

Uncertainty-Aware Legal Query Routing with Small Language Models

Amirreza Mohseni

Department of Advanced Computing Sciences

Faculty of Science and Engineering

Maastricht University

Maastricht, The Netherlands

i6356969

Abstract—Conversational assistants increasingly receive user queries that touch on legal issues, and reliable routing is a prerequisite for downstream legal-tech pipelines: the system must decide whether a conversation asks for legal guidance and, if so, which legal topic it belongs to. Frontier models such as GPT-5.4 perform this routing well but are expensive to call at scale.

This thesis studies whether smaller, cheaper specialists can learn legal query routing from frontier-model annotations. We construct a GPT-5.4-labeled benchmark of legal and non-legal conversations. A manually verified audit set confirms that GPT-5.4 is a strong silver labeler (93% overall, 88% on legal-guidance queries), while also showing that its errors concentrate in uncertain borderline cases. We then compare a mix of encoder classifiers and small language models, including ModernBERT and Legal-BERT classifiers and Qwen and Llama variants trained and evaluated with supervised fine-tuning and RL with verifiable rewards.

The main outcome is that small specialist models can approximate the frontier labeler for legal routing, but are most useful when paired with explicit uncertainty handling. The strongest 4B instruction-tuned model reaches 91.2% legal-guidance agreement with GPT-5.4 labels on the held-out test split, compared with 75.7% for the base model. On the Legal-Questions subset, reasoning prompts improve topic routing, with gains up to +12.0pp after RL fine-tuning. Finally, adding a Don't-Know option moves more incorrect predictions into high-uncertainty rollouts, making abstention a practical mechanism for escalation. These results suggest that legal-query routing can be performed accurately with small language models and classifiers, offering a cheaper and more sustainable path to large-scale annotation, specialized legal-model development, and the study of user-LLM behavior in law.

Index Terms—legal query routing, legal information needs, text classification, instruction-tuned language models, chain-of-thought, abstention, uncertainty, ModernBERT, reinforcement learning

I. INTRODUCTION

Conversational AI systems are now used for a broad mix of everyday and work-related tasks, including high-stakes domains [1]. In medicine, recent work has shown that users ask LLMs for concrete advice and that these conversations can be mined to study real information needs [2], [3]. Law is a similar high-stakes setting: users ask about contracts,

immigration, housing, work, family, consumer disputes, and other issues where a wrong answer or routing decision can have real-world consequences [4], [5]. This makes legal query routing an important first step for any legal-tech pipeline built on conversational data.

In this work, a *legal query router* has two jobs. First, it decides whether the user is seeking legal guidance: for example, asking how a law applies to their situation, what rights or obligations they may have, what steps to take in a dispute, or how to interpret or prepare a legal document. Second, when the conversation seeks legal guidance, it assigns the request to a primary legal topic. This differs from classifying formal legal documents: the input is an informal, multi-turn user–assistant conversation, and the legally relevant signal may appear only after context-setting, clarification, or follow-up messages. A good router should therefore be accurate, cheap enough to run at scale, and robust on unseen conversations. If a specialist can closely reproduce GPT-5.4 annotations on held-out data, it may offer a more cost-effective and sustainable way to label and analyze large conversational datasets. In addition, if the router can express uncertainty when it is likely to be wrong, those uncertainty levels can indicate when model outputs should be treated more cautiously rather than trusting every label equally.

One way to build such a dataset is to use a strong frontier model as a labeler. Frontier models can follow detailed labeling instructions and produce structured outputs, but calling them at scale is expensive, slower than running a small model, and dependent on an external provider. The central question of this thesis is therefore whether a smaller specialist can approximate the behavior of a frontier labeler on legal-guidance detection and topic routing, while also exposing useful uncertainty signals.

We approach this problem in three steps. First, we construct a legal-classification dataset by labeling conversations from WildChat-1M [6] with GPT-5.4 and rebalancing the sample to contain both legal and non-legal conversations. A manually verified audit set (Table II, in Section III-C) shows that GPT-5.4 reaches 93.0% overall accuracy on this sample: 88.0% on legal-guidance queries and 98.0% on non-legal queries. However, its errors concentrate on medium-uncertainty examples, where accuracy drops to 85.5%, compared with 96.4% on low-uncertainty examples. This motivates two design choices:

This thesis was prepared in partial fulfillment of the requirements for the Degree of Bachelor of Computer Science, Maastricht University. Supervisor: Jerry Spanakis. Second examiner: Aki Härmä.

using GPT-5.4 as a silver labeler and training a smaller specialist model as the production router.

Second, we train specialist models under three paradigms on the resulting data. *Encoder training* fine-tunes ModernBERT-base and Legal-BERT with hyperparameter sweeps over input representation and batch settings. SFT trains Qwen3.5-4B to predict the GPT-5.4 labels for each task input. RLVR fine-tunes the same Qwen model using task-level rewards over structured rollouts, allowing us to study reasoning as a training signal. We additionally use Llama-3.2-3B-Instruct for a reasoning ablation on the `Legal-Questions` split. Because it is instruction-tuned rather than RLVR-trained, this model helps separate prompting-based reasoning effects from reasoning effects introduced through training.

Third, we evaluate whether reasoning and abstention improve this routing task. Prior work shows that chain-of-thought (CoT) prompting and reasoning-oriented training can improve LMs on tasks that benefit from intermediate reasoning [7]–[9]. Here, we test whether the same idea helps legal classification by comparing a NO THINK mode, where the model answers with labels only, against a THINK ON mode, where it may reason before producing those labels. We also test whether an explicit Don’t-Know option and an uncertainty tag can move more of the model’s wrong answers into high-uncertainty cases, which would make uncertainty useful for selective routing.

Problem statement. Frontier models can label legal routing tasks accurately, but calling them at scale is expensive and slow. This thesis asks whether smaller specialist models can make legal query routing over informal conversations accurate, efficient, and uncertainty-aware without relying on a frontier model at inference time.

This leads to four research questions:

- RQ1** Can a small specialist model closely approximate GPT-5.4 annotations for legal-guidance detection and topic routing under realistic compute constraints?
- RQ2** Does enabling LMs to reason improve legal classification, both as a prompting strategy and as an RL training signal?
- RQ3** Does giving the model an explicit uncertainty option let it route more of its wrong answers into high-uncertainty rollouts, mirroring the calibration pattern observed in GPT-5.4?
- RQ4** Does routing performance improve when classifiers see the full user–assistant conversation rather than user turns alone?

Our main contributions are: (1) the `V2-Balanced` legal-classification dataset based on `WildChat-1M`, the `Legal-Questions` companion dataset, and a visualization tool for viewing conversations with added metadata (detailed in Appendix K); the released dataset and model resources are listed in Appendix A; (2) a controlled comparison of encoder classifiers (ModernBERT and Legal-BERT), label SFT, and RL-tuned language models evaluated against GPT-5.4 silver reference labels on the same routing task; (3) evidence that explicit reasoning helps both prompted and RL-tuned small LMs in legal classification; and (4) evidence that incentivizing

uncertainty can make likely misclassifications easier to identify and support selective routing.

II. RELATED WORK

A. Conversational datasets and specialist models

Large language models such as GPT-3 and PaLM helped establish open-ended conversational interaction with models [10], [11]. This makes real user–assistant chat logs useful for studying how deployed assistants are used. `WildChat-1M` is central for this thesis because it contains real ChatGPT conversations across many topics [6], and broader usage studies show that people increasingly use conversational AI for everyday, work-related, and high-stakes tasks [1].

Medical-advice-seeking work shows a pattern that this thesis follows for law: isolate a high-stakes subset from general chat data, characterize the user needs, and build a task-specific dataset [2], [3]. This also connects to work showing that smaller models can become capable when trained on carefully selected or high-quality data rather than only larger, noisier corpora [12].

B. Legal NLP and legal information needs

Legal NLP has often focused on formal legal texts such as statutes, court cases, and contracts, with downstream tasks including retrieval, classification, and prediction [13]. Access-to-justice research instead emphasizes the everyday legal problems people face before they become formal cases [14]. This thesis sits between these perspectives by studying informal user–assistant conversations and identifying whether they contain legal-guidance-seeking behavior.

Question-oriented legal datasets such as BSARD map citizen questions to relevant statutory provisions and report topic structures that overlap with everyday legal needs [5]. LLM-based legal topic classification provides a complementary direction: Sargeant et al. [15] evaluate large language models for assigning legal texts to a bespoke topic taxonomy and report strong expert-evaluated performance. Our setting moves from formal legal texts and isolated citizen questions to noisy multi-turn conversations, requiring a two-stage routing setup: first detecting legal guidance, then assigning a primary legal topic.

C. Long-context encoders

BERT and Legal-BERT are strong baselines for fine-tuned text classification [16], [17], but many chatbot conversations exceed 512 tokens. ModernBERT keeps the efficient encoder-classifier setup while extending context length to 8192 tokens with a more recent architecture and training recipe [18], making it a natural baseline for long multi-turn conversations.

D. Chain-of-thought and reasoning

LLMs can improve on reasoning-heavy tasks when prompted to reason before answering, including zero-shot and few-shot chain-of-thought prompting [7], [8]. Recent work extends this idea into training through reinforcement learning

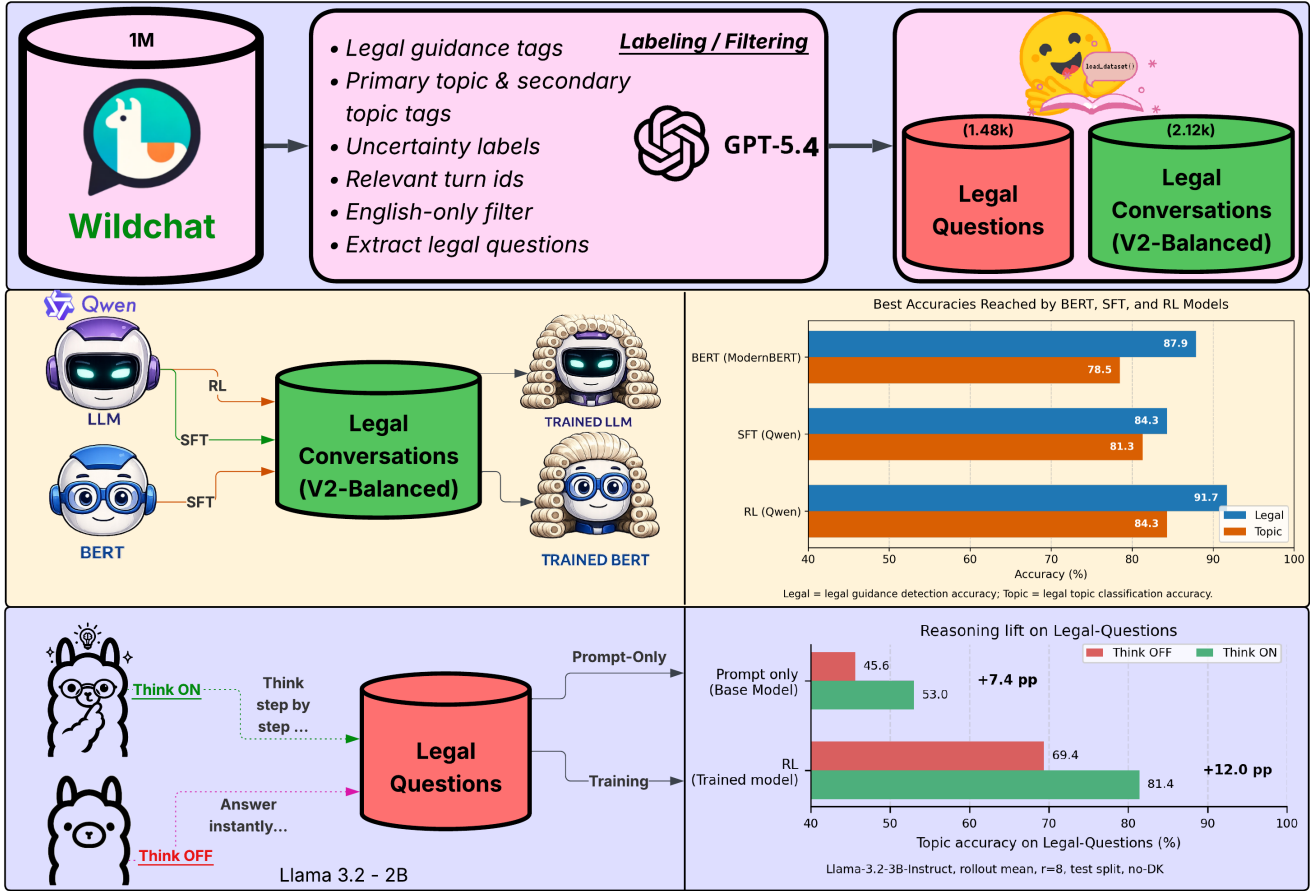


Fig. 1. Overview of the thesis pipeline. GPT-5.4 labels WildChat-1M conversations into legal routing datasets, after which encoder classifiers, label SFT, and RL-tuned small LMs are compared on legal-guidance detection, topic routing, reasoning, and abstention behavior.

with verifiable rewards, as in Tulu 3, DeepSeek-R1-style systems, and DeepSeekMath’s GRPO approach [9], [19], [20].

