LEPREC: Reasoning as Classification over Structured Factors for Assessing Relevance of Legal Issues

Fanyu Wang[♠], **Xiaoxi Kang**[♢], **Paul Burgess**[♣], **Aashish Srivastava**[♣], **Chetan Arora**[♠], **Adnan Trakic**[♡], **Lay-Ki Soon**[♢], **Md Khalid Hossain**[♠], **Lizhen Qu**[♠*]

[♠] Faculty of Information Technology, Monash University, Australia
[♢] School of Information Technology, Monash University Malaysia
[♡] School of Business, Monash University Malaysia
[♣] Faculty of Law, Monash University, Australia
{firstname.lastname}@monash.edu

Abstract

More than half of the global population struggles to meet their civil justice needs due to limited legal resources. While Large Language Models (LLMs) have demonstrated impressive reasoning capabilities, significant challenges remain even at the foundational step of legal issue identification. To investigate LLMs' capabilities in this task, we constructed a dataset from 769 real-world Malaysian Contract Act court cases, using GPT-4o to extract facts and generate candidate legal issues, annotated by senior legal experts, which reveals a critical limitation: while LLMs generate diverse issue candidates, their precision remains inadequate (GPT-4o achieves only 62%). To address this gap, we propose LEPREC (Legal Professional-inspired Reasoning Elicitation and Classification), a neuro-symbolic framework combining neural generation with structured statistical reasoning. LEPREC consists of: (1) a **Neuro** component leverages LLMs to transform legal descriptions into question–answer pairs representing diverse analytical factors, and (2) a **Symbolic** component applies sparse linear models over these discrete features, learning explicit algebraic weights that identify the most informative reasoning factors. Unlike end-to-end neural approaches, LEPREC achieves interpretability through transparent feature weighting while maintaining data efficiency through correlation-based statistical classification. Experiments show a 30–40% improvement over advanced LLM baselines, including GPT-4o and Claude, confirming that correlation-based factor–issue analysis offers a more data-efficient solution for relevance decisions.

1 Introduction

Access to justice remains a global challenge, with over half of individuals worldwide unable to meet their civil justice needs (Gutiérrez Patiño et al., 2019). The shortage of legal expertise underscores the need for automated legal reasoning tools. Within the IRAC framework (Issue, Rule, Application, Conclusion) (Stockmeyer, 2021; Kang et al., 2023), legal issue identification, comprising both generating candidate issues and assessing their relevance, is the crucial first step, determining which legal questions arise from given facts. Given a set of facts, large language models (LLMs) show promise for *generating candidate legal issues* due to their strong language capabilities (Siino et al., 2025; Bernsohn et al., 2024), however, their assessment of *issue relevance* remains imprecise in real-world contexts (Schroeder and Lindholm, 2023; Magesh et al., 2025).

Assessing the relevance of generated issues requires both realistic legal scenarios and expert evaluation, yet such resources remain scarce. Existing benchmarks are limited to simplified or synthetic settings: Guha et al. (2023) avoid full legal case transcripts, while Kang et al. (2024) evaluate zero- and few-shot LLM performance on textbook-derived examples. The lack of real-world court case datasets hinders rigorous evaluation of models' ability to identify legally relevant issues, leaving a major gap in assessing their practical utility (Magesh et al., 2025).

To enable rigorous evaluation, we curate a Legal Issue dataset on Court cases, coined LIC, comprising 769 Malaysian Contract Act decisions. Herein, GPT-4o is used to extract facts and generate issue candidates, which are then annotated for relevance by senior lawyers, legal academics, and advanced law students. This is the first expert-validated, real-world benchmark for legal issue relevance assessment. Our analysis on this dataset shows that while LLMs generate diverse issue candidates, their precision remains low, simply prompting state-of-the-art LLMs achieves only 62% precision. The difficulty lies in reasoning beyond surface similarity among facts, as legal professionals assess relevance by considering multiple factors, including the core controversy and legal principles within the correct

