	2.40	13.21	20.10	82.95	8.35
Ours	14.85	2.52	14.42	20.80	83.05	6.10

Table 1. Experimental results.

CONCLUSION

In legal QA systems, efficient resource allocation can deliver reliable answers at lower cost while preserving accountable human oversight. In this paper, we introduce a novel CoT-Guided Two-Stage Routing Framework for Legal QA. The framework operates in two phases: first, a coverage-constrained deferral gate routes high-risk or out-of-scope queries to human experts. Second, a contextual-bandit selector dynamically chooses the most cost-effective automated model from a pool. We enhance routing intelligence by using GRPO-trained CoT rationales as features, injecting legal-aware reasoning signals that generic routers often miss. Experiments on a marriage-law dataset validate the performance and cost-effectiveness of our proposed framework.

Contact

rjyao@mail.nankai.edu.cn; xliu14@wpi.edu