Legal-domain reasoning work points in a similar direction but on different tasks: UniLaw-R1 applies R1-style SFT and RL to legal reasoning benchmarks, while Legal Δ uses a chain-of-thought-guided information-gain objective to improve legal reasoning [21], [22]. At the same time, *Reasoning’s Razor* shows that RLVR-style classification training can make models more overconfident even when aggregate accuracy improves, creating a potential limitation for routing settings where confident mistakes are costly [23]. This motivates evaluating reasoning empirically for legal routing and testing whether adding a Don’t-Know option during training can mitigate overconfident errors.

E. LLM fine-tuning

Fine-tuning turns a general model into a specialist. SFT and RLHF-style methods improve instruction following [24], while LoRA reduces adaptation cost by training low-rank adapters instead of all parameters [25]. This is practical for RL with verifiable rewards, where outcome rewards provide

lower-bandwidth learning signals and low-rank updates may be sufficient [26].

For legal routing, SFT can imitate input-label demonstrations, but it remains off-policy with respect to the model’s own sampled outputs. Prior work links SFT to memorization and out-of-distribution (OOD) forgetting, while RL with outcome-based rewards can improve generalization or restore some OOD performance [27], [28]. This motivates comparing RLVR against SFT for legal routing, with the goal of assessing whether outcome-based optimization can improve generalization on held-out data.

F. Abstention and uncertainty

A core issue in LLM deployment is that models often answer even when uncertain. Kalai et al. argue that hallucinations are partly sustained by training and evaluation procedures that reward guessing over acknowledging uncertainty [29].

TruthRL addresses this problem with a ternary reward: correct answers are rewarded, wrong answers are penalized, and abstention receives an intermediate outcome instead of being treated as ordinary failure [30]. Our Don’t-Know setup is inspired by this idea, but our task is more constrained because

TABLE I
SUMMARY OF THE 14 LEGAL TOPIC CATEGORIES USED FOR
PRIMARY-TOPIC ROUTING. FULL CATEGORY DEFINITIONS ARE IN
TABLE XII (APPENDIX C).

Topic	Topic
Family	Housing
Work	Public Benefits
Criminal Justice	Consumer Debt
Contracts	Immigration
Business	Data Privacy
Intellectual Property	Civil Rights
International / Cross-Border	Other

legal-guidance and topic labels can be evaluated by exact matching. We also measure whether self-reported uncertainty predicts errors. LLM-as-a-judge work such as MT-Bench and Chatbot Arena is relevant for open-ended evaluation [31], but exact matching is preferable for this discrete-label routing task.

III. DATASET

This section describes the `V2-Balanced` legal-guidance routing dataset. We first summarize how it was derived from `WildChat-1M` [6], then define the routing task and label format, and finally audit the GPT-5.4 teacher annotations used as silver reference labels.

A. Dataset Construction

We constructed the dataset in three steps:

- 1) **Label conversations.** We labeled conversations from the public `WildChat-1M` corpus with GPT-5.4 using a structured prompt that elicits the legal-guidance flag, one primary topic, a list of all legal topics that appear in the conversation, and the model’s own uncertainty (LOW, MEDIUM, or HIGH). We include both the primary topic and the all-topics list because some user requests span multiple legal domains; the primary topic provides a single routing target, while the list preserves secondary topics for analysis and auditing.
- 2) **Balance the dataset.** We retained all 1,061 conversations that GPT-5.4 labeled as legal guidance and sampled an equal number of non-guidance conversations uniformly at random from the English non-guidance pool, producing the 2,122-example `V2-Balanced` dataset.
- 3) **Create held-out splits.** We split the balanced dataset with a 90/5/5 train/validation/test ratio, yielding 1,909 train, 106 validation, and 107 test examples.

The specialist models are trained and evaluated against these GPT-5.4 annotations as silver reference labels; the manual annotations described below are used to audit the teacher labels, not as validation or test labels for the trained models. Complete prompt and release details are provided in the appendix; the released `V2-Balanced` classification dataset and `Legal-Questions` companion dataset are listed in Appendix A.

TABLE II
GPT-5.4 ACCURACY ON THE MANUALLY VERIFIED AUDIT SET,
STRATIFIED BY LEGAL-GUIDANCE CLASS AND BY GPT-5.4’S OWN
UNCERTAINTY TAG.

Evaluation group	Correct	Incorrect	Accuracy
Overall	186	14	93.0 %
Legal-guidance queries	88	12	88.0 %
Non-legal-guidance cases	98	2	98.0 %
Low-uncertainty cases	133	5	96.4 %
Medium-uncertainty cases	53	9	85.5 %

B. Task Definition

Each example is a multi-turn user–assistant conversation. The router must produce three structured fields: (i) a legal-guidance flag indicating whether the user is seeking legal advice or guidance from the model, (ii) a primary topic from the 14-category legal topic taxonomy summarized in Table I (full definitions in Table XII, Appendix C), and (iii) an uncertainty tag in {LOW, MEDIUM, HIGH}. In the silver reference label format, non-legal examples are represented as NO/N/A, while legal-guidance examples are represented as YES plus a primary topic. The no-DK model variants must choose between YES and NO; the +DK variants may instead output Don’t-Know. We use the term Don’t-Know in training and evaluation terminology throughout the paper.

C. GPT-5.4 Manual Audit

To validate the labeling pipeline, we randomly sampled a balanced set of legal and non-legal conversations and re-labeled them manually to measure the performance of GPT-5.4 on the task. This manually annotated audit estimates the quality of the teacher annotations before they are used as silver reference labels for training, validation, and test splits.

Table II reports accuracy stratified by ground-truth class and by GPT-5.4’s own uncertainty tag. There is no high-uncertainty row because GPT-5.4 almost never emitted HIGH for this task, which is plausible for a large model on a coarse routing decision. Three observations motivate the rest of the thesis:

- 1) GPT-5.4 is a strong teacher model: 93.0% overall, 88.0% on legal-guidance, 98.0% on non-legal. This supports using its annotations as silver reference labels for the downstream training and evaluation splits.
- 2) GPT-5.4 detects non-legal cases more confidently, which suggests that frontier labelers can help reduce large conversational datasets to smaller task-relevant subsets.
- 3) Errors are not uniformly distributed. Accuracy is 96.4% on cases GPT-5.4 marks as low uncertainty and only 85.5% on medium-uncertainty cases. *Uncertainty correlates with error*, which is exactly the property we want our smaller specialist to inherit and the focus of RQ3.

IV. METHODS

Having described the dataset in Section III, we now describe the models, training paradigms, and evaluation protocol used to train specialist routers from GPT-5.4 silver reference labels.

A. Models

We compare three training paradigms on the same silver reference labels: encoder classifiers, supervised fine-tuning (SFT), and reinforcement learning (RL). Within the generative family, we additionally study reasoning prompts on Llama-3.2-3B-Instruct.

Encoder classifiers (BERT training). We fine-tune both Legal-BERT [17] and ModernBERT-base [18] as encoder baselines. Legal-BERT is trained and evaluated with a 512-token sequence length, truncating any remaining turns. ModernBERT-base supports longer inputs; we train the reported checkpoints at 4096 tokens for cost and speed, then evaluate them at both 4096 and 8192 tokens. For each encoder family, we train separate classifiers for legal-guidance detection and primary-topic classification. We run 24-trial hyperparameter sweeps over learning rate, batch size, and input mode (user turns only vs. full user–assistant conversation), select the binary legal-guidance checkpoint by validation weighted F1, and select the topic-classification checkpoint by validation accuracy. The deployed checkpoints use the best input format per task (Appendix F).

Label SFT. As a generative baseline that imitates the teacher directly, we fine-tune Qwen3.5-4B with LoRA ($r = 64$) on silver reference `<legal_guidance>` and `<primary_topic>` tags using the NO THINK (fast) prompt style. Training uses the same V2-Balanced split as the RL runs. This baseline uses the same stable tag schema and dataset as RL, but optimizes cross-entropy on fixed labels rather than verifiable rollout rewards.

Instruction-tuned LMs. We train Llama-3.2-3B-Instruct on `Legal-Questions` to measure how explicit reasoning affects model performance for an instruction-tuned model. Because this model was not explicitly trained to reason, changes from NO THINK prompting to THINK ON prompting provide a clearer signal of how much reasoning helps on this topic-routing task.

Reasoning LMs (RL training). Separately from the Llama instruction-tuned study, we use Qwen3.5-4B for the main reasoning-LM experiments because it is a stronger baseline model with reasoning behavior already present in its training. For Qwen3.5-4B, we train a no-DK variant that is not explicitly instructed, rewarded, or penalized for abstaining, and a +DK variant that allows the model to abstain by outputting Don’t-Know (Section IV-C). RL training uses a verifiable rubric reward over the structured output.

Reference. GPT-5.4 provides the silver reference labels used for training and for the validation/test splits. Its manually audited accuracy in Section III-C is therefore a teacher-quality check rather than a direct benchmark that trained models are expected to exceed.

B. Prompting: No Think vs. Think On

We compare two prompt styles. In NO THINK mode, the model is asked to answer immediately with the final XML-style labels. In THINK ON mode, the model may first write an explicit reasoning block before producing the same final labels. Thus the experimental contrast is not a different label space, but whether the model is allowed to reason before committing to the labels. In the reported generative evaluations, the prompt also asks the model to self-verbalize its confidence as LOW, MEDIUM, or HIGH before the final task labels. When a model does not finish its reasoning and reach a conclusion within the token budget, we mark the output as HIGH uncertainty because it did not produce a confident answer in time.

C. Abstention and Don’t-Know

When abstention is enabled, the legal-guidance label may take the value Don’t-Know. Under this configuration the wrong-answer penalty is set to $w = -3$, rather than the TruthRL-inspired pilot value $w = -1$, so a confidently wrong answer is strictly worse than abstaining.

D. Reward Design

The RL objective follows the same binary-versus-ternary distinction used in TruthRL [30]. Let $y_L \in \{\text{YES}, \text{NO}\}$ denote the silver reference legal-guidance label and let y_T denote the silver reference primary topic, with $y_T = \text{N/A}$ when $y_L = \text{NO}$. The model outputs \hat{y}_L and \hat{y}_T , and receives an additive reward

$$R = R_{\text{legal}} + R_{\text{topic}}.$$

The uncertainty tag is recorded for calibration analysis, but it is not itself part of the accuracy reward. The topic component rewards a topic decision only when the legal-guidance decision is also correct. Therefore the topic component directly captures whether both required routing decisions are correct.

No-DK binary reward. The model must commit to both fields:

$$R_{\text{legal}}^{\text{bin}} = \begin{cases} +1, & \text{if } \hat{y}_L = y_L, \\ 0, & \text{otherwise,} \end{cases}$$

$$R_{\text{topic}}^{\text{bin}} = \begin{cases} +1, & \text{if } \hat{y}_L = y_L \text{ and } \hat{y}_T = y_T, \\ 0, & \text{otherwise.} \end{cases}$$

+DK ternary reward. The same topic scoring is used, but Don’t-Know receives zero reward:

$$R_{\text{legal}}^{\text{DK}} = \begin{cases} +1, & \text{if the committed legal label is correct,} \\ 0, & \text{if the legal label is Don’t-Know,} \\ w, & \text{if the legal label is wrong or invalid,} \end{cases}$$

$$R_{\text{topic}}^{\text{DK}} = \begin{cases} +1, & \text{if } \hat{y}_L = y_L \text{ and the committed topic label is correct,} \\ 0, & \text{if the topic label is Don’t-Know,} \\ w, & \text{otherwise.} \end{cases}$$

Final runs use $w = -3$. Pilot runs followed TruthRL by setting $w = -1$, but after experiments we increased the penalty so incorrect committed answers are further below abstentions.

The point of the ternary design is to separate abstention from hallucination. For either reward component, if y_0 is correct, y_1 abstains, and y_2 commits to a wrong label, then

$$R_*(x, y_0) = 1 > R_*(x, y_1) = 0 > w = R_*(x, y_2).$$

Under a group-relative update, this gives abstention a higher relative advantage than a hallucinated commitment, without making abstention as good as a correct answer.

We also changed how unfinished completions are handled after pilot runs. The first version gave completions a reward of 0 when they ran out of tokens before producing final tags. This made non-completion resemble Don’t-Know behavior and caused some policies to stop finishing outputs as training progressed. The final version assigns unfinished completions the same $w = -3$ penalty as malformed or wrong outputs. This removed the accidental abstention loophole and made the reward closer to deployment, but it also increased forced-commit hallucination errors relative to the pilot because the model could no longer avoid difficult examples by not finishing. Unlike TruthRL, we did not observe broad over-abstention in the final +DK runs.

E. Evaluation Protocol

For generative evaluations, each example is sampled with $r = 8$ rollouts for both the main legal-classification task and the Legal-Questions companion split. Metrics are computed at the *rollout level* rather than by majority vote: each rollout is scored independently and accuracies are means over all rollouts. Reported model results use the held-out **test** split; GPT-5.4 annotations are the silver reference labels, while the manually annotated audit set is used only to estimate teacher quality. We report reward-free legal-guidance accuracy and end-to-end primary-topic accuracy so that runs with different reward configurations remain directly comparable. Malformed or unparsed outputs count as wrong in these reward-free metrics; they are not credited as Don’t-Know. The topic metric captures both routing decisions: for non-legal silver-reference examples, it requires NO/N/A; for legal silver-reference examples, it requires YES and the correct primary topic.