*Corresponding author: lizhen.qu@monash.edu

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7701–7736
July 2-7, 2026

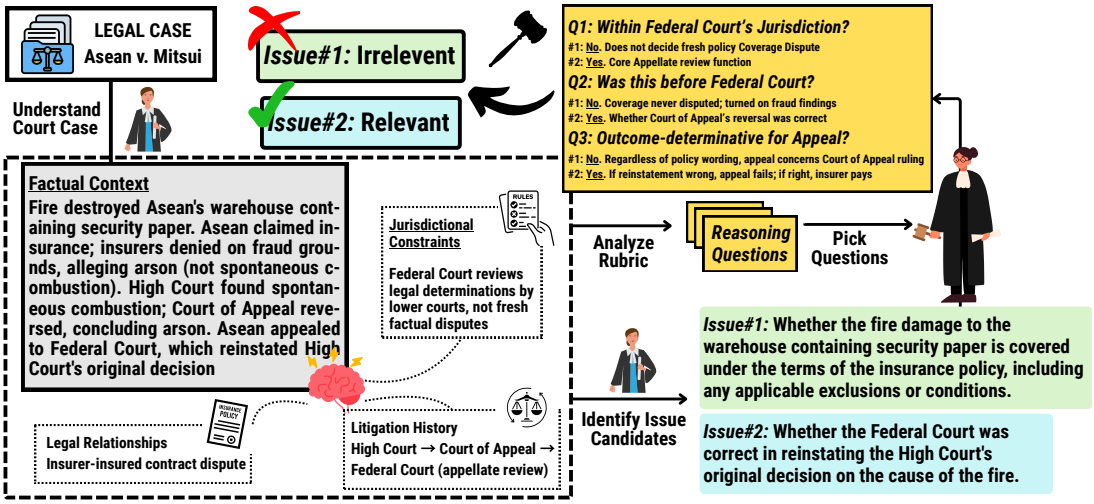

Figure 1: Legal issue relevance assessment requires understanding multiple contextual layers beyond surface-level fact matching. In this appeal, a candidate issue about policy coverage appears factually relevant but is **irrelevant** because the Federal Court (an appellate body) does not make fresh policy determinations. The truly **relevant** issue addresses what this specific court is being asked to decide: whether the lower court's ruling was legally correct. Lawyers employ reasoning questions that capture jurisdictional constraints, procedural context, and case-specific factors to distinguish relevant from irrelevant, indicating that a nuanced judgment requires expert legal knowledge.

jurisdictional and procedural context (Fig. 1).

Legal professionals approach such assessments through a natural two-stage process: i) they *identify key analytical factors* by brainstorming jurisdictional constraints, procedural context, and case-specific considerations; ii) they *weigh these factors* to reach a final relevance judgment. This decomposition, separating factor extraction from factor-based reasoning, mirrors the neuro-symbolic paradigm of combining neural comprehension with symbolic deliberation, motivating the design of our framework, named LEPREC (Legal Professional-inspired Reasoning Elicitation and Classification). It reframes the reasoning-based relevance assessment task as a neuro-symbolic framework combining neural factor extraction with statistical classification over symbolic features. LEPREC consists of a two-stage framework: i) the **neural component** leverages LLMs' language understanding to generate and answer rich reasoning Yes/No questions, transforming unstructured fact-issue descriptions into symbolic features capturing diverse analytical factors, and ii) the **symbolic component** employs sparse linear models to learn explicit algebraic weights over these discrete features, identifying the most informative reasoning cues for relevance assessment. This formulation transforms the task from evaluating fact–issue relationships to assessing factor–issue relevance, enabling similarity-based reasoning at an abstraction level, where similar factors yield similar relevance judgments. It is data-efficient because the number of model parameters is comparable to the size of the training data. Our contributions are threefold:

- We present LIC, the *first* large annotated dataset for legal issue relevance assessment, comprising 769 real-world court cases drawn from authentic judicial decisions.
- We propose LEPREC, a neuro-symbolic approach that mirrors legal professionals' analysis, transforming text-based legal reasoning into interpretable statistical classification over structured features by separating neural comprehension from symbolic deliberation.
- We conduct extensive experiments demonstrating that LEPREC achieves 30-40% improvement over cutting-edge LLM baselines. Ablation studies confirm the effectiveness of structured features and sparse linear models for relevance assessment. The interpretable statistical analysis of linear models offers deeper insight into how informative QA pairs contribute to the final relevance decisions.

2 The LIC Dataset

To enable rigorous evaluation of legal issue relevance assessment in real-world settings, we construct LIC, a corpus built from 769 Malaysian Contract Act court cases. Malaysia's Commonwealth legal system shares Commonwealth law founda-

tions with 50+ jurisdictions, including the UK, Australia, and Singapore (Greenleaf et al., 2013). Contract law exhibits universal reasoning principles, jurisdictional constraints, procedural posture, and factual relevance that transcend specific legal systems (Kazantsev, 2022). These characteristics suggest strong generalization ability in assessing legal understanding and judgment for the targeted legal analysis approaches. This section describes our dataset construction process, emphasizing the diversity of issue generation and the critical precision challenges revealed through expert annotation (The algorithm is illustrated in Appendix A.4).

Dataset Construction Overview. We collect 769 court cases from the *Current Law Journal (CLJ)*, focusing on *illegality under Section 24 of the Contracts Act Malaysia* and the *formation of contracts*. Starting with 243 Federal and High Court judgments, prioritized for their citation reputation, we expand the dataset by tracing cited cases within each judgment, yielding approximately 20 related cases per primary case. The dataset spans judgments from the 1990s to the present, capturing diverse legal scenarios and judicial writing styles.

