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
xliu14@wpi.edu

Abstract

Legal question-answering systems powered by Large Language Models 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

Large Language Models (LLMs) have rapidly advanced the field of Legal Question Answering (Legal QA) and are capable of functioning as agentic components within enterprise legal workflows (Yao et al. 2025b; Wu et al. 2023). For instance, LLMs can provide powerful semantic understanding and knowledge generation capabilities for automated legal QA systems, thereby greatly accelerating consultation tasks (Yao et al. 2025a). 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 (Ong et al. 2025). Furthermore, even the outputs of the most state-of-the-art LLMs may still be unreliable, unsubstantiated, or non-compliant with specific legal standards (Huang et al. 2025). In legal settings, erroneous answers can lead to significant economic losses and severe legal liabilities. 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 (Mozannar and Sontag 2020; Narasimhan et al. 2022; Feng, Shen, and You 2025).

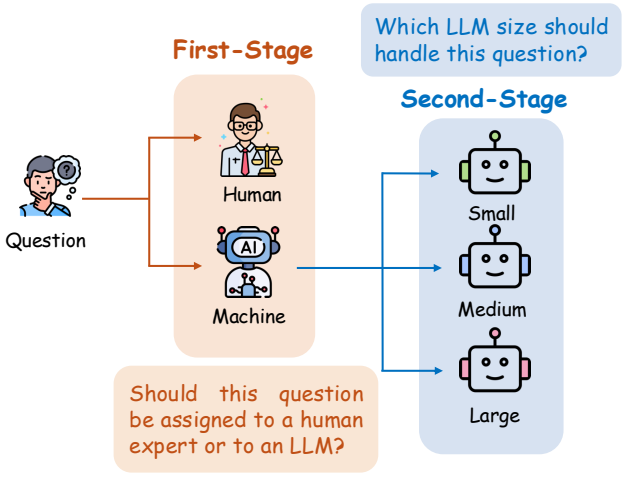

Figure 1: An illustration of our framework.

A wide range of routing approaches has been proposed to address similar challenges (Zhao, Jin, and Mao 2024; Ding et al. 2024; Yang et al. 2025). Confidence-based routers are a prevalent approach, utilizing metrics such as softmax probabilities, entropy, or verifier scores to determine whether to defer or escalate a query (Chuang et al. 2024). Alternatively, rule- or classifier-based routing methods select models based on question type, topic, or heuristic complexity (Srivatsa, Maurya, and Kochmar 2024). More recently, lightweight or meta-model routers have emerged, which predict the optimal model for a given query by leveraging historical cross-model performance (Chen, Zaharia, and Zou 2023). However, the majority of existing routers are developed for general-purpose settings and conspicuously lack legal-aware signals. Furthermore, these solutions primarily focus on selection among machine models and generally offer no principled mechanism for seamlessly deferring high-risk or out-of-scope queries to human experts, a critical requirement in practical legal workflows. While learning-to-defer frameworks explicitly optimize for machine–human assignment, prior work in this area typically assumes a single machine model interacting with a human. They do not

*Corresponding author.

Proceedings of Bridge between Artificial Intelligence and Law 2026 (Short papers), pages 219–224
Januray 21, 2026, AAAI 2026 Bridge Program, Singapore

account for routing among diverse machine models, which are essential for robust deployment in the legal sector (Hemmer et al. 2023).

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. To better align with the complex reasoning requirements of the legal domain, we integrate CoT into the routing process. This allows the system to leverage structured, legal-aware reasoning traces as pivotal features for making more informed and reliable routing decisions.

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.

Related Work

Learning to Defer

Learning to Defer addresses the when-to-answer problem in automated systems. Its core mechanism is to trade automated coverage for reduced risk by escalating uncertain or high-stakes instances to human experts.

Contemporary Learning to Defer methods can be grouped into four methodological families. End-to-end surrogate losses jointly optimize the classifier and the deferrer under a single objective, explicitly encoding the coverage–risk trade-off (Mozannar and Sontag 2020). Building on this, follow-up work tackles limited expert supervision (Hemmer et al. 2023; Charusaie et al. 2022). Uncertainty- and confidence-based approaches avoid training a separate deferrer and instead trigger deferral using internal signals, often combined with human–AI complementarity or triage frameworks (Steyvers et al. 2022). Conformal-prediction approaches provide training-free, distribution-free guarantees on error or hallucination rates at test time, yielding principled abstention thresholds (Quach et al. 2023; Yadkori et al. 2024). Additionally, several studies focus on deployment issues, including calibrating abstention policies to curb over-abstention while preserving safety (Srinivasan et al. 2024), predicting human–model disagreement to guide escalation in clinical pipelines (Sanchez et al. 2023), and using LLM-based evidence elicitation to inform deferral decisions (Strong, Men, and Noble 2025).