We additionally evaluate on Legal-Questions, where every row seeks legal guidance, so only topic accuracy is reported. Qwen checkpoints are trained on V2-Balanced only, while Legal-Questions test evaluations use the same reasoning prompt and rollout protocol. The Llama-3.2-3B-Instruct reasoning-lift study is trained and evaluated on the Legal-Questions test split with the prompt style used during RL: fast-trained checkpoints are evaluated with NO THINK, and reasoning-trained checkpoints are evaluated with THINK ON. A *cross-eval* runs the no-DK-trained checkpoint with `allow_dont_know=true` (+DK eval config) so it can be

TABLE III

MAIN RESULTS ON V2-BALANCED TEST (ROLLOUT MEAN, $r = 8$ FOR LMS). LEGAL AND TOPIC REPORT PERCENTAGE ACCURACY AGAINST GPT-5.4 SILVER REFERENCE LABELS; ABSTAIN IS THE PERCENTAGE OF ROLLOUTS THAT OUTPUT DON’T-KNOW. TOPIC REQUIRES THE CORRECT LEGAL-GUIDANCE DECISION AND THE CORRECT PRIMARY-TOPIC DECISION. ENCODER ROWS USE THE VALIDATION-SELECTED CLASSIFIERS DESCRIBED IN SECTION IV-A. THE RL *no-DK* (*cross +DK*) ROW IS A DIAGNOSTIC RE-EVALUATION OF THE NO-DK CHECKPOINT UNDER THE +DK PROMPT/CONFIG.

Model	Legal	Topic	Abstain
ModernBERT classifiers	87.9	78.5	N/A
Legal-BERT classifiers	84.1	72.0	N/A
Label SFT	84.3	81.3	0.0
Qwen3.5-4B base	75.7	71.1	0.0
Qwen3.5-4B RL no-DK	91.2	83.4	0.0
Qwen3.5-4B base +DK	74.5	68.1	0.0
Qwen3.5-4B RL +DK	87.0	80.3	1.1
Qwen3.5-4B RL no-DK (<i>cross +DK</i>)	91.7	84.3	0.0

compared directly against the +DK-trained checkpoint under identical evaluation settings.

Training runs. Run-specific training hyperparameters, including Prime configuration files, batch sizes, rollout counts, generation budgets, and LoRA settings, are reported in Appendix G.

V. EXPERIMENTS

The main experimental matrix on V2-Balanced crosses encoder sweeps (ModernBERT and Legal-BERT; user-only vs. full-conversation input), a label SFT baseline on Qwen3.5-4B, and generative RL. For Qwen RL, we cross $\{\text{BASE,RL}\} \times \{\text{NO-DK,+DK}\}$, giving four runs per split. Headline +DK cross-evaluation and Legal-Questions transfer results are reported in Section VI, with complete uncertainty breakdowns in the appendix. The ModernBERT and Legal-BERT classifiers are fine-tuned on the same training split and reported under their canonical inference setup. The reasoning-lift study uses Llama-3.2-3B-Instruct on the Legal-Questions split and crosses $\{\text{BASE,RL}\} \times \{\text{NO THINK,THINK ON}\}$ with matched prompt styles.

VI. RESULTS

A. RQ1: Can a small specialist approximate GPT-5.4 labels?

Table III reports the main comparison against GPT-5.4 silver reference labels on the test split. For legal-guidance detection, Qwen3.5-4B RL reaches 91.2% agreement with GPT-5.4 labels, while the base model reaches 75.7%. The same pattern holds on the stricter Topic metric, which moves from 71.1% to 83.4%. Because Topic already requires the correct legal-guidance decision and the correct primary-topic decision, it is the end-to-end routing metric reported here. These numbers should be read as agreement with the teacher annotations, not as performance above GPT-5.4 on the manually annotated audit set.

TABLE IV

ENCODER PIPELINE COMPARISON ON V2-BALANCED TEST. MODERNBERT IS TRAINED AT 4096 TOKENS AND EVALUATED AT BOTH 4096- AND 8192-TOKEN CONTEXT LENGTHS; LEGAL-BERT TRUNCATES AT 512 TOKENS. ANY-TOPIC COUNTS A TOPIC PREDICTION AS CORRECT IF IT APPEARS ANYWHERE IN THE GPT-5.4 TOPIC LIST.

Encoder	Max len	Legal	Topic	Any-topic
ModernBERT-base	4096	87.9	78.5	80.4
ModernBERT-base	8192	87.9	78.5	81.3
Legal-BERT	512	84.1	72.0	74.8

1) *Encoder family comparison*: Table IV compares the two encoder families (Fig. 3 in Appendix F). ModernBERT leads Legal-BERT on both heads when evaluated at the training length of 4096 tokens: +3.8pp on legal-guidance detection and +6.5pp on end-to-end topic routing (78.5% vs. 72.0%). Evaluating the same ModernBERT checkpoints at 8192 tokens leaves headline legal and topic accuracy unchanged and only slightly improves Any-topic accuracy (80.4% \rightarrow 81.3%).

2) *Training paradigm comparison*: Table III also summarizes the three training paradigms (Fig. 4 in Appendix F). RL is strongest among the canonical trained configurations on both heads. Label SFT closes most of the remaining topic gap relative to BERT (+2.8pp topic vs. ModernBERT) while matching Legal-BERT on legal (+0.4pp).

3) *RQ4: User vs. full conversation input*: The best input mode is encoder-dependent, so we select it from each encoder’s validation sweep rather than from the test split. Legal-BERT uses user turns, while ModernBERT uses the full conversation. The held-out test comparison confirms each validation choice: the sweep-selected input mode wins *both* heads for each encoder (Legal-BERT user-only 84.1/78.6 vs. 75.7/75.2 legal/topic F1; ModernBERT full 87.6/83.2 vs. 82.3/60.9). The per-input-mode sweep and test tables are in Appendix F (Tables XVI and XVII).

Under identical +DK evaluation settings, the no-DK-trained RL checkpoint evaluated in cross mode outperforms the +DK-trained checkpoint on V2-Balanced test (Table III): 91.7/84.3 versus 87.0/80.3 on legal/topic accuracy. This cross row is a separate stochastic evaluation with a different +DK prompt, so the small gain over the canonical no-DK row should not be over-interpreted; its main use is comparing the same no-DK-trained policy against the +DK-trained policy under matched +DK inference.

On the out-of-domain Legal-Questions test split ($n = 148$ legal-guidance rows), the same no-DK RL checkpoint reaches 84.5% topic accuracy (Table XVIII; Fig. 7), compared with 49.1% for the base model. The +DK-trained checkpoint trails both under matched +DK evaluation (71.3% topic), while cross-evaluating the no-DK checkpoint with Don’t-Know enabled still reaches 82.8%. Complete per-configuration uncertainty breakdowns and extended Legal-Questions tables are in Appendix H.

TABLE V

REASONING LIFT: TOPIC ACCURACY ON LEGAL-QUESTIONS TEST (LLAMA-3.2-3B-INSTRUCT, ROLLOUT MEAN, $r = 8$). NO THINK ASKS FOR LABELS ONLY; THINK ON ALLOWS A REASONING BLOCK BEFORE THE LABELS.

Model	No Think	Think On	Lift
Llama-3.2-3B base	45.6	53.0	+7.4
Llama-3.2-3B RL	69.4	81.4	+12.0

TABLE VI

UNCERTAINTY DISTRIBUTION ON *wrong* ROLLOUTS (LEGAL-GUIDANCE OR TOPIC PREDICTION INCORRECT) ON V2-BALANCED TEST (ROLLOUT MEAN, $r = 8$; MISSING TAGS MAPPED TO HIGH). n IS THE NUMBER OF WRONG ROLLOUTS.

Model	LOW	MED	HIGH	n
Qwen3.5-4B base	54.3	7.7	38.1	247
Qwen3.5-4B base +DK	56.0	1.1	42.9	273
Qwen3.5-4B RL	99.3	0.7	0.0	142
Qwen3.5-4B RL +DK	76.9	0.0	23.1	169

B. RQ2: Does reasoning help?

Table V reports the reasoning-lift study on the Legal-Questions test split. The base Llama-3.2-3B-Instruct reaches 45.6% topic accuracy in NO THINK mode and 53.0% in THINK ON mode (+7.4pp). RL fine-tuning on Legal-Questions raises NO THINK accuracy to 69.4%, and THINK ON adds a further +12.0pp to 81.4%. The full Any-topic breakdown and plot are in Appendix I.

C. RQ3: Does Don’t-Know flag wrong answers as uncertain?

Table VI supports the RQ3 analysis, with the corresponding legal-head plots moved to Appendix J. The base Qwen3.5-4B already has a non-trivial uncertainty signal, and enabling Don’t-Know during evaluation increases the share of wrong outputs tagged HIGH while slightly reducing Topic accuracy (71.1 \rightarrow 68.1).

The RL-fine-tuned model presents a starker version of the same picture. Without Don’t-Know, the policy collapses uncertainty to LOW on essentially every rollout, even when wrong. Retraining with Don’t-Know enabled restores a meaningful uncertainty signal: over all wrong rollouts, HIGH uncertainty rises from 0.0% to 23.1%, with Topic accuracy moving from 83.4% to 80.3%.

Figure 2 illustrates the reward-design trade-off directly. Moving the wrong-answer penalty from $w = -1$ toward $w = -3$ increases the model’s use of Don’t-Know.

D. Efficiency

Table VII reports list pricing for the hosted LM calls used in the pipeline, plus an approximate self-hosted input-token estimate for ModernBERT-base. At these rates, Qwen3.5-4B is about $25\times$ cheaper than GPT-5.4 on input tokens and $50\times$ cheaper on output tokens. The ModernBERT output cost is

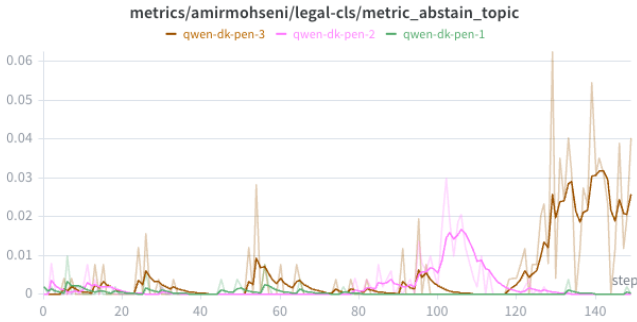


Fig. 2. Topic abstention rate for Qwen3.5-4B +DK runs with wrong-answer penalties $w = -1$, $w = -2$, and $w = -3$. Increasing the penalty makes abstention more attractive relative to an incorrect committed answer, with the strongest penalty producing the largest sustained rise in Don’t-Know outputs.

TABLE VII
INFERENCE PRICING (USD PER 1M TOKENS, MAY 2026). PRIME
INFERENCE LIST PRICING IS SHOWN FOR HOSTED LMS; MODERNBERT
USES AN APPROXIMATE SELF-HOSTED INPUT-TOKEN ESTIMATE.

Model	Input \$/1M	Output \$/1M
GPT-5.4 (reference)	2.50	15.00
Qwen3.5-4B	0.10	0.30
Llama-3.2-3B-Instruct	0.10	0.34
ModernBERT-base classifiers	0.001 (est.)	≈ 0

effectively zero because the encoder classifier emits fixed labels rather than generated text.

VII. DISCUSSION

ModernBERT vs. Legal-BERT. The encoder comparison suggests that context length matters more than legal-domain pretraining for this routing task. Legal-BERT remains a useful legal-domain baseline, but its 512-token limit can remove later turns that carry user intent. ModernBERT’s 4096-token training setup already captures much more of the conversation, and evaluating at 8192 tokens gives only a small Any-topic gain on this split.

BERT vs. label SFT vs. RL. The three training paradigms occupy different points in the trade-off between cost, flexibility, and accuracy. Encoders are best for high-volume, fixed routing. Label SFT is most attractive when the model is mainly being used for this task-specific routing problem and must follow a stable tag schema, consistent with findings that SFT can memorize teacher demonstrations [27]. RL is preferable when the base model can already obtain non-zero reward and the additional rollout cost is acceptable; in this setting, on-policy learning from structured rollout feedback appears to generalize better than fixed-label imitation [28].

Input mode is model-dependent. The input-mode results should not be read as a universal preference for user-only or full-conversation input. For Legal-BERT, user-only input likely helps because it compresses user intent into a 512-token

window. Serializing the full conversation can drop later user turns and add assistant text that dilutes the signal.

ModernBERT shows the opposite pattern: its longer context allows the classifier to use assistant clarification turns and surrounding context. The practical lesson is to choose the serialization jointly with the encoder context length rather than treating input mode as an independent preprocessing choice.

When is ModernBERT enough? The practical choice between ModernBERT and a small generative LM depends less on raw accuracy alone than on how fixed the routing task is. If the task has a stable label set, a long-context encoder such as ModernBERT is attractive: it is smaller, cheaper to run, scalable, and simple to fine-tune for supervised classification, with a well-understood training recipe. This makes ModernBERT a strong choice for narrow, repeated decisions such as legal-guidance detection or topic routing.

The trade-off is flexibility and context. ModernBERT can be retrained or adapted for another supervised task, but it cannot simply be prompted at inference time to add a new tag or follow a changed output schema. Qwen3.5-4B also has a much larger 256k-token context window than ModernBERT, although our experiments use a 4096-token generation budget. LLMs are more expensive, but the same model can often be reused across related routing, explanation, abstention, and instruction-following tasks by changing the prompt or fine-tuning objective; it may also learn new tasks through in-context learning. Therefore, ModernBERT is preferable when the task is fixed and cost matters most, whereas the LM route is preferable when the same model should be easily repurposed for nearby tasks or when much longer context is required.

Reasoning helps, but not always in every metric. The reasoning-lift study shows that explicit reasoning before committing to labels improves topic routing for both the base and RL-trained 3B models. This is consistent with the broader CoT literature and with recent legal-reasoning RL studies such as UniLaw-R1 and Legal Δ [21], [22].