We employ GPT-4o to automatically extract facts and legal issues from case PDFs, following best practices in prompt engineering (Wang et al., 2024; Lin et al., 2023). This process yields 5,690 issues and 7,397 facts (Prompt templates in Appendix A.1). We refer to these GPT-extracted issues as *silver truth issues* to distinguish them from the *ground truth labels*, i.e., binary relevance annotations (Relevant / Irrelevant) assigned by senior legal experts and used as the supervision signal for classification. A single datapoint consists of a (fact set, candidate issue) pair as input and a ground truth label as output; at inference time, the model observes only the fact set and candidate issue, and predicts its binary relevance label. To validate extraction quality, we engage a team of four annotators, including junior lawyers and high-performing law students, who evaluate randomly sampled outputs against original case documents using structured criteria. *65.1% of outputs achieve "High Distinction" ratings and 30.2% receive "Pass" ratings, with facts demonstrating the highest inter-annotator agreement* (Detailed analysis in Appendix A.2).

Incremental Issue Generation for Diversity. While extracted issues from case documents are reliable, legal disputes often admit multiple valid interpretations. A single case may yield different perspectives depending on how issues are framed, making diverse issue generation essential. Moreover, negative (non-applicable) issues play a crucial role in training, as even correctly extracted issues are not uniquely determined by facts.

Inspired by the principle that LLMs perform best with sufficient and necessary information (Feng et al., 2023), we propose an *incremental generation strategy* to produce diverse issue candidates while avoiding spurious correlations. Given a fact list $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, we generate incrementally:

1. Generate issues $\hat{\mathcal{Y}}_1$ using only $\mathbf{x}_1$, initializing $\hat{\mathcal{Y}} = \hat{\mathcal{Y}}_1$.
2. Generate issues $\hat{\mathcal{Y}}_2$ using $[\mathbf{x}_1, \mathbf{x}_2]$, updating $\hat{\mathcal{Y}} = \hat{\mathcal{Y}} \cup \hat{\mathcal{Y}}_2$.
3. Repeat by incrementally adding facts until all are used: $\hat{\mathcal{Y}} = \bigcup_{i=1}^{m} \hat{\mathcal{Y}}_i$.

By varying the "depth" of context, this strategy encourages the LLM to attend to progressively richer fact combinations, uncovering nuanced issue candidates that single-pass generation might miss. We craft prompts using GPT-4o and refine them with Claude (see Appendix A.10).

Accordingly, we compare against a *baseline* that feeds all m facts to the LLM in one pass with m times sampling, producing $\hat{\mathcal{Y}}_m \subseteq \hat{\mathcal{Y}} = \bigcup_{i=1}^{m} \hat{\mathcal{Y}}_i$. We evaluate both **quality** (whether generated issues semantically cover silver truth issues) and **diversity** (whether candidates differ meaningfully from each other). We employ five metrics: Fréchet BERT Distance (FBD) and BERT Embedding Distance (EMBD) (Alihosseini et al., 2019) for quality, and Self-EMBD, Self-BLEU, Distinct-N (Zhu et al., 2018; Li et al., 2015) for diversity. Table 1 shows that our incremental method substantially outperforms the baseline, achieving better coverage of silver truth issues while generating more diverse candidates (Detailed in Appendix A.7).

Expert Annotation of Issue Relevance. Legal issue relevance assessment is inherently subjective, with experts often reaching divergent conclusions. To establish consistent annotation standards, we convened multiple discussion rounds with three legal scholars, each with over 15 years of courtroom or academic experience, resulting rubric centers on a core principle: *an issue is deemed relevant only if it directly addresses the main dispute or core facts of the case, not merely if it relates to the scenario*, which ensures annotators capture precise

Table 1: Results of Issue Generation Strategies

Methods	Quality		Diversity		
	FBD	EMBD	Self-EMBD	Self-BLEU (n=3,4,5)	Distinct-N (n=3,4,5)
Baseline	1311	1354	211.8	11.85/14.52/16.64	8250/8559/8604
Ours	**1177**	**1227**	**225.1**	**24.90/30.36/34.95**	**11341/11790/11766**

Table 2: Statistics of Issue-Facts Pairs LIC Dataset

Type	Total	Gold	Rele.	Irrele.
LIC_U	3,903	752	–	–
LIC_L	1,188	213	607	368

factual-legal linkages rather than restating background principles. All the annotation processes are under regulation, with detailed explanation in Appendix A.3. We recruited three annotators with strong backgrounds in Commonwealth law jurisdictions through formal interviews, led by a senior solicitor with seven years of practice. Each annotator independently evaluated issue-fact pair relevance in the test set. Inter-annotator agreement measured by Fleiss' κ is 0.659, with pairwise Cohen's κ scores of 0.647, 0.746, and 0.584, indicating substantial agreement despite the task's inherent subjectivity (Gwet, 2008; McHugh, 2012).

Dataset Statistics and LLM Precision Analysis. The LIC corpus comprises two subsets: a large unlabeled collection (LIC_U), consisting of extracted issue and fact pairs from court judgments treated as highly relevant by construction, and a rigorously annotated test set (LIC_L), where each (fact set, candidate issue) pair carries an expert ground truth label. Table 2 summarizes key statistics.