Despite these advances, most Learning to Defer methods assume a single automated model and employ domain-agnostic signals, such as confidence or entropy, providing limited leverage of legal-specific structure and no mechanism for choosing among multiple automated endpoints once automation is selected.

LLM Routing

LLM routing is the problem of assigning each query to an appropriate model or toolchain subject to explicit quality, cost, and latency constraints.

Cascading strategies invoke a cheaper model first and escalate only when the initial response is insufficient, yielding substantial cost savings with minimal quality degradation (Chen, Zaharia, and Zou 2023). To enable rigorous, apples-to-apples comparison across providers and cost profiles, benchmarks such as RouterBench standardize tasks, metrics, and evaluation protocols for multi-LLM selection (Hu et al. 2024). Learning-based routers leverage preference or feedback signals to switch online between weaker and stronger models based on input features and interaction history (Ong et al. 2025; Stripelis et al. 2024). Systems research explores hybrid architectures that explicitly trade privacy and latency by choosing between local small language models (SLMs) and hosted LLM services (Ding et al. 2024). A parallel direction models routing as a budgeted reinforcement-learning problem, optimizing long-horizon cost–quality trade-offs and enabling conservative escalation under uncertainty (Wei et al. 2025; Zhang et al. 2024; Li 2025; Wang et al. 2025b). Training-free routers use global or local Elo-style ranking to select models without per-task fine-tuning, offering a lightweight and scalable alternative when labeled routing data are scarce (Zhao, Jin, and Mao 2024).

Methodology

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

GRPO-based CoT Generation

Effective routing in the legal domain hinges on a nuanced understanding of a query's latent characteristics. Generic routing solutions are insufficient for law, as they fail to capture critical dimensions. A query's difficulty is not merely semantic; it is deeply tied to the structure of the legal system itself.

To address this gap, we propose using Chain-of-Thought (CoT) rationales as rich, structured features for routing. A well-formed CoT can externalize the implicit reasoning steps needed to approach a legal question. This provides a far more informative signal for downstream routing decisions than a dense vector alone.

For a given legal query q, our goal is to train a policy LLM $\pi_\theta(\mathbf{y} \mid q)$. The model generates a chain of thought (CoT):

$$\mathbf{y} = (y_1, \ldots, y_T). \tag{1}$$

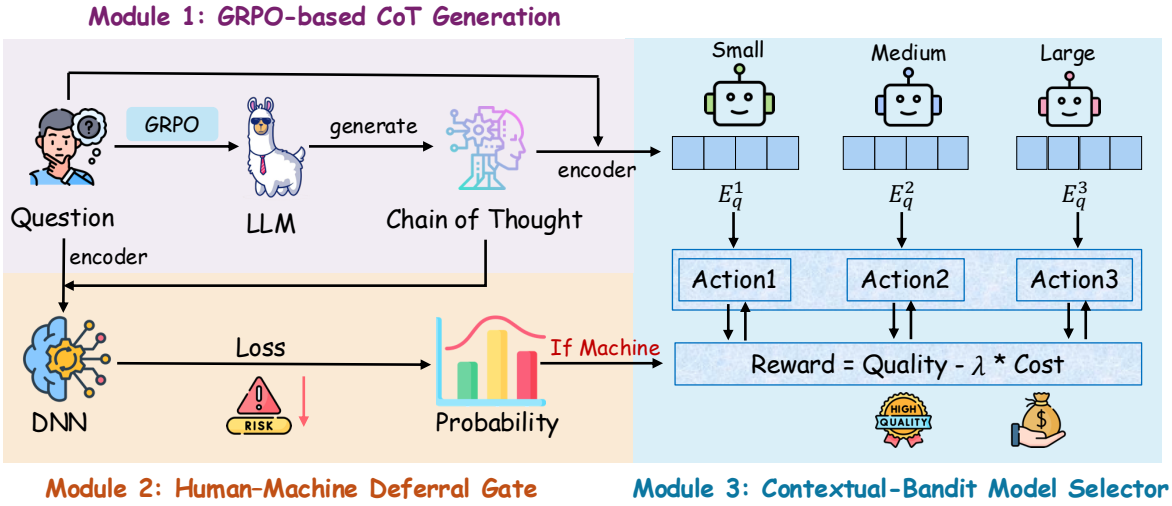

Figure 2: The overall framework.