The caveat is that reasoning is not automatically better on every axis. Reasoning’s Razor shows that reasoning can raise aggregate accuracy while making models more overconfident and eroding recall at important safety thresholds [23]. Their task differs from ours, but we observe a related pattern: the strongest no-DK Qwen3.5-4B RL policy is highly accurate while collapsing its uncertainty tag to LOW on essentially every rollout. For that reason, we interpret the reasoning lift together with RQ3’s uncertainty and Don’t-Know analysis. Our setting is narrower than legal-reasoning systems, because the model is not asked to solve legal problems; it is asked to route conversations into a verifiable legal-guidance taxonomy.

Role of the ternary reward. The RQ3 results indicate that an explicit Don’t-Know option, combined with a required uncertainty tag and a stronger wrong-answer penalty, gives the model a cleaner mechanism for selective routing. The value of the ternary reward is not that it eliminates errors, but that it makes some errors easier to identify and set aside for additional review.

This directly counteracts the reasoning-induced overconfidence noted above. Reasoning and abstention are not in opposition: reasoning supplies the accuracy gains, while a ternary reward with an explicit uncertainty requirement restores some of the calibration that reasoning erodes. A concrete next step is to tune the reward so the no-DK accuracy is preserved while the +DK calibration benefit is retained.

Limitations. First, the main benchmark uses GPT-5.4 annotations as silver reference labels rather than ground truth. The manually verified audit set gives evidence that these labels are useful, but it does not remove teacher noise or prove legal correctness. Second, the task is treated as single-label classification even though legal information needs can be ambiguous: a conversation may reasonably fit multiple topics, and human annotators may disagree. This study does not measure inter-annotator agreement or model label ambiguity directly. Third, the training budget is limited. The reported RL and classifier runs should therefore be read as evidence that small specialists can work in this setting, not as a fully optimized comparison across all model families and hyperparameters. Fourth, the data and methodology are English-language and based on the WildChat/Legal-Questions setup used here; multilingual and jurisdiction-specific routing are left for future work. Fifth, the reported cost comparison uses list prices and approximate encoder estimates rather than a full systems benchmark.

Finally, we do not perform broader human evaluation of RL outputs beyond the audit set, including qualitative assessment of generated reasoning traces. This also applies to Don’t-Know detection: abstention is evaluated through exact-match labels, uncertainty buckets, and rollout metrics, not through LLM-assisted judging or human assessment of whether each abstention was substantively appropriate. This mirrors a limitation noted by UniLaw-R1: final-answer accuracy can improve under RL even when step-by-step legal reasoning has not been evaluated in depth [21].

VIII. CONCLUSION

This thesis investigated whether small specialist models can support legal query routing when trained from frontier-model annotations. Within the silver-label setting studied here, the results are encouraging but conditional. For RQ1, a 4B instruction-tuned LM with reasoning and RL reaches 91.2% legal-guidance agreement with GPT-5.4 labels on the test split, suggesting that a much cheaper specialist can approximate the teacher for this routing task. This should not be interpreted as proving legal correctness, because the labels remain GPT-5.4-derived rather than independent ground truth.

For RQ2, the experiments show that explicit reasoning can materially improve small-model routing performance on the Legal-Questions test split: topic accuracy increases by +7.4 pp for the base 3B model and by +12.0 pp after RL fine-tuning when switching from NO THINK to THINK ON. For RQ3, the ternary Don’t-Know reward provides a practical abstention mechanism by moving a larger share of wrong answers into high-uncertainty rollouts. Its contribution is therefore selective routing rather than automatic correctness: uncertain cases can

be escalated instead of being forced through the small model. For RQ4, longer context appears to help when it lets the model retain more of the conversation: full-conversation input helps ModernBERT, while Legal-BERT’s 512-token limit makes user-only input more effective.

The practical conclusion is that no single model family dominates all settings. ModernBERT is a strong option for stable classification tasks where the label set is fixed and low cost matters most. Small generative LMs are more expensive, but they are easier to reuse when the same model must support prompting, changed schemas, in-context examples, abstention, or related downstream tasks. Future work should replace or supplement GPT-5.4 silver labels with multi-annotator human labels, measure disagreement explicitly, test multilingual and jurisdiction-specific settings, and explore larger training budgets before drawing stronger claims about legal reliability.

REFERENCES

- [1] A. K. Chatterji, J. De Freitas, W. Li, M. Haslberger, S. Zheng, D. Baxter, M. Pieler, J. Robinson, A. Drake, B. Cira, A. Aumjaud, R. Willer, E. Brynjolfsson, J. J. Horton, L. King, R. Koning, S. Noy, B. Zhang, N. Jia, M. Zhang, M. Kim, P. Tambe, S. Zhou, P. Choudhury, Y. Bigman, E. Kalliamvakou, R. Doherty, K. Fahrback, A. R. Doshi, J. Reich, C. Turco, D. Wedel, L. Lucas, S. Chilazi, E. Karger, K. Karbownik, L. Bossi, V. Capraro, Z. Nabi, R. Li, W. Yan, R. Uuk, R. Gosline, K. C. Kellogg, M. Lavanchy, H. Lifshitz-Assaf, E. Lyons, R. K. Lyons, B. S. Manning, H. Peng, S. Noy, S. Peng, M. Raj, P. Santhanam, S. W. Si, L. Sivan, S. Skiena, D. Sraer, I. Tewari, B. Wang, A. Wood, S. Yamane, Z. Yang, R. Zhang, F. Zhu, and O. E. R. Team, “How people use chatgpt,” National Bureau of Economic Research, Working Paper 34255, September 2025. [Online]. Available: <https://www.nber.org/papers/w34255>
- [2] Q. Mei, Y. Xie, W. Yuan, and M. L. McDermott, “Real-world medical advice seeking with large language models,” 2025. [Online]. Available: <https://arxiv.org/abs/2506.18183>
- [3] A. Paruchuri, A. Kupke, G. Siwo, A. F. Cooper, and J. Zou, “What’s up, doc? a dataset for studying medical advice in llms,” 2025, accepted to Findings of EMNLP 2025. [Online]. Available: <https://arxiv.org/abs/2506.21532>
- [4] J. Kuk, V. Khovanskaya, and M. A. Madaio, “LLMs & legal aid: Understanding legal needs exhibited through user queries,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.01711>
- [5] A. Louis and G. Spanakis, “A statutory article retrieval dataset in French,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 6789–6803. [Online]. Available: <https://aclanthology.org/2022.acl-long.468/>
- [6] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng, “Wildchat: 1m chatgpt interaction logs in the wild,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.01470>
- [7] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” 2023. [Online]. Available: <https://arxiv.org/abs/2205.11916>
- [8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou, “Chain of thought prompting elicits reasoning in large language models,” *CoRR*, vol. abs/2201.11903, 2022. [Online]. Available: <https://arxiv.org/abs/2201.11903>
- [9] N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu, Y. Gu, S. Malik, V. Graf, J. D. Hwang, J. Yang, R. L. Bras, O. Taffjord, C. Wilhelm, L. Soldaini, N. A. Smith, Y. Wang, P. Dasigi, and H. Hajishirzi, “Tulu 3: Pushing frontiers in open language model post-training,” 2025. [Online]. Available: <https://arxiv.org/abs/2411.15124>
- [10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford,

- I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901. [Online]. Available: <https://arxiv.org/abs/2005.14165>
- [11] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, “PaLM: Scaling language modeling with pathways,” 2022. [Online]. Available: <https://arxiv.org/abs/2204.02311>
- [12] S. Gunasekar, Y. Zhang, J. Anreja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. S. Behl, X. Wang, S. Bubeck, R. Eldan, A. T. Kalai, Y. T. Lee, and Y. Li, “Textbooks are all you need,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.11644>
- [13] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, “How does NLP benefit legal system: A summary of legal artificial intelligence,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5218–5230. [Online]. Available: <https://aclanthology.org/2020.acl-main.466>
- [14] R. L. Sandefur, “Accessing justice in the contemporary USA: Findings from the community needs and services study,” American Bar Foundation, Tech. Rep., 2014. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2478040
- [15] H. Sargeant, A. Izzidien, and F. Steffek, “Topic classification of case law using a large language model and a new taxonomy for UK law: AI insights into summary judgment,” *Artificial Intelligence and Law*, 2025. [Online]. Available: <https://doi.org/10.1007/s10506-025-09434-0>
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>
- [17] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “LEGAL-BERT: The muppets straight out of law school,” 2020. [Online]. Available: <https://arxiv.org/abs/2010.02559>
- [18] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli, “Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference,” 2024. [Online]. Available: <https://arxiv.org/abs/2412.13663>
- [19] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Xu, H. Ding, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Chen, J. Yuan, J. Tu, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. You, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Zhou, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma,
- Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang, “Deepseek-r1 incentivizes reasoning in llms through reinforcement learning,” *Nature*, vol. 645, no. 8081, pp. 633–638, September 2025. [Online]. Available: <http://dx.doi.org/10.1038/s41586-025-09422-z>
- [20] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.03300>
- [21] H. Cai, S. Zhao, L. Zhang, X. Shen, Q. Xu, W. Shen, Z. Wen, and T. Ban, “Unilaw-r1: A large language model for legal reasoning with reinforcement learning and iterative inference,” 2025. [Online]. Available: <https://arxiv.org/abs/2510.10072>
- [22] X. Dai, B. Xu, Z. Liu, Y. Yan, H. Xie, X. Yi, S. Wang, and G. Yu, “Legal δ : Enhancing legal reasoning in llms via reinforcement learning with chain-of-thought guided information gain,” 2026. [Online]. Available: <https://arxiv.org/abs/2508.12281>
- [23] A. Chegini, H. Kazemi, G. Souza, M. Safi, Y. Song, S. Bengio, S. Williamson, and M. Farajtabar, “Reasoning’s razor: Reasoning improves accuracy but can hurt recall at critical operating points in safety and hallucination detection,” 2025. [Online]. Available: <https://arxiv.org/abs/2510.21049>
- [24] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, “Training language models to follow instructions with human feedback,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27730–27744. [Online]. Available: <https://arxiv.org/abs/2203.02155>
- [25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” 2021. [Online]. Available: <https://arxiv.org/abs/2106.09685>
- [26] Thinking Machines Lab, “Lora without regret,” <https://thinkingmachines.ai/blog/lora/>, 2025, accessed: 2026-05-23.
- [27] T. Chu, Y. Zhai, J. Yang, S. Tong, S. Xie, D. Schuurmans, Q. V. Le, S. Levine, and Y. Ma, “Sft memorizes, rl generalizes: A comparative study of foundation model post-training,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.17161>
- [28] H. Jin, S. Luan, S. Lyu, G. Rabusseau, R. Rabbany, D. Precup, and M. Hamdaqa, “Rl fine-tuning heals ood forgetting in sft,” 2025. [Online]. Available: <https://arxiv.org/abs/2509.12235>
- [29] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, “Why language models hallucinate,” 2025. [Online]. Available: <https://arxiv.org/abs/2509.04664>
- [30] Y. Wei, Y. Wang, Z. Wang, C. Li, H. Yan, X. Qu, and Y. Cheng, “Truthrl: Incentivizing truthfulness in llms via reinforcement learning,” 2025. [Online]. Available: <https://arxiv.org/abs/2509.25760>
- [31] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging LLM-as-a-judge with MT-Bench and chatbot arena,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023. [Online]. Available: <https://arxiv.org/abs/2306.05685>

APPENDIX A

RELEASED RESOURCES AND MODEL IDENTIFIERS

- **Balanced legal-classification dataset:** V2-Balanced dataset.
- **Legal-Questions companion dataset:** Legal-Questions dataset.
- **Visualization embeddings:** WildChat-Legal-Visualization-V2.
- **Legal Conversation Explorer:** Hugging Face Space.
- **Manually annotated GPT-5.4 audit sheet:** GPT-5.4 manual audit annotations.
- **ModernBERT-base checkpoint:** ModernBERT-base.
- **Legal-guidance ModernBERT classifier:** seeks-guidance classifier.
- **Primary-topic ModernBERT classifier:** primary-topic classifier.
- **Legal-guidance Legal-BERT classifier:** Legal-BERT seeks-guidance classifier.
- **Primary-topic Legal-BERT classifier:** Legal-BERT primary-topic classifier.
- **Qwen3.5-4B no-DK RL checkpoint:** base Qwen/Qwen3.5-4B, suffix jpbcbq5fsepwcuu364gb6shta.
- **Qwen3.5-4B +DK RL checkpoint:** base Qwen/Qwen3.5-4B, suffix fuidmk9mg03ggr0pqn5sq71m.

The Qwen RL checkpoint strings are Prime Inference deployment identifiers rather than Hugging Face repository identifiers.

A. Model and Checkpoint Identifiers

TABLE VIII

EXACT MODEL AND CHECKPOINT IDENTIFIERS USED IN THE PAPER.