Through expert annotation, we reveal a critical limitation in current LLM approaches to issue relevance assessment. While our incremental generation strategy successfully produces diverse issue candidates, the *precision of LLM-generated issues remains inadequate*. When we apply state-of-the-art models, including GPT-4o, Claude, and others to identify relevant issues from the generated candidates, they achieve only 62.26% precision on LIC_L. This low precision stems from LLMs' inability to distinguish between problems that are merely fact-related and those that truly address the case's core controversy, a nuanced judgment requiring deep legal expertise. Even closely fact-related issues may be irrelevant if they fail to address the dispute's central controversy, as demonstrated in Fig. 1. Unless noted otherwise, all evaluations in this study are performed on LIC_L, while the benefits of leveraging LIC_U are explored in Sec. 3.

Algorithm 1: LEPREC Framework

Input: A (fact set, candidate issue) pair $\langle \mathbf{X}, \hat{Y}_j \rangle$

Output: Binary relevance label $y_j \in \{\text{Relevant}, \text{Irrelevant}\}$

`// Neural Component: Question Generation`

For each pair $\langle \mathbf{X}_i, \hat{Y}_j \rangle$ in LIC_U, apply LLM to generate binary reasoning questions $\mathcal{Q}_{i,j} = \{q_1, \ldots, q_h\}$;

Accumulate into shared question pool $\mathcal{Q} = \bigcup_{i,j} \mathcal{Q}_{i,j}$, where $h = |\mathcal{Q}| = 2{,}486$;

Note: $\mathcal{Q}$ is generated from LIC_U *and shared across all cases*;

`// Neural Component: Question Answering`

For each question $q_t \in \mathcal{Q}$, apply generative verifier G to compute answer probability $G_{q_t}(\mathbf{X}, \hat{Y}_j) \in (0, 1)$;

Collect scores into feature vector $\mathbf{f} = G_{\mathcal{Q}}(\mathbf{X}, \hat{Y}_j) \in \mathbb{R}^h$;

`// Symbolic Component: Linear Classification`

Predict relevance label via learned linear model: $\hat{y}_j = \text{sign}(\mathbf{w}^\top \mathbf{f})$;

return $\hat{y}_j \in \{\textit{Relevant}, \textit{Irrelevant}\}$

3 Neuro-Symbolic Framework for Assessing Relevance of Legal Issues

We operationalize legal professionals' factor-based reasoning through a neuro-symbolic framework: neural question generation transforms unstructured legal text into symbolic features, followed by a linear classifier for interpretable relevance assessment over a structured feature space.

Problem Formulation. Given a fact set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and an issue candidate $\hat{\mathcal{Y}}$, our goal is to predict a binary relevance label indicating whether $\hat{\mathcal{Y}}$ is legally germane to $\mathbf{X}$.

Opening Challenges. The annotation results in Sec.2 indicate that LLMs struggle with nuanced judgments regarding relevance in identifying legal

issues. Our preliminary results in Sec.4 demonstrate that LLMs not only have difficulty generating relevant issue candidates but also identifying them accurately: Claude and GPT-4o achieve only ≈55% and ≈58% F1-Score, respectively. These shortfalls underscore a central challenge: *the conventional "black-box" strategy of mapping inputs directly to outputs falls short, whereas neuro-symbolic methods that mirror legal professionals' step-wise reasoning hold greater promise.*

Methodology Inspiration. By consulting legal experts, we observe their analysis follows a two-stage process with three operations: i) generating comprehensive analytical questions or rubrics (*factor extraction*), ii) selecting a focused subset pertinent to case context (*factor selection*), and iii) making decisions based on selected questions (*factor weighting*), where we provide a case study of LEPREC in Appendix A.5.1 aligned with Fig.1.

This naturally embodies neuro-symbolic decomposition: operation (i) requires language understanding to extract structured factors from unstructured text (*neuro*), while operations (ii)-(iii), which select informative factors and learn their weights, constitute symbolic deliberation (*symbolic*) through explicit algebraic operations (L1 regularization for selection, linear weighting for assessment), distinguishing them from neural black-box processing.

***Neural* Generation of Reasoning Questions.** Legal professionals examine issues by eliciting diverse reasoning questions representing different rubrics and concerns. We operationalize this through LLMs, leveraging their language understanding to generate sets of binary questions on different issue-fact pairs, forming question pool $\mathcal{Q}$:

1. Apply the LLM to produce several contextualized questions $\mathcal{Q}_{i,j} = \{\mathbf{q}_1, ..., \mathbf{q}_h\}$ for pairs of facts $\mathbf{X}_i$ and corresponding legal issue $\mathcal{Y}_j$. Then update the question pool $\mathcal{Q} = \mathcal{Q} \cup \mathcal{Q}_{i,j}$.
2. Given a question $\mathbf{q}_t \in \mathcal{Q}$ and a pair $\langle \mathbf{X}_i, \mathcal{Y}_j \rangle$, generation method $\mathcal{G}$ outputs $\mathcal{G}_{\mathbf{q}_t}(\mathbf{X}_i, \mathcal{Y}_j) \in (0, 1)$. Collectively, these scores form an h-dimensional feature vector $\mathcal{G}_{\mathbf{Q}}$ for each data point, where $h = |\mathcal{Q}|$.