The CoT supports high-quality legal reasoning and exposes routing-relevant signals. We fine-tune π_θ from a fixed reference policy π_{ref} using Group Relative Policy Optimization (GRPO).

Specifically, for each query q, we first sample J candidate CoTs from the current policy:

$$\mathbf{y}_j \sim \pi_\theta(\cdot \mid q), \quad j = 1, \ldots, J. \tag{2}$$

A larger LLM with a fixed rubric acts as a judge to score each $\mathbf{y}_j$:

$$s_j \in [0, 1]. \tag{3}$$

We then compute the group mean and the group-relative advantage:

$$\bar{s} = \frac{1}{J} \sum_{j=1}^{J} s_j, \tag{4}$$

$$A_j = s_j - \bar{s}, \tag{5}$$

where $s_j \in [0, 1]$ is the judge's quality score for the j-th CoT, $\bar{s}$ is the mean score over the J sampled CoTs, and A_j is the group-relative advantage. The advantage A_j measures the quality of $\mathbf{y}_j$ relative to the other CoTs in the batch.

Finally, we update θ by maximizing the GRPO objective. After training, we use the trained model to generate a CoT for each query, which then serves as auxiliary information for the two-stage router.

Human–Machine Deferral Gate

To stabilize human workload and avoid automation imbalance (over-automation or under-automation), we introduce a coverage-constrained, risk-driven router at the human–machine deferral stage. The router estimates a per-instance risk from automatic quality signals and minimizes expected risk under a target auto-answer rate ρ. This stage decides only whether to answer by machine or defer to a human.

Let (q_i, A_i) denote a legal QA instance, where q_i is the user query and A_i is the gold answer. We learn a router g_ϕ that outputs the probability of answering automatically:

$$p_i \;=\; g_\phi(\text{Enc}([q_i,\, z_i])) \in (0, 1), \tag{6}$$

where $\text{Enc}(\cdot)$ denotes the encoder, and z_i denotes CoT.

Given a user-specified target auto-coverage $\rho \in (0, 1)$, we minimize the expected risk over instances routed to the automatic channel subject to the coverage constraint:

$$\min_{\phi} \quad \frac{1}{N} \sum_{i=1}^{N} p_i R_i + \beta \Big(\frac{1}{N} \sum_{i=1}^{N} p_i - \rho \Big)^2, \qquad \beta > 0. \tag{7}$$

where $R_i \in [0, 1]$ is an automatic risk proxy.

We instantiate g_ϕ as a multilayer perceptron (MLP):

$$p_i \;=\; \sigma\big(\text{MLP}\big(\text{Enc}([q_i,\, z_i])\big)\big)\,. \tag{8}$$

Contextual-Bandit Model Selector

For queries that pass the human–machine deferral gate and are deemed suitable for automated response, the final challenge is to select the model that maximizes expected net utility, balancing answer quality against inference cost. A one-size-fits-all approach, such as always using the most powerful LLM, would be suboptimal.

To address this, we formulate the model selection task as a contextual-bandit problem. In this paradigm, each available LLM in our model pool, $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$, represents an arm that the agent can pull. Each model $\mathcal{M}_k$ is associated with a known cost C_k.

In the paper, we implement the model selector using NeuralUCB. The decision of which arm to pull is conditioned on the context of the incoming query q_i. This context vector, $\mathbf{c}(q_i)$, is constructed from the query's encoding and the CoT's encoding:

$$\mathbf{c}(q_i) = \text{Enc}([q_i, z_i]). \tag{9}$$

At each time step t, for an incoming query q_t, the NeuralUCB agent selects the arm k_t that maximizes the Upper Confidence Bound (UCB):

$$k_t = \arg \max_{k \in \{1, \ldots, K\}} \left(\hat{\mu}_k(\mathbf{c}(q_t)) + \alpha \cdot U_k(\mathbf{c}(q_t)) \right), \tag{10}$$

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.

where $\hat{\mu}_k(\mathbf{c}(q_t))$ is the expected reward for arm k as predicted by a neural network f_ψ. The network takes the context $\mathbf{c}(q_t)$ as input and outputs a reward estimate for each of the K arms; $U_k(\mathbf{c}(q_t))$ is the exploration bonus, quantifying the uncertainty of the reward estimate; and $\alpha \geq 0$ is a hyperparameter that balances the trade-off between exploitation (choosing the current best option) and exploration (trying new options).