Role	Identifier
Teacher / silver labeler	gpt-5.4
Qwen student	Qwen/Qwen3.5-4B
Llama student	meta-llama/Llama-3.2-3B-Instruct
ModernBERT base	answerdotai/ModernBERT-base
Qwen no-DK RL	Qwen/Qwen3.5-4B + jpbcbq5fsepwcuu364gb6shta
Qwen +DK RL	Qwen/Qwen3.5-4B + fuidmk9mg03ggr0pqn5sq71m

APPENDIX B

DATASET CONSTRUCTION DETAILS

The V2-Balanced dataset is built in two GPT-5.4 passes. A low-effort first pass triages WildChat-1M conversations into legal, non-legal, and maybe-legal candidates. A high-effort second pass applies the full schema to legal and maybe-legal rows, including `contains_legal_content`, `seeks_legal_guidance`, `personalized_or_operational`, topic labels, primary topic, uncertainty, and relevant turn IDs. Short rationales

are generated during labeling but are not published. An English-dominance filter removes false positives before balancing. The pre-balance stage-2 artifact contains 7,256 rows: 1,062 legal and 6,194 non-legal. The released balanced split keeps all 1,061 legal-guidance rows and samples 1,061 non-guidance rows uniformly at random from the English non-guidance pool (seed 43), then shuffles and splits with seed 42.

TABLE IX
PUBLISHED V2-BALANCED SPLIT SIZES.

Split	Total	Legal	Non-legal
Total	2,122	1,061	1,061
Train	1,909	950	959
Validation	106	55	51
Test	107	56	51

TABLE X
GPT-5.4 UNCERTAINTY DISTRIBUTION IN THE PUBLISHED V2-BALANCED DATASET. THERE ARE NO HIGH-UNCERTAINTY ROWS IN VALIDATION OR TEST.

Split	All rows low/med/high	Non-legal low/med/high
Train	1,309 / 600 / 0	933 / 26 / 0
Validation	75 / 31 / 0	49 / 2 / 0
Test	65 / 42 / 0	48 / 3 / 0

GPT-5.4’s lack of HIGH-uncertainty labels is not a post-processing merge. High uncertainty was simply very rare for the teacher on this coarse routing task; for example, only 8 of the 6,194 non-legal stage-2 rows received HIGH. This is plausible because the large teacher model treated most examples as either clear legal-guidance requests or clear non-guidance cases.

TABLE XI
PRIMARY-TOPIC DISTRIBUTION FOR LEGAL-GUIDANCE ROWS IN TRAIN, VALIDATION, AND TEST ($n = 950 / 55 / 56$).

Topic	Train	Val	Test
Immigration	118 (12.4%)	12 (21.8%)	7 (12.5%)
Other	156 (16.4%)	7 (12.7%)	5 (8.9%)
Intellectual Property	99 (10.4%)	7 (12.7%)	3 (5.4%)
Work	72 (7.6%)	6 (10.9%)	7 (12.5%)
Housing	60 (6.3%)	6 (10.9%)	4 (7.1%)
Criminal Justice	113 (11.9%)	4 (7.3%)	5 (8.9%)
Consumer Debt	51 (5.4%)	3 (5.5%)	5 (8.9%)
Business	141 (14.8%)	2 (3.6%)	14 (25.0%)
Contracts	34 (3.6%)	2 (3.6%)	2 (3.6%)
Civil Rights	6 (0.6%)	2 (3.6%)	1 (1.8%)
Public Benefits	20 (2.1%)	2 (3.6%)	1 (1.8%)
Family	34 (3.6%)	1 (1.8%)	1 (1.8%)
Data Privacy	26 (2.7%)	1 (1.8%)	0 (0.0%)
International / Cross-Border	20 (2.1%)	0 (0.0%)	1 (1.8%)

The Legal-Questions companion dataset contains 1,475 rows derived from 1,061 source conversations and uses an 80/10/10 split of 1,180 / 147 / 148 with seed 42.

APPENDIX C TOPIC TAXONOMY

APPENDIX D GPT-5.4 LABELING PROMPT

The GPT-5.4 teacher uses the OpenAI API model identifier `gpt-5.4`, temperature 0.5, and the structured JSON schema `LegalClassificationOutput`. The topic block in the prompt is filled with the 14-topic taxonomy in Table XII. The prompt instructs the model to base decisions primarily on user messages, using assistant messages only when they help recover an elliptical user request.

The labeling instructions define `contains_legal_content` as discussion of law, legal rights, obligations, courts, legal documents, legal procedures, regulations, enforcement, or official appeals. They define `seeks_legal_guidance` as a user request about legality, rights or duties, legal procedures or deadlines, required legal documents, or practical legal next steps. The prompt marks exam questions, translation or summarization tasks, coding/tooling requests, policy discussion without a request about legal requirements, and fictional roleplay as non-guidance cases.

For topic routing, the prompt asks GPT-5.4 to assign one to three directly necessary topics, choose one primary topic, use `OTHER` only when no taxonomy category fits, and return no topics when `seeks_legal_guidance=false`. The required JSON fields are `contains_legal_content`, `seeks_legal_guidance`, `personalized_or_operational`, `topics`, `primary_topic`, relevant user and assistant turn IDs, `uncertainty`, and a short rationale. The rationale is used internally during labeling but is not published in the released dataset.

APPENDIX E STUDENT TASK PROMPTS

The student prompts are generated from the repository prompt builders used by both training and evaluation. The boxes below transcribe the written prompt templates from the legal-classification and Legal-Questions environments, with `{topic_block}`, `{conversation_text}`, and `{question}` left as placeholders. The reported generative evaluations set `require_uncertainty=true`, so the model must self-report `LOW`, `MEDIUM`, or `HIGH` before the final label tags. When `allow_dont_know=true`, the same template adds `DONT_KNOW` as an allowed value and tells the model to consider it when `uncertainty` is high.

A. Legal Conversation Classification

Legal conversation classification prompt

You classify a user–assistant conversation.

Task 1 – Legal guidance: Is the USER seeking legal guidance or legal interpretation? Answer `<legal>` with YES or NO [or `DONT_KNOW` if abstention is enabled]. Focus on user intent.

Task 2 – Topic: If `<legal>` is YES, pick exactly one topic from the list below. If `<legal>` is NO, write N/A for `<topic>`.

Legal guidance examples: Legal guidance (YES) means the user wants legal interpretation, rights/obligations, or legal procedure for a real situation, similar to “Can my landlord do this under Dutch law?”, “What are my rights if I was fired without notice?”, or “Is this contract clause enforceable?”. Answer NO when the user is not seeking legal guidance, such as exam-style multiple-choice questions, pure summarization or translation of legal text, or tooling/dataset tasks, even if the text mentions law.

When DONT_KNOW is enabled: use `DONT_KNOW` for `<legal>` when the legal-guidance decision cannot be made confidently. If `<legal>` is `DONT_KNOW`, also set `<topic>` to `DONT_KNOW`. Use `DONT_KNOW` for `<topic>` when `<legal>` is YES but a topic cannot be picked confidently.

Uncertainty: Before `<legal>`, declare confidence with `<uncertainty>LOW</uncertainty>`, `<uncertainty>MEDIUM</uncertainty>`, or `<uncertainty>HIGH</uncertainty>`. `LOW` means confident, `MEDIUM` means somewhat unsure, and `HIGH` means very unsure.

Topic list and descriptions: `{topic_block}`

Exact values: copy these strings verbatim; no other spelling is accepted. `<legal>` must be one of YES, NO [and optionally `DONT_KNOW`]. `<topic>` must be one of the taxonomy labels, N/A [and optionally `DONT_KNOW`]. `<uncertainty>` must be one of `LOW`, `MEDIUM`, or `HIGH`.

Formatting rules: use `ALL_CAPS_WITH_UNDERSCORES` for topic names; N/A uses a forward slash; `DONT_KNOW` uses an underscore; topic names are graded case-insensitively, but the prompt asks the model to use all caps.

Input: `{conversation_text}`.

No THINK response format:
`<uncertainty>LOW|MEDIUM|HIGH</uncertainty>`
then `<legal>YES or NO [or DONT_KNOW]</legal>`
and `<topic>exact topic name or N/A [or DONT_KNOW]</topic>`. The prompt includes examples for a legal employment question (YES/WORK) and a non-legal query (NO/N/A).

THINK ON response format: `<think>...</think>` followed by the same uncertainty, legal, and topic tags. The prompt asks the model to reason step by step inside `<think>` tags and to write nothing after `</topic>`.

B. Legal-Questions Topic Classification

Legal-Questions prompt

You classify a standalone legal question.

Task – Topic: Pick exactly one topic from the list below that best matches the question.

When DONT_KNOW is enabled: use `DONT_KNOW` when you cannot confidently identify a best topic for this question. Pick a specific topic only when you are reasonably confident it matches the question.

Uncertainty: Before `<topic>`, declare confidence with `<uncertainty>LOW</uncertainty>`, `<uncertainty>MEDIUM</uncertainty>`, or `<uncertainty>HIGH</uncertainty>`. `LOW` means confident, `MEDIUM` means somewhat unsure, and `HIGH` means

TABLE XII
LEGAL TOPIC TAXONOMY USED FOR PRIMARY-TOPIC ROUTING.

Topic	Description
Family	Marriage, divorce, child custody, child support, alimony, adoption, guardianship, domestic violence, parentage, family-status disputes.
Housing	Rent, eviction, landlord-tenant disputes, habitability, deposits, mortgages, foreclosure, neighbors, housing subsidies.
Work	Employment contracts, wages, dismissal, discrimination at work, leave, workplace safety, severance, freelancers when the main issue is labor rights.
Public Benefits	Unemployment benefits, disability, pensions, welfare, public assistance, eligibility, reductions, sanctions, appeals on benefits.
Criminal Justice	Police, arrest, criminal charges, fines, prosecution, defense, victims' rights, probation, criminal procedure.
Consumer Debt	Purchases, warranties, subscriptions, refunds, scams, debt collection, loans, bankruptcy, credit, repossession, consumer finance.
Contracts	Private civil agreements and breach or interpretation issues not better covered by work, housing, consumer, or business.
Immigration	Visas, residence permits, asylum, citizenship, deportation, family migration, immigration status and related procedures.
Business	Company formation, shareholder issues, commercial compliance, business operations, B2B disputes, self-employment when the main issue is business law.
Data Privacy	Personal data, surveillance, GDPR/privacy rights, data deletion, consent, monitoring, platform data practices.
Intellectual Property	Copyright, trademark, patent, trade secrets, licensing, infringement, ownership of creative or technical works.
Civil Rights	Discrimination outside employment/housing, free speech, due process, equal treatment, constitutional or human-rights-style claims.
International / Cross-Border	Choice of law, jurisdiction, treaty-based questions, cross-border enforcement, multi-country disputes where cross-border law is central.
Other	Genuinely legal but not covered above.

very unsure. With DONT_KNOW enabled, high uncertainty is a cue to consider abstaining.

Topic list and descriptions: {topic_block}

Exact values: copy one of these strings verbatim, no other spelling is accepted. <topic> must be one of the taxonomy labels [and optionally DONT_KNOW]. <uncertainty> must be one of LOW, MEDIUM, or HIGH.

Formatting rules: use ALL_CAPS_WITH_UNDERSCORES, for example INTELLECTUAL_PROPERTY, PUBLIC_BENEFITS, or DATA_PRIVACY. If enabled, DONT_KNOW uses an underscore.

Input: optional context followed by {question}.

No THINK response format:
<uncertainty>LOW|MEDIUM|HIGH</uncertainty>
then <topic>exact topic name [or DONT_KNOW]</topic>. The prompt includes examples for WORK and HOUSING.

THINK ON response format:
<reasoning>...</reasoning> followed by the same uncertainty and topic tag. The prompt asks the model to reason step by step inside <reasoning> tags and to write nothing after </topic>.

Legal conversation, NO THINK:
<uncertainty>LOW|MEDIUM|HIGH</uncertainty>
followed by <legal>...</legal> and <topic>...</topic>.

Legal conversation, THINK ON: <think>...</think> followed by the same uncertainty, legal, and topic tags.

Legal-Questions, NO THINK:
<uncertainty>LOW|MEDIUM|HIGH</uncertainty>
followed by <topic>...</topic>.

Legal-Questions, THINK ON:
<reasoning>...</reasoning> followed by the same uncertainty and topic tags.

APPENDIX F ENCODER TRAINING AND SWEEPS

TABLE XIII
MODERNBERT FINE-TUNING SETUP FOR THE REPORTED ENCODER CLASSIFIERS. BOTH USE THE MODERNBERT-BASE TOKENIZER AND ARE SELECTED BY THE BEST VALIDATION WEIGHTED F1 (SEEKS) OR ACCURACY (TOPIC).

Classifier	Input	Max len	LR	Batch
seeks-guidance	full conversation	4096	8×10^{-5}	32
primary-topic	full conversation	4096	10^{-4}	32

Student output formats

NO THINK corresponds to prompt_style=fast; THINK ON corresponds to prompt_style=reasoning. The uncertainty tag is required in the reported generative evaluations.

Both ModernBERT classifiers are trained for 8 epochs. The effective batch size is 32 for both checkpoints (batch 8 with 4 gradient-accumulation steps). The primary-topic checkpoint is selected by best validation accuracy (83.96% on V2-Balanced validation); the seeks-guidance checkpoint is selected by best validation weighted F1 (94.34%). Joint pipeline test accuracy at 4096 tokens is 87.9% legal and 78.5% topic (Table III); extending to 8192 tokens does not change headline legal or topic accuracy on this split.

TABLE XIV

LEGAL-BERT FINE-TUNING SETUP. BOTH CLASSIFIERS USE 512-TOKEN TRUNCATION, 8 EPOCHS, AND A 24-TRIAL SWEEP OVER LEARNING RATE ($\{2 \times 10^{-5}, 5 \times 10^{-5}, 8 \times 10^{-5}, 10^{-4}\}$), INPUT MODE (USER VS. FULL), AND EFFECTIVE BATCH SIZE ($\{32, 64, 128\}$). SELECTION: VALIDATION WEIGHTED F1 FOR SEEKS_GUIDANCE, VALIDATION ACCURACY FOR PRIMARY_TOPIC.

Classifier	Input	Max len	LR	Batch
seeks-guidance	user turns only	512	10^{-4}	32
primary-topic	user turns only	512	10^{-4}	128

The winning Legal-BERT seeks_guidance checkpoint reaches 91.5% validation F1 with user input; the best full-conversation trial reaches 88.7% validation F1. Joint pipeline test accuracy on V2-Balanced test is 84.1% legal and 72.0% topic (Table III).

Table XV reports the complete 24-trial validation sweep for both Legal-BERT classifiers. For seeks-guidance, user input dominates (best 91.5% vs. 88.7% for the full conversation, and a higher mean of 90.0% vs. 87.3%), and all of the top trials use user input. For primary-topic, the two input modes are close (best 78.4% user vs. 77.9% full). The deployed Legal-BERT checkpoints therefore both use user input.

Table XVI summarizes both encoders’ input-mode validation sweeps, and Table XVII reports the held-out test matrix where each encoder is evaluated with user-only and full-conversation serialization on both the seeks_guidance and primary_topic checkpoints. For ModernBERT, the full conversation wins both heads of the sweep (best/mean val F1 93.4/92.4 seeks and 84.9/81.3 topic, vs. 91.5/90.5 and 76.4/74.3 for user input); the deployed checkpoints are the earlier 4096-token models, whose validation F1 (94.3% seeks, 84.9% topic) the newer long-context input-mode sweep did not exceed. For Legal-BERT, user input wins seeks decisively and is the better choice on topic (highest best F1), so both deployed checkpoints use user turns. In both cases the sweep-selected input mode also wins both heads on test (Table XVII).

A. Headline Comparison Figures

Figures 3 and 4 visualize the encoder-family and training-paradigm comparisons whose numbers are reported in Tables IV and III, respectively.

TABLE XV

LEGAL-BERT VALIDATION SWEEP (NEWER 24-TRIAL PASS). VALIDATION WEIGHTED F1 (%) ON V2-BALANCED FOR THE SEEKS-GUIDANCE AND PRIMARY-TOPIC CLASSIFIERS ACROSS INPUT MODE, LEARNING RATE, AND EFFECTIVE BATCH SIZE. BEST PER COLUMN IN BOLD.

Input	LR	Batch	Seeks F1	Topic F1
User	2×10^{-5}	32	89.6	71.7
	2×10^{-5}	64	90.5	69.8
	2×10^{-5}	128	88.7	60.3
	5×10^{-5}	32	90.6	74.8
	5×10^{-5}	64	90.6	73.2
	5×10^{-5}	128	87.7	70.8
	8×10^{-5}	32	91.5	76.3
	8×10^{-5}	64	90.5	76.6
	8×10^{-5}	128	88.6	44.1
	10^{-4}	32	91.5	74.7
	10^{-4}	64	90.6	77.7
	10^{-4}	128	89.6	78.4
Full	2×10^{-5}	32	87.7	71.7
	2×10^{-5}	64	84.9	68.4
	2×10^{-5}	128	88.7	63.7
	5×10^{-5}	32	87.7	77.7
	5×10^{-5}	64	88.7	75.2
	5×10^{-5}	128	87.7	64.7
	8×10^{-5}	32	86.8	77.9
	8×10^{-5}	64	87.7	73.8
	8×10^{-5}	128	85.9	73.5
	10^{-4}	32	88.7	76.0
	10^{-4}	64	87.7	76.0
	10^{-4}	128	84.8	70.2

TABLE XVI

ENCODER INPUT-MODE VALIDATION SWEEPS ON V2-BALANCED: WEIGHTED F1 (%) PER HEAD BY INPUT MODE. MODERNBERT SWEEPS LEARNING RATE $\{5 \times 10^{-5}, 10^{-4}\} \times$ BATCH $\{32, 64\}$ WITH AN 8192-TOKEN CONTEXT (4 TRIALS PER CELL); LEGAL-BERT SWEEPS LEARNING RATE $\{2, 5, 8\} \times 10^{-5}, 10^{-4} \times$ BATCH $\{32, 64, 128\}$ AT 512 TOKENS (12 TRIALS PER CELL; FULL GRID IN TABLE XV). BOLD MARKS THE HIGHER VALUE PER ENCODER WITHIN EACH COLUMN.

Encoder	Input	Seeks F1		Topic F1	
		Best	Mean	Best	Mean
ModernBERT-base	full conversation	93.4	92.4	84.9	81.3
	user turns only	91.5	90.5	76.4	74.3
Legal-BERT	user turns only	91.5	90.0	78.4	70.7
	full conversation	88.7	87.3	77.9	72.4

APPENDIX G

GENERATIVE TRAINING AND EVALUATION SETUP

The reported Qwen3.5-4B Prime RL checkpoints use legal-cls-reasoning.toml for no-DK and legal-cls-reasoning-dk.toml for +DK. Both run for 150 steps with batch size 256, 8 rollouts per example, and a 4096-token generation budget; the +DK run uses the $w = -3$ wrong-answer penalty described in Section IV-D.

The label SFT baseline uses the same V2-Balanced split

TABLE XVII

ENCODER INPUT-MODE COMPARISON ON V2-BALANCED TEST ($n = 107$). EACH ROW EVALUATES THE SAME ENCODER CHECKPOINTS WITH USER-ONLY VS. FULL-CONVERSATION SERIALIZATION AT INFERENCE (MODERNBERT MAX LENGTH 4096; LEGAL-BERT 512). BOLD MARKS THE SWEEP-SELECTED (DEPLOYED) INPUT MODE PER ENCODER, WHICH WINS BOTH HEADS ON TEST.

Encoder	Input	Legal F1	Topic F1
ModernBERT-base	full conversation	87.6	83.2
	user turns only	82.3	60.9
Legal-BERT	full conversation	75.7	75.2
	user turns only	84.1	78.6

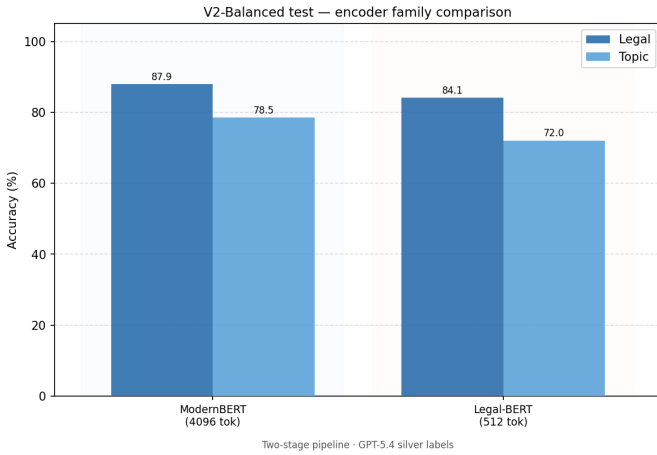


Fig. 3. ModernBERT vs. Legal-BERT on V2-Balanced test (legal-guidance and primary-topic accuracy).

and NO THINK prompt as the Qwen RL runs, with LoRA rank 64, 2 epochs, learning rate 10^{-4} , and batch size 32.

The Llama-3.2-3B-Instruct Legal-Questions reasoning run uses legal-questions-reasoning.toml, 150 steps, batch size 512, 8 rollouts per example, and a 2048-token generation budget.

Generative evaluation uses $r = 8$ rollouts per example for the main legal-classification evaluation and the Legal-Questions companion split.

APPENDIX H LEGAL-QUESTIONS TRANSFER (TEST)

The Legal-Questions companion split contains only conversations seeking legal guidance ($n = 148$ test rows). We report topic accuracy against GPT-5.4 silver reference labels. Qwen RL checkpoints are trained on V2-Balanced only; ModernBERT uses the same joint pipeline as on the main split.

On the test split, the no-DK-trained checkpoint generalizes strongly to Legal-Questions (84.5% topic under no-DK eval) and still beats the +DK-trained checkpoint under +DK eval (82.8 vs. 71.3% topic), mirroring the pattern on V2-Balanced.

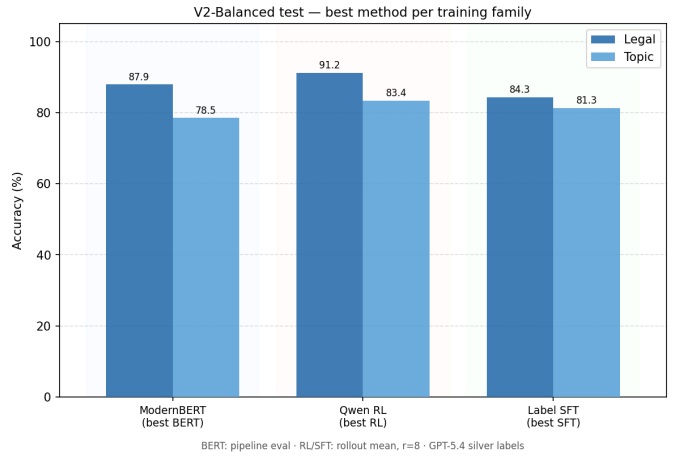


Fig. 4. Best BERT encoder vs. label SFT vs. RL on V2-Balanced test.

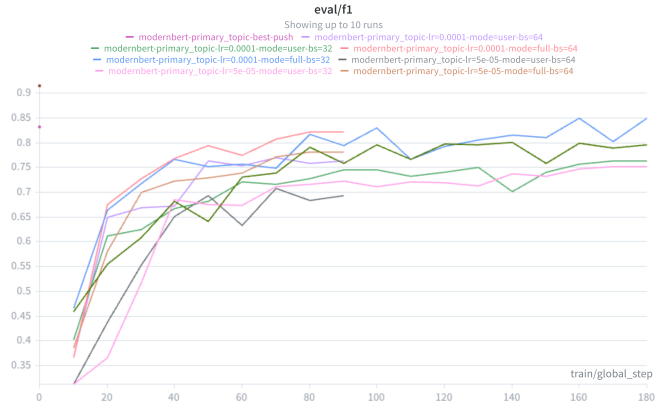


Fig. 5. Validation F1 curves for the ModernBERT and Legal-BERT legal-classification sweeps, shown together across encoder runs.

APPENDIX I REASONING-LIFT DETAIL (LLAMA-3.2-3B)

APPENDIX J UNCERTAINTY BREAKDOWNS (V2-BALANCED TEST)

A. Legal-guidance Uncertainty

B. Topic-level Uncertainty

C. +DK Checkpoint Comparison

Under identical +DK evaluation settings (`allow_dont_know=true`), the no-DK-trained checkpoint evaluated in the cross-evaluation setting outperforms the +DK-trained checkpoint on both legal-guidance and topic accuracy (Table III).

D. Qwen Uncertainty and Abstention Traces

APPENDIX K LEGAL CONVERSATION EXPLORER

We release an interactive *Legal Conversation Explorer* as a Hugging Face Space (Appendix A) for exploring the labeled dataset in embedding space. Conversations from

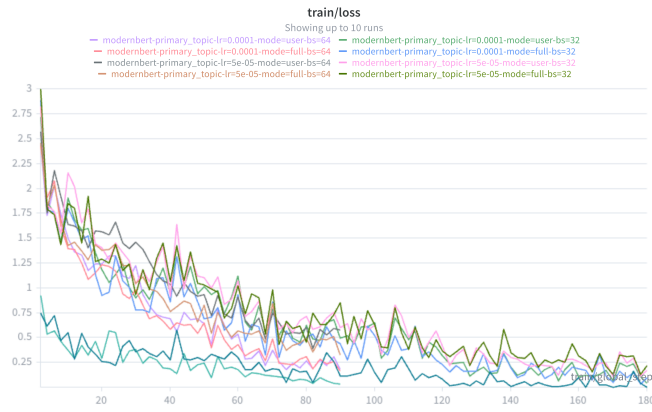


Fig. 6. Training-loss curves for the ModernBERT and Legal-BERT legal-classification sweeps.

TABLE XVIII

QWEN3.5-4B ON LEGAL-QUESTIONS TEST (ROLLOUT MEAN, $r = 8$). ANY-TOPIC INDICATES THAT THE PREDICTION APPEARS ANYWHERE IN THE GPT-5.4 TOPIC LIST; ABSTAIN INDICATES THE % OF ROLLOUTS OUTPUTTING DON'T-KNOW.

Configuration	Topic	Any-topic	Abstain
Qwen3.5-4B base – no-DK	49.1	50.4	0.0
Qwen3.5-4B base – +DK	50.7	51.9	0.6
Qwen3.5-4B RL – +DK	71.3	76.0	1.8
Qwen3.5-4B RL no-DK – no-DK	84.5	89.8	0.0
Qwen3.5-4B RL no-DK – +DK (cross)	82.8	89.3	0.8

WildChat-Legal-Visualization-V2 are embedded and projected to 2D with UMAP. The tool supports:

- **Embedding overview:** conversations plotted in 2D, colored by primary legal topic.
- **Topic filtering:** restrict the view to one or more of the 14 legal topic categories.
- **Semantic search:** type a natural-language query to find similar conversations by embedding distance.
- **Conversation detail:** click any point to read the full multi-turn conversation and its GPT-5.4 metadata.

The server loads precomputed embeddings from Hugging Face by default; local regeneration is documented in the repository `visualization/` directory.

A companion ModernBERT classifier demo allows users to provide a conversation or browse existing conversations and run the trained encoder checkpoints directly. The interface reports the legal-guidance decision and, when applicable, the primary-topic prediction with model probabilities alongside the GPT-5.4 reference label.

Topic accuracy on Legal-Questions (Qwen3.5-4B, rollout mean, $r=8$, test split)

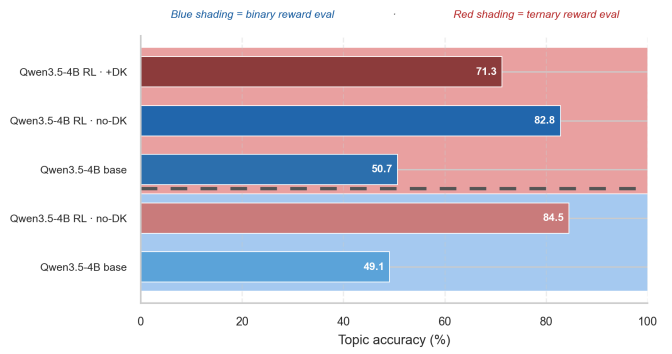


Fig. 7. Topic accuracy on Legal-Questions test for all five Qwen configurations (rollout mean, $r = 8$).

TABLE XIX

MODERNBERT JOINT PIPELINE ON LEGAL-QUESTIONS TEST ($n = 148$). ALL ROWS SEEK LEGAL GUIDANCE.

Metric	Legal	Topic
Accuracy (%)	87.8	71.6

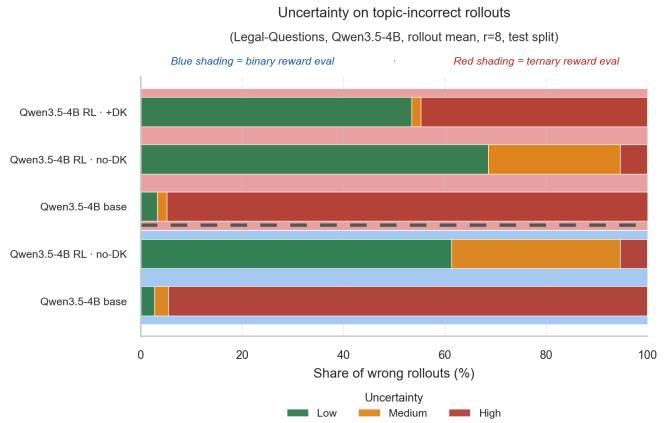


Fig. 8. Uncertainty on topic-incorrect rollouts, Legal-Questions test.

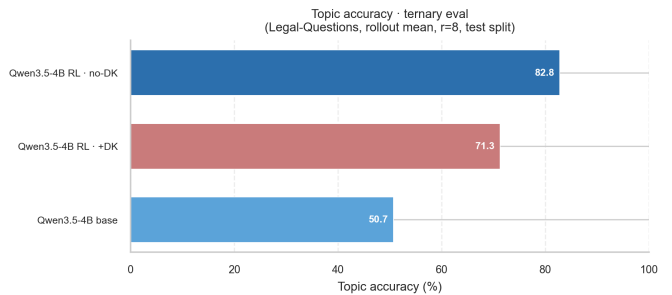


Fig. 9. Topic accuracy for +DK-trained vs. no-DK cross-eval on Legal-Questions test.

TABLE XX

FULL LLAMA-3.2-3B LEGAL-QUESTIONS TEST MATRIX (ROLLOUT MEAN, $r = 8$, NO-DK). ANY-TOPIC COUNTS A ROLLOUT AS CORRECT IF THE PREDICTED TOPIC APPEARS ANYWHERE IN THE GPT-5.4 TOPIC LIST.

Run	Topic	Any-topic
Base – NO THINK	45.6	54.1
Base – THINK ON	53.0	62.2
RL (fast-trained) – NO THINK	69.4	80.8
RL (reasoning-trained) – THINK ON	81.4	88.3

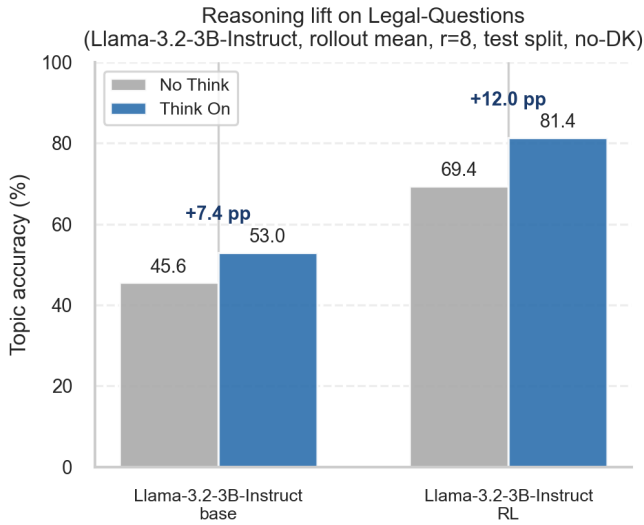


Fig. 10. Reasoning lift on the Legal-Questions test split (Llama-3.2-3B-Instruct, rollout mean, $r = 8$, no abstention). Switching from NO THINK to THINK ON improves accuracy by 7.4 pp for the base model and 12.0 pp for the RL-fine-tuned model.

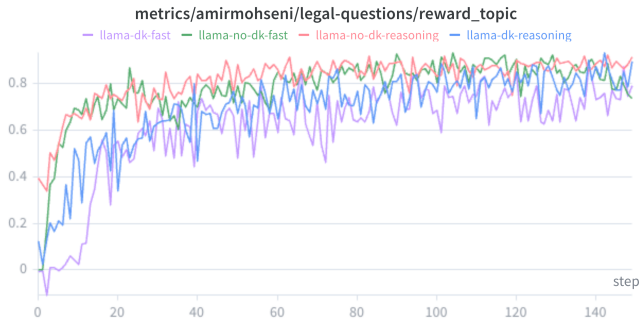


Fig. 11. Reward trace for Llama-3.2-3B-Instruct trained on the Legal-Questions topic-routing task.

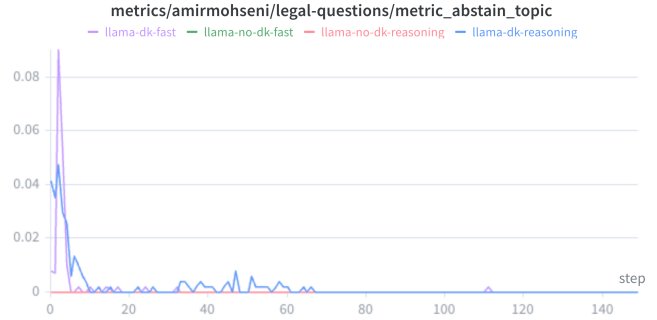


Fig. 12. Abstention trace for the Llama-3.2-3B-Instruct Legal-Questions runs.

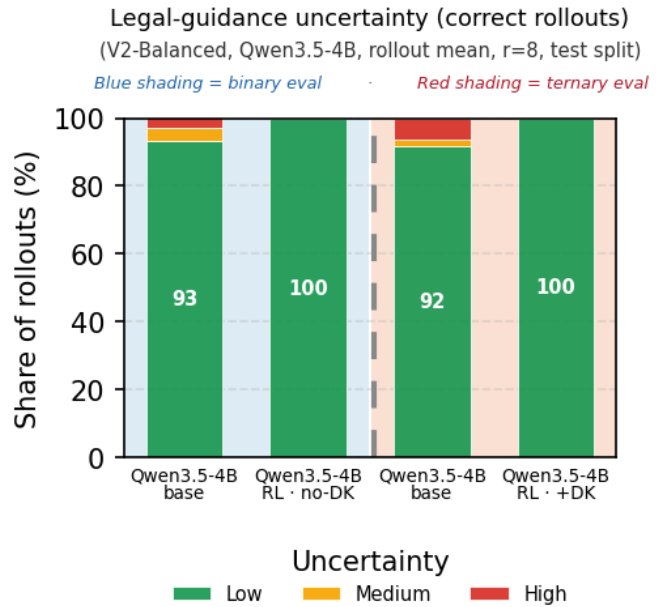


Fig. 13. Uncertainty distribution on *correct* legal-guidance rollouts on V2-Balanced test (rollout mean, $r = 8$). Nearly all correct rollouts are tagged LOW across all configurations. Missing tags are mapped to HIGH.

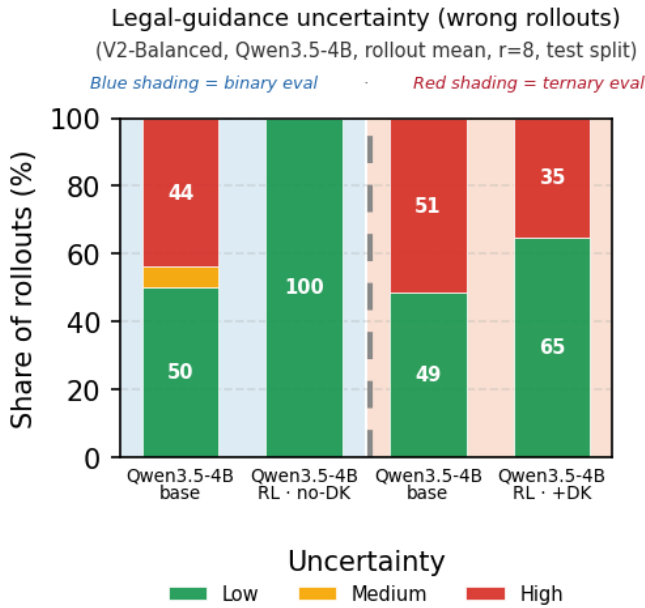


Fig. 14. Uncertainty distribution on *wrong* legal-guidance rollouts on V2-Balanced test (rollout mean, $r = 8$). The +DK variants push a larger share of wrong rollouts into the HIGH bucket, making model uncertainty a usable abstention signal. Missing tags are mapped to HIGH.

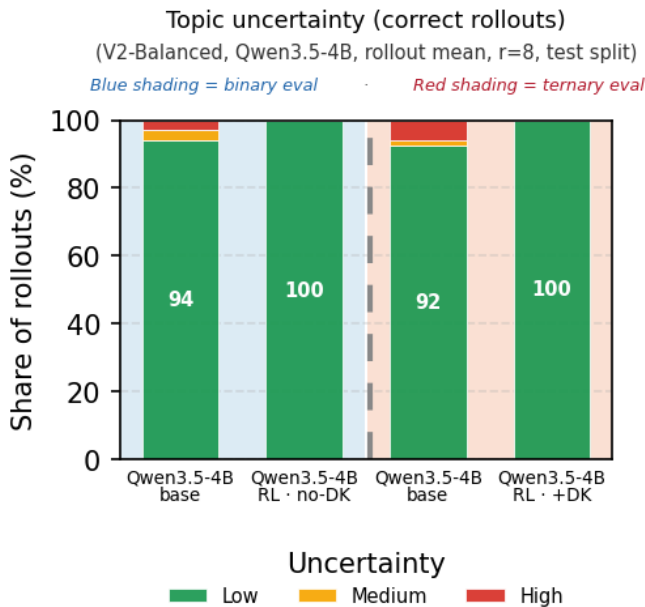


Fig. 15. Topic uncertainty on *correct* rollouts on V2-Balanced test (rollout mean, $r = 8$). Nearly all correct topic rollouts are tagged LOW.

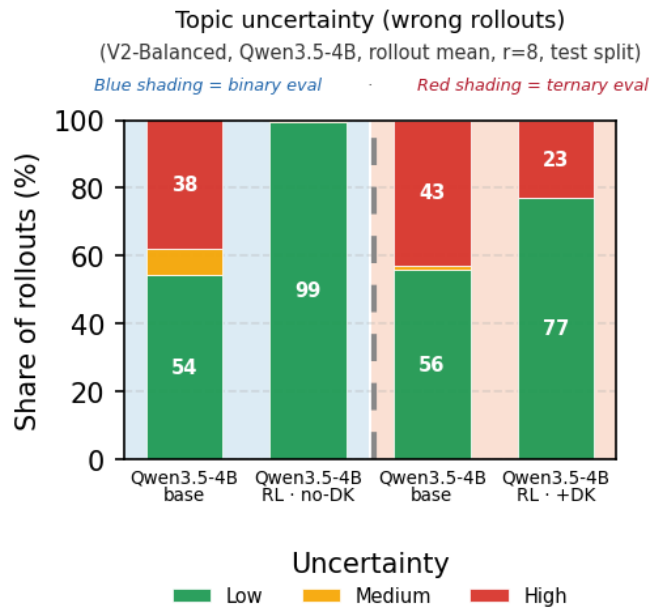


Fig. 16. Topic uncertainty on *wrong* rollouts on V2-Balanced test (rollout mean, $r = 8$). The +DK variants shift wrong rollouts toward HIGH on the topic head as well, complementing the legal-head result in Fig. 14.

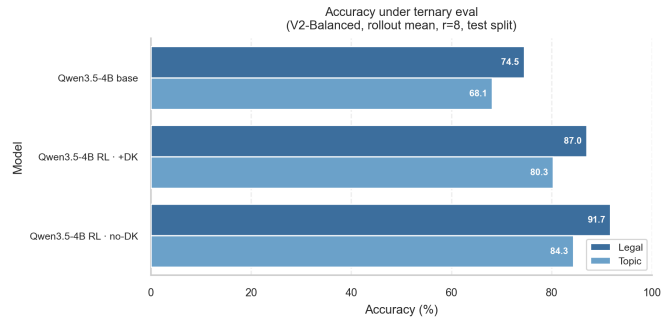


Fig. 17. Legal and topic accuracy for the +DK-trained vs. no-DK cross-eval checkpoints on V2-Balanced test.

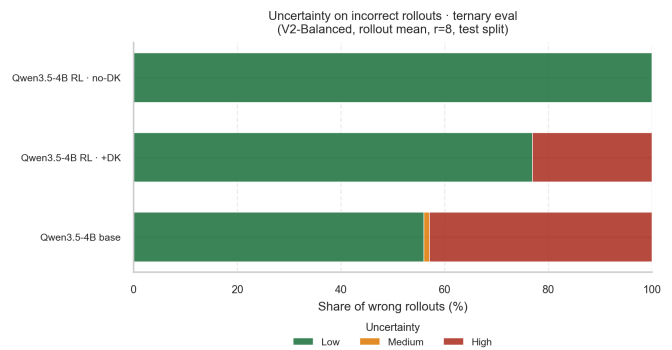


Fig. 18. Uncertainty mix on wrong rollouts for the +DK checkpoint comparison on V2-Balanced test.

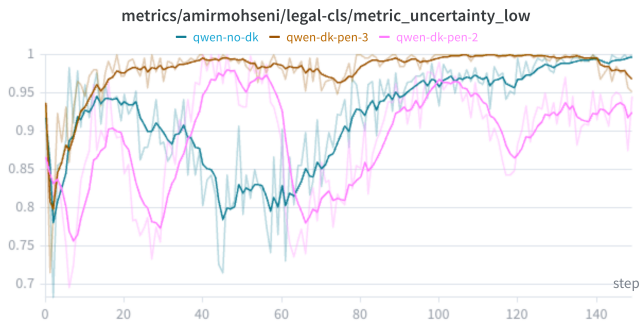


Fig. 19. Share of Qwen legal-classification rollouts tagged LOW uncertainty across the no-DK run and the +DK runs with different wrong-answer penalties.

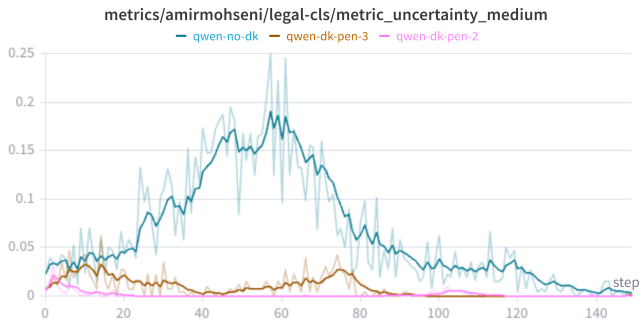


Fig. 20. Share of Qwen legal-classification rollouts tagged MEDIUM uncertainty across the no-DK run and the +DK penalty variants.

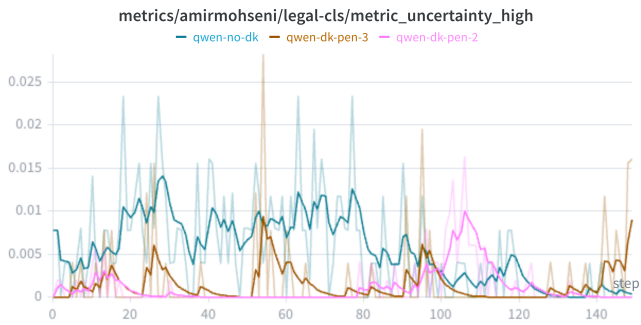


Fig. 21. Share of Qwen legal-classification rollouts tagged HIGH uncertainty across the no-DK run and the +DK penalty variants.

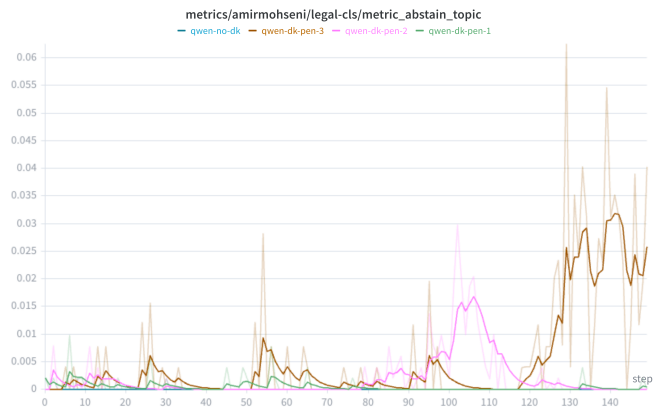


Fig. 22. Topic abstention rate for Qwen legal-classification runs, comparing the no-DK configuration with +DK configurations under the tested wrong-answer penalties.

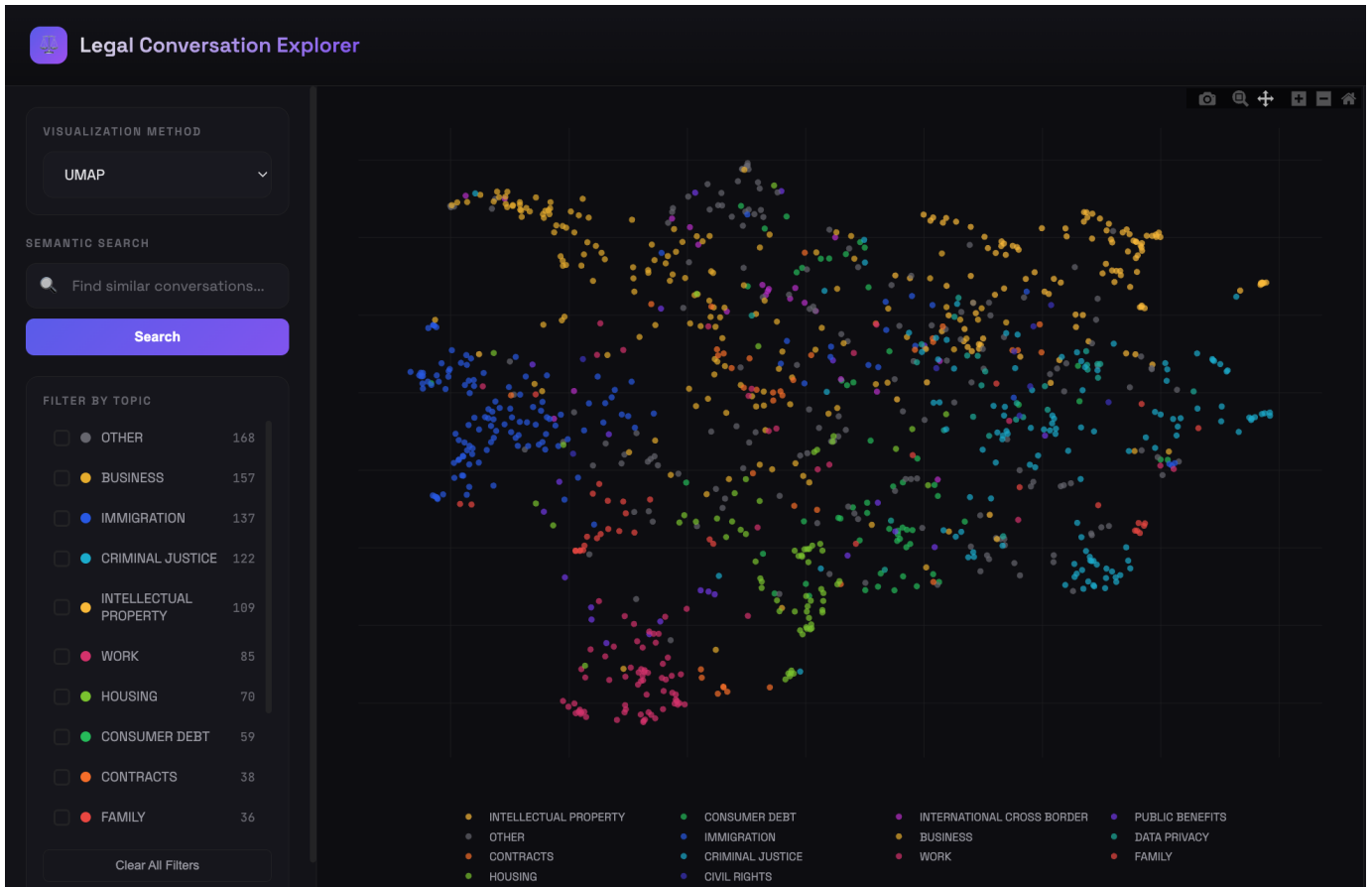


Fig. 23. Legal Conversation Explorer overview. The left panel exposes the UMAP projection, semantic search, and topic filters, while the main panel shows topic-colored conversations in embedding space.

Conversation ✕

PUBLIC BENEFITS **FAMILY** **UNCERTAINTY: LOW**

The user asks what government subsidies they may be eligible for in their real situation as a single father, which is primarily a public-benefits issue with family status relevant to eligibility.

[User]:
what are some subsidies I can claim as a single father

[Assistant]:
As a single father, there are several subsidies and benefits you may be eligible for depending on your location and circumstances. Here are some common ones:

1. Child Tax Credit: Many countries offer child tax credits or benefits to support families with dependent children.
2. Earned Income Tax Credit (EITC): In the United States, the EITC provides financial assistance to low and moderate-income families.
3. Childcare Subsidies: Some governments provide childcare subsidies to help with the cost of daycare or after-school programs.

LEGAL INCIDENTS (1)

INCIDENT 1 **PUBLIC BENEFITS** **FAMILY**

Turns: User [0] · Assistant [1]

What subsidies or benefits can I claim as a single father?

Fig. 24. Legal Conversation Explorer detail view for a public-benefits conversation. The drawer shows topic chips, GPT-5.4 uncertainty, the labeling rationale, the transcript, and the extracted legal incident metadata.

Conversation ✕

CONSUMER DEBT **UNCERTAINTY: LOW**

The user asks about options for forgiving private student loans, which is a practical question about debt obligations and possible legal/financial remedies.

[User]:
In Optometry private practice, if a practice grosses \$1 million, how much is the owner's take home usually?

[Assistant]:
The amount that the owner of an optometry private practice takes home can vary significantly and depends on several factors such as overhead expenses, operating costs, staff salaries, taxes, and the owner's role in the practice.

Typically, a general rule of thumb for healthcare practices is that the owner's take-home pay ranges from 30% to 50% of the gross revenue. However, it's essential to note that this is a broad estimate and may not be accurate for every practice.

It is crucial to consider that a portion of the gross revenue will be allocated towards various practice expenses like rent, utilities, insurance, equipment, marketing, and employee salaries. Additionally, taxes

LEGAL INCIDENTS (1)

INCIDENT 1 **CONSUMER DEBT**

Turns: User [4] · Assistant [5]

Is there any way to forgive private student loans?

Fig. 25. Legal Conversation Explorer detail view for a consumer-debt incident identified within a longer conversation.

What tax is paid if Wyoming Corporation forms a North Carolina LLC to operate in North Carolina would tax still be owed to Wyoming off income may be the LLC?

If a Wyoming Corporation forms a North Carolina LLC to operate in North Carolina, the LLC will be considered a separate entity for tax purposes. As a result, the LLC will be subject to North Carolina state taxes on the income it earns in North Carolina, which includes income from sales, services, and tangible property within the state.

The Wyoming Corporation itself, as a separate entity, will still be subject to Wyoming state taxes on its income derived from Wyoming. However, it may be possible for the Wyoming Corporation to avoid double taxation by filing as an S corporation or by establishing a limited liability company (LLC) taxed as a partnership, in which case the profits and losses would pass through to the shareholders or members and be taxed only at the individual level.

It's important to note that different states have different tax laws, and it's important to comply with all applicable tax laws to avoid penalties and fines. To ensure proper compliance, it's recommended to consult with a tax professional or accountant who is knowledgeable about the tax laws of both Wyoming and North Carolina and how they apply to your specific situation.

Next conversation Predict

RESULTS

FIELD	GROUND TRUTH	PREDICTION
Seeks legal guidance	Yes	True · 100%
Primary topic	BUSINESS · also tagged: INTERNATIONAL_CROSS_BORDER	BUSINESS · 100%

Fig. 26. ModernBERT classifier demo for a legal-guidance conversation. The legal-guidance classifier predicts True and the topic classifier predicts BUSINESS, each with the displayed probability.

Conversation

What is the author?

The concept of the author is complicated and multifaceted. In traditional literary criticism, the author is often thought of as the individual who created the written work and who is responsible for its ideas and content. However, Michel Foucault, a poststructuralist philosopher, challenges this notion, suggesting that the author is not an individual figure but rather a social construction.

Foucault argues that the author is an effect of discourses and power relations that operate within society. He suggests that the author is not the origin or the source of meaning but rather a product of a system of norms, values, and institutions that regulate the production and circulation of texts.

Similarly, Roland Barthes, another poststructuralist philosopher, proposed that the author is not a "genius" who creates original works, but rather a figure who has absorbed various cultural and social influences and incorporated them into their writing.

In summary, while the traditional understanding of the author as a singular, individual figure who creates original works persists in some fields, poststructuralist and other critical theories complicate this view by emphasizing the role of broader social, cultural, and institutional factors in shaping the production and understanding of written works.

Next conversation Predict

RESULTS

FIELD	GROUND TRUTH	PREDICTION
Seeks legal guidance	No	False · 100%
Primary topic	(non-guidance)	(non-guidance) · 100%

Fig. 27. ModernBERT classifier demo for a non-guidance conversation. The classifier predicts False for legal guidance and suppresses topic routing because the conversation is non-guidance.