We generate questions from LIC_U (since extracted issues in LIC_L are insufficient), resulting in 2,486 questions. For the generation method $\mathcal{G}(\cdot)$, we adopt probability-based Generative Verifier (Zhang et al., 2024) rather than direct binary LLM responses, as preliminary results indicate direct answers are unreliable (Detailed in Sec. 4).

Although only GPT-4O is used, question generation is a model-agnostic process that can be applied to different LLMs, as subsequent sparse feature selection automatically retains only the most predictive factors with observable weights, thereby significantly reducing the model-specific noise. This neural stage transforms unstructured legal text into structured features, creating a symbolic-like representation where each dimension corresponds to one analytical factor.

Correlation-Aware *Symbolic* Prediction. The question generation creates a large pool of diagnostic questions covering a wide range of analytical perspectives. However, using all questions directly presents two challenges. **Challenge 1**, semantically similar questions are expected to collaboratively yield consistent results. However, due to LLM unreliability, similar questions often produce conflicting outcomes, turning potential collaboration into noise. **Challenge 2**, some questions are highly domain-specific, leading to noise when applied to unrelated fact-issue pairs. E.g., the question *"Is the issue central to the insurer's stated reason for denying the claim?"* pertains only to insurance disputes. We cannot simply discard these narrow questions, as doing so would eliminate crucial information when their domain is relevant.

We adopt linear models on symbolic features to demonstrate how explicit algebraic operations on analytical factors address the identified challenges. Linear models exemplify key symbolic properties, explicit weight coefficients and transparent algebraic combination, while offering practical advantages: competitive performance with data efficiency and interpretable analysis of which specific questions contribute to legal reasoning, how their weights reveal relative importance, and how they activate across different case contexts. Through learned coefficients, linear models implement correlation-based feature weighting where the optimization process distributes weights according to marginal contributions. The linear function $\mathbf{w}^\top \mathbf{f}$ learns a weighted combination of structured features, which addresses **Challenge 1** by automatically downweighting noisy or redundant features through learned coefficients, treating correlated features as collective signals. For **Challenge 2**, standard linear models employ adaptive weighting without removal: retaining all features with

non-zero weights ($w_j \neq 0$), learning coefficients that reflect marginal contributions. The linear combination enables implicit contextualization where domain-specific features amplify when relevant context is present and attenuate otherwise. In contrast, L1-regularized variants induce sparsity by setting $w_j = 0$ for selected features, eliminating them globally, creating tension with preserving domain-specific questions needed in specialized contexts. We empirically compare these strategies to evaluate their trade-offs in handling correlated features while maintaining symbolic interpretability.

4 Experiments

We systematically evaluate our correlation-aware framework through three research questions: **RQ1**: *How do state-of-the-art LLMs perform on legal issue relevance classification?* We benchmark cutting-edge LLMs on LIC to establish baseline performance and reveal fundamental limitations of direct LLM judgment approaches. **RQ2**: *How does our correlation-aware linear framework compare against baselines?* We evaluate LEPREC against alternatives, demonstrating that correlation-aware linear weighting outperforms sophisticated end-to-end LLM approaches. **RQ3**: *What is essential in relevance classification?* We investigate whether stable, privileged questions drive legal reasoning through stability analysis of feature selection methods across multiple configurations, and verify our findings through an interview-based analysis with legal practitioners.

Experimental Setup. We report the mean and standard deviation of accuracy, macro-F1, precision, and recall from stratified 5-fold cross-validation. We use LIC_L with outer testset and divide the train and validation set into 0.7 and 0.3. Hyperparameter details appear in Appendix A.12.

RQ1: SOTA Methods on LIC. We evaluate two types of methods: reward models and LLMs as judges. Detailed settings and prompt templates appear in Appendix A.12 and A.13.

SOTA Baselines: **Reward Models**: (i) Generative verifier (Zhang et al., 2024), simplifying reward tasks to next token prediction on "yes" and "no" tokens ("Gen"); (ii) Prometheus (Kim et al., 2024), an open-source evaluation model using absolute rewarding mode ("Prom"); (iii) BERT-based Classifier (Chalkidis et al., 2020), a legal-specialized pre-trained LM with binary classifier head ("LBERT"). We initialize models with GPT-oss-20B ("oss"), Microsoft Phi-4 ("Phi"), and Qwen3-14B ("Qwen"). **Large Language Models**: Following LLMs-as-judge (Zheng et al., 2023), we evaluate GPT-4o and Claude3.5 ("GPT4o" and "Claude") (Prompt in Appendix A.13). Besides, we also use GPT-oss-20B, Microsoft Phi-4 and Qwen3-14B as LLMs as judges.

Table 3: Comparison of Current LLMs on LIC (RQ1)

Methods	F1	Acc.	Prec.	Rec.
Claude	$54.55_{\pm 1.85}$	$70.91_{\pm 1.16}$	$66.00_{\pm 3.48}$	$56.19_{\pm 1.36}$
GPT4o	$57.80_{\pm 1.98}$	$70.91_{\pm 1.16}$	$64.46_{\pm 2.52}$	$58.07_{\pm 1.56}$
GPT-OSS	$40.51_{\pm 0.11}$	$68.10_{\pm 0.30}$	$34.37_{\pm 0.01}$	$49.33_{\pm 0.23}$
Phi-4	$54.03_{\pm 2.65}$	$58.59_{\pm 2.60}$	$54.12_{\pm 2.50}$	$54.50_{\pm 2.74}$
Qwen3	$55.33_{\pm 2.62}$	$71.63_{\pm 1.31}$	$68.73_{\pm 4.39}$	$56.91_{\pm 1.81}$
LBERT	$52.31_{\pm 13.4}$	$41.28_{\pm 7.91}$	$52.10_{\pm 5.52}$	$50.79_{\pm 2.14}$
Prom	$48.63_{\pm 1.74}$	$62.76_{\pm 1.31}$	$50.05_{\pm 2.24}$	$44.96_{\pm 1.81}$
Gen_{oss}	$45.15_{\pm 2.47}$	$51.96_{\pm 5.31}$	$44.73_{\pm 1.84}$	$50.03_{\pm 1.45}$
Gen_{Phi}	$54.03_{\pm 2.65}$	$58.59_{\pm 2.60}$	$54.12_{\pm 2.50}$	$54.50_{\pm 2.74}$
Gen_{Qwen}	$63.70_{\pm 3.15}$	$68.59_{\pm 3.34}$	$63.84_{\pm 3.30}$	$63.92_{\pm 3.03}$

RQ1 Results: Table 3 shows Gen_{Qwen} achieves the highest F_1 (63.70%), followed by GPT-4o (57.80%) and Qwen3 (55.33%). The generative verifier framework's performance varies substantially with backbone model quality, as Gen_{Qwen} outperforms Gen_{Phi} (54.03%) and Gen_{oss} (45.15%). LegalBERT exhibits high variance ($\text{F}_1 = 52.31_{\pm 13.4}$) due to its thirst for training data. Critically, even the best-performing method achieves only 63.70% F_1, indicating they struggle in understanding the boundary of relevance in legal issue identification.

> ***Answer to RQ1:*** *Current LLMs cannot precisely judge the relevance of issue candidates.*

RQ2: LEPREC on LIC. We systematically evaluate alternatives at each key step and present results in Table 4[1]. **RQ2-1**: *How does reasoning question generation improve relevant issue identification?* We compare classification methods with SOTA LLMs to validate reasoning question generation efficacy. **RQ2-2**: *How do different selection methods handle legal reasoning questions?* By comparing with RQ2-1, we investigate the L1 regression and LLM-based selection methods to explore whether the feature selection can maintain a smaller size with comparable performance. RQ2 primarily focuses on the research-level performance. We further provide a discussion from the deployment-level in Appendix A.6.

Alternative Methods: From RQ1, three local LLMs (GPT-oss-20B, Microsoft Phi-4, Qwen3-14B) achieve close or better performance than com-

[1]For each method, we run results on three feature generation models and report their best performance. Check Appendix A.14 for all results.

Table 4: The Bests of the Alternatives (RQ2)