The agent learns by observing a reward signal after each action. The reward is defined to balance answer quality against model cost:

$$r(q_t, k_t) = \text{Quality}(a_{k_t}(q_t), A_t) - \lambda C_{k_t}, \qquad (11)$$

where $a_{k_t}(q_t)$ is the answer produced by the chosen model $\mathcal{M}_{k_t}$, $\text{Quality}(\cdot,\cdot)$ is an automatic quality metric, and λ is a cost-balancing hyperparameter.

The network parameters ψ are updated online via gradient descent. After observing the true reward r_t, the model minimizes the squared error between its prediction and the observation for the chosen arm:

$$\mathcal{L} = (\hat{\mu}_{k_t}(\mathbf{c}(q_t)) - r_t)^2. \qquad (12)$$

This approach allows our framework to learn a sophisticated policy that reserves the most powerful and expensive models for queries that genuinely demand their capabilities, thereby maximizing the system's overall cost-effectiveness.

Experiments

Dataset

To accurately evaluate our proposed two-stage routing framework, we collected 11,611 marriage-related legal Q&A instances from JUSTIA[1], consisting of authentic user questions and expert attorney answers. We randomly split the data into training and test sets with an 8:2 ratio.

Baselines and Experimental Settings

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. To facilitate a thorough comparison, we consider the following baseline methods: Random Router, LLM Router, ICL-Router (Wang et al. 2025a), RouterDC (Chen et al. 2024), and EmbedLLM (Zhuang et al. 2024).

[1]https://www.justia.com/

Method	METEOR	BERTScore	Cost
Ours (full)	20.80	83.05	6.10
w/o CoT	20.13	82.86	7.85
w/o Gate	20.62	82.98	8.40

Table 2: Ablation results.

To comprehensively evaluate the performance of our framework, we assess it along two core dimensions: Answer Quality and Cost-Effectiveness. For answer quality, we employ ROUGE, METEOR, and BERTScore. We adopt the cost-accounting methodology of Li (2025). For our two-stage routing framework, we use Qwen2.5-7B-Instruct as the CoT generator. A two-layer MLP serves as the deferral gate.

Experimental Results

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. Table 2 presents the ablation results, demonstrating the necessity of each module in our framework.

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.

References

Charusaie, M.-A.; Mozannar, H.; Sontag, D.; and Samadi, S. 2022. Sample efficient learning of predictors that complement humans. In *International Conference on Machine Learning*, 2972–3005. PMLR.

Chen, L.; Zaharia, M.; and Zou, J. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.

Chen, S.; Jiang, W.; Lin, B.; Kwok, J.; and Zhang, Y. 2024. Routerdc: Query-based router by dual contrastive learning for assembling large language models. *Advances in Neural Information Processing Systems*, 37: 66305–66328.

Chuang, Y.-N.; Sarma, P. K.; Gopalan, P.; Boccio, J.; Bolouki, S.; Hu, X.; and Zhou, H. 2024. Learning to route llms with confidence tokens. *arXiv preprint arXiv:2410.13284*.

Ding, D.; Mallick, A.; Wang, C.; Sim, R.; Mukherjee, S.; Ruhle, V.; Lakshmanan, L. V.; and Awadallah, A. H. 2024. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*.

Feng, T.; Shen, Y.; and You, J. 2025. GraphRouter: A Graph-based Router for LLM Selections. In *The Thirteenth International Conference on Learning Representations*.

Hemmer, P.; Thede, L.; Vössing, M.; Jakubik, J.; and Kühl, N. 2023. Learning to defer with limited expert predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6002–6011.

Hu, Q. J.; Bieker, J.; Li, X.; Jiang, N.; Keigwin, B.; Ranganath, G.; Keutzer, K.; and Upadhyay, S. K. 2024. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031*.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Li, Y. 2025. LLM Bandit: Cost-Efficient LLM Generation via Preference-Conditioned Dynamic Routing. *arXiv preprint arXiv:2502.02743*.

Mozannar, H.; and Sontag, D. 2020. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, 7076–7087. PMLR.

Narasimhan, H.; Jitkrittum, W.; Menon, A. K.; Rawat, A.; and Kumar, S. 2022. Post-hoc estimators for learning to defer to an expert. *Advances in Neural Information Processing Systems*, 35: 29292–29304.

Ong, I.; Almahairi, A.; Wu, V.; Chiang, W.-L.; Wu, T.; Gonzalez, J. E.; Kadous, M. W.; and Stoica, I. 2025. RouteLLM: Learning to Route LLMs from Preference Data. In *The Thirteenth International Conference on Learning Representations*.

Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J. H.; Jaakkola, T. S.; and Barzilay, R. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

Sanchez, M.; Alford, K.; Krishna, V.; Huynh, T. M.; Nguyen, C. D.; Lungren, M. P.; Truong, S. Q.; and Rajpurkar, P. 2023. AI-clinician collaboration via disagreement prediction: A decision pipeline and retrospective analysis of real-world radiologist-AI interactions. *Cell Reports Medicine*, 4(10).

Srinivasan, T.; Hessel, J.; Gupta, T.; Lin, B. Y.; Choi, Y.; Thomason, J.; and Chandu, K. 2024. Selective "Selective Prediction": Reducing Unnecessary Abstention in Vision-Language Reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, 12935–12948.

Srivatsa, K.; Maurya, K. K.; and Kochmar, E. 2024. Harnessing the power of multiple minds: Lessons learned from LLM routing. *arXiv preprint arXiv:2405.00467*.

Steyvers, M.; Tejeda, H.; Kerrigan, G.; and Smyth, P. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119.

Stripelis, D.; Hu, Z.; Zhang, J.; Xu, Z.; Shah, A. D.; Jin, H.; Yao, Y.; Avestimehr, S.; and He, C. 2024. Tensoropera router: A multi-model router for efficient llm inference. *arXiv preprint arXiv:2408.12320*.

Strong, J.; Men, Q.; and Noble, J. A. 2025. Trustworthy and Practical AI for Healthcare: A Guided Deferral System with Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28413–28421.

Wang, C.; Li, H.; Zhang, Y.; Chen, L.; Chen, J.; Jian, P.; Ye, P.; Zhang, Q.; and Hu, S. 2025a. ICL-Router: In-Context Learned Model Representations for LLM Routing. *arXiv preprint arXiv:2510.09719*.

Wang, X.; Liu, Y.; Cheng, W.; Zhao, X.; Chen, Z.; Yu, W.; Fu, Y.; and Chen, H. 2025b. Mixllm: Dynamic routing in mixed large language models. *arXiv preprint arXiv:2502.18482*.

Wei, W.; Yang, T.; Chen, H.; Zhao, Y.; Dernoncourt, F.; Rossi, R. A.; and Eldardiry, H. 2025. Learning to Route LLMs from Bandit Feedback: One Policy, Many Trade-offs. *arXiv preprint arXiv:2510.07429*.

Wu, Y.; Zhou, S.; Liu, Y.; Lu, W.; Liu, X.; Zhang, Y.; Sun, C.; Wu, F.; and Kuang, K. 2023. Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12060–12075.

Yadkori, Y. A.; Kuzborskij, I.; Stutz, D.; György, A.; Fisch, A.; Doucet, A.; Beloshapka, I.; Weng, W.-H.; Yang, Y.-Y.; Szepesvári, C.; et al. 2024. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.

Yang, J.; Wu, Q.; Feng, Z.; Zhou, Z.; Guo, D.; and Chen, X. 2025. Quality-of-Service Aware LLM Routing for Edge Computing with Multiple Experts. *IEEE Transactions on Mobile Computing*, (99): 1–15.

Yao, R.; Wu, Y.; Wang, C.; Xiong, J.; Wang, F.; and Liu, X. 2025a. Elevating Legal LLM Responses: Harnessing Trainable Logical Structures and Semantic Knowledge with Legal Reasoning. In *Proceedings of the 2025 Conference of*

the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 5630–5642.

Yao, R.; Wu, Y.; Zhang, T.; Zhang, X.; Huang, Y.; Wu, Y.; Yang, J.; Sun, C.; Wang, F.; and Liu, X. 2025b. Intelligent Legal Assistant: An Interactive Clarification System for Legal Question Answering. In *Companion Proceedings of the ACM on Web Conference 2025*, 2935–2938.

Zhang, X.; Huang, Z.; Taga, E. O.; Joe-Wong, C.; Oymak, S.; and Chen, J. 2024. Efficient contextual llm cascades through budget-constrained policy learning. *Advances in Neural Information Processing Systems*, 37: 91691–91722.

Zhao, Z.; Jin, S.; and Mao, Z. M. 2024. Eagle: Efficient training-free router for multi-llm inference. *arXiv preprint arXiv:2409.15518*.

Zhuang, R.; Wu, T.; Wen, Z.; Li, A.; Jiao, J.; and Ramchandran, K. 2024. EmbedLLM: Learning compact representations of large language models. *arXiv preprint arXiv:2410.02223*.