Methods	F1	Acc.	Prec.	Rec.
RQ2-1: Question Generation				
SVC_{Phi}	$\mathbf{80.19}_{\pm 2.83}$	$82.66_{\pm 2.38}$	$79.67_{\pm 2.70}$	$\mathbf{81.01}_{\pm 3.13}$
LR_{Phi}	$79.70_{\pm 2.93}$	$82.49_{\pm 2.41}$	$79.58_{\pm 2.89}$	$80.05_{\pm 3.23}$
$Ridge_{Phi}$	$80.10_{\pm 2.86}$	$82.91_{\pm 2.41}$	$80.06_{\pm 2.89}$	$80.28_{\pm 3.05}$
KNN_{Qwen}	$74.53_{\pm 2.06}$	$79.12_{\pm 1.81}$	$76.06_{\pm 2.61}$	$73.66_{\pm 2.01}$
RF_{Qwen}	$74.45_{\pm 2.94}$	$79.63_{\pm 2.74}$	$77.52_{\pm 4.56}$	$73.04_{\pm 2.42}$
DT_{Qwen}	$67.81_{\pm 4.88}$	$71.79_{\pm 5.57}$	$68.72_{\pm 4.76}$	$68.67_{\pm 4.08}$
NB_{Qwen}	$65.53_{\pm 4.02}$	$69.61_{\pm 3.86}$	$65.49_{\pm 3.86}$	$66.24_{\pm 3.93}$
SGD_{Qwen}	$75.24_{\pm 2.94}$	$79.37_{\pm 3.08}$	$76.65_{\pm 4.39}$	$74.58_{\pm 2.28}$
LDA_{Phi}	$79.56_{\pm 4.01}$	$83.50_{\pm 2.69}$	$81.77_{\pm 2.97}$	$78.39_{\pm 4.30}$
FFN_{Qwen}	$75.65_{\pm 2.57}$	$79.29_{\pm 2.43}$	$76.10_{\pm 2.84}$	$75.57_{\pm 2.47}$
CNN_{Qwen}	$43.33_{\pm 3.08}$	$69.78_{\pm 0.96}$	$54.78_{\pm 24.83}$	$51.22_{\pm 1.51}$
$Tranf._{Qwen}$	$75.44_{\pm 1.82}$	$80.14_{\pm 2.02}$	$78.22_{\pm 3.64}$	$74.39_{\pm 2.10}$
$LSTM_{Qwen}$	$63.48_{\pm 2.94}$	$65.90_{\pm 3.79}$	$64.04_{\pm 1.93}$	$65.81_{\pm 1.84}$
$ResNet_{Phi}$	$67.33_{\pm 3.61}$	$72.48_{\pm 4.19}$	$68.84_{\pm 4.89}$	$67.33_{\pm 3.40}$
RQ2-2: Question Selection				
$L1Reg_{Phi}$	$80.01_{\pm 3.61}$	$\mathbf{83.34}_{\pm 3.06}$	$\mathbf{81.13}_{\pm 3.93}$	$79.32_{\pm 3.45}$
$L1SVC_{Phi}$	$77.60_{\pm 2.58}$	$80.89_{\pm 1.97}$	$77.68_{\pm 2.23}$	$77.62_{\pm 2.93}$
$LMSel._{SVC}$	$57.78_{\pm 1.57}$	$66.33_{\pm 0.82}$	$58.90_{\pm 1.20}$	$57.64_{\pm 1.52}$
$LMSel._{L}$	$45.63_{\pm 1.79}$	$50.25_{\pm 1.72}$	$46.27_{\pm 1.76}$	$45.84_{\pm 1.92}$
$LMSel._{P}^{LR}$	$74.86_{\pm 2.85}$	$77.44_{\pm 2.71}$	$74.24_{\pm 2.71}$	$76.32_{\pm 2.85}$
GCI	$64.32_{\pm 5.29}$	$66.50_{\pm 7.18}$	$69.43_{\pm 8.96}$	$65.52_{\pm 3.53}$
GCI_{Chain}	$64.07_{\pm 5.16}$	$69.20_{\pm 2.38}$	$72.55_{\pm 6.11}$	$65.32_{\pm 4.10}$
GCI_{Cons}	$67.63_{\pm 8.09}$	$74.19_{\pm 5.05}$	$80.72_{\pm 5.62}$	$68.11_{\pm 6.37}$

mercial LLMs. We adopt them as $\mathcal{G}(\cdot)$ for generating features, yielding three alternatives. **RQ2-1 Alternatives.** (1) LLM output types: continuous scores (default) or binary labels (underlined). (2) Classifier families: *Embedding-based* (FFN, CNN, LSTM, Transformer, ResNet) and *Traditional ML* (SVM, LR, KNN, RF, GB, Ridge, SGD, DT), representing approaches with different inductive biases. **RQ2-2 Alternatives.** (1) *L1-regularized methods* (LR_{L1}, SVC_{L1}) perform explicit feature selection through sparsity penalties. (2) *LLM-Select (LMSel.)* (Jeong et al., 2024): SOTA feature selection leveraging LLMs with standard binary labels (LMSel.), improved variant using probabilities ($LMSel._{P}$), and LLM-based selective classification ($LMSel._{L}$) (Please find the implementation details in Appendix A.12). Moreover, CGI (Liu et al., 2021) is a neuro-symbolic approach that combines causal discovery algorithms with neural networks for legal charge disambiguation. We evaluate three variants: the standard GCI based on causal discovery; GCI_{Chain} (CausalChain) leverages recurrent neural networks on extracted causal chains; and GCI_{Cons} (Bi-LSTM+Att+Cons) constrains neural attention using causal strength estimates.

RQ2-1 Results: Results validate our framework through clear performance stratification. **Linear models** (Linear SVC, Standard LR, Ridge: 79.70–80.19% F1) achieve remarkably consistent performance, demonstrating that simple linear weighting addresses **Challenge 1** with data efficiency and interpretability. Near-identical results across variants validate that linear combination itself, not particular regularization, drives performance. **Tree/distance methods** (RF, KNN, DT: 67–75% F1) and **deep learning** show competitive but slightly lower or more variable performance. Feature generator patterns: Phi-4 pairs well with linear models (all ≈80%), while Qwen excels with tree-based and deep methods, aligning with RQ1 where Gen_{Qwen} (63.70%) outperformed Gen_{Phi} (54.03%). Notably, continuous probability scores uniformly outperform binary variants, confirming probabilistic information is critical.

RQ2-2 Results: Feature selection results reveal the challenge of identifying truly important questions. **LLM-Select** methods fail dramatically, revealing fundamental misalignment: LLMs prioritize questions based on factual salience rather than empirical predictiveness. This bottleneck from RQ1 propagates through selection, demonstrating LLMs cannot reliably identify which questions lawyers find predictive. **GCI series**, as a hybrid model, even improves over standard causal inference, but still significantly underperforms L1-based methods. As GCI's strict causal discovery overly restricts the feature space, sacrificing significant recall, this confirms that legal relevance relies on a broader set of correlational signals that strict causal graphs may discard, further validating the efficiency of correlation-based sparse selection.

In contrast, **L1-based methods** achieve competitive performance through selection, while **$L1SVC_{Phi}$** (77.60%) drops only 2.5 points. This near-parity initially appears to contradict **Challenge 2**, but closer examination reveals that L1 methods eliminate questions based on correlation, including domain-specific questions that contribute to specialized cases, whereas linear models reduce weights rather than eliminate them. Critically, L1's competitive performance raises an intriguing possibility: if sparsity-based selection identifies predictive subsets without relying on LLM judgment, the selected questions and their learned weights may reveal which aspects of legal reasoning drive relevance classification, motivating a deeper investigation into what linear models learn: which questions receive high weights, how weights relate to legal doctrine, and whether coefficient patterns align with human legal reasoning priorities.

Answer to RQ2: *For relevant legal issue identification, • simple linear models with correlation-aware weighting achieve competitive performance (79–80% F1) while enabling interpretable analysis; • LLMs cannot reliably identify predictive questions through explicit selection; • adaptive weighting outperforms elimination: retention with*

context-dependent weights preserves domain-specific reasoning.

RQ3: What is Essential in Relevance Classification. L1's competitive aggregate performance in RQ2-2 (L1 LR: 80.01%, L1 SVC: 77.60%) despite eliminating features raises a critical question: do sparsity-inducing methods identify a stable core of essential reasoning questions? We investigate this through two complementary analyses: a **quantitative stability analysis** of feature selection methods, and a **qualitative practitioner study** that grounds our findings in real legal reasoning. Together, they reveal that the absence of a universal question set is not a limitation of our framework, but a fundamental characteristic of legal issue relevance classification itself: even experienced lawyers do not reason from a fixed checklist, but from a broad, context-sensitive coverage of analytical factors.

Quantitative: Extreme Selection Instability: We conduct stability analysis across multiple threshold configurations using 100-iteration bootstrap subsampling within 5-fold CV to investigate whether L1-based selection reveals consistent question subsets. Results reveal no universal essential question set exists, but validate that multiple question subsets can effectively capture legal reasoning (See details in Appendix B).

L1 Logistic Regression exhibits extreme instability: only 0.04–0.53% of features are consistently selected across all 5 folds (Table 5). Critically, these L1-selected features receive near-zero coefficients in Standard LR (e.g., f_{1914}: β=0.061, p=0.9998), proving L1 selects arbitrary representatives from correlated clusters rather than genuinely important questions. Hyperparameter sensitivity causes 8–10$\times$ variation in feature counts across folds, yet each fold's different subset achieves reasonable classification, demonstrating many question combinations work equivalently. **L1 SVC** shows different behavior: 32$\times$ more consistently selected features than L1 LR. Critically, only 38% overlap between L1 LR and L1 SVC selections proves different methods identify different "important" features, yet both achieve similar reasoning capability. This demonstrates that legal reasoning does not depend on privileged questions but on covering sufficient reasoning dimensions.

Three findings explain why no universal essential questions exist: (1) *Extreme instability*: only 0.04–6.4% consistently selected across methods/thresholds, proving multiple subsets capture legal reasoning equivalently. (2) *Method disagreement*: 62% non-overlap between L1 LR and L1 SVC selections, yet both function effectively. (3) *Subset equivalence*: 8–11$\times$ variation in feature counts produces similar reasoning capability, validating massive redundancy in our 2,464 questions. Our contextualized question generation ($\approx$8 per case-issue pair) explains this: questions target specific legal scenarios rather than a universal reasoning bank. While no single golden question set exists, domain-specific questions are essential components of effective subsets. L1 methods arbitrarily eliminate such questions when they appear noisy globally, but these questions contribute valuable signal in specialized contexts. This is why adaptive weighting outperforms elimination-based selection, echoing **Challenge 2**: effective legal reasoning requires comprehensive coverage where domain-specific questions activate conditionally rather than fixed selection of privileged questions.

Qualitative: Practitioner Validation: Practitioner feedback from our human usability study (Appendix A.5.2) confirms that the absence of a universal question set reflects how legal reasoning actually works. We picked a court case from LIC and invited two legal practitioners with substantial experience in contract law: a lawyer with over 10 years of experience in civil and commercial litigation from China, and a law professor from Australia. These two legal practitioners have extensive practical experience and are from different countries outside Malaysia, thereby demonstrating the generalizability of LIC and LEPREC in other countries. Despite this, their feedback converges on two key messages that directly align with our quantitative findings. For further details on the study design and full interview transcripts, see Appendix A.5.2.

(i) Lawyers do not reason from a fixed checklist. Both practitioners confirmed that legal reasoning in practice is context-sensitive and does not follow a prescribed sequence of questions. Lawyer #1 noted that not every reasoning question reflects a step they would naturally take in every case. Lawyer #2 echoed this, observing that the process is "much more intuitive than considering a list of particular questions in turn" and that the relevance of individual questions depends heavily on the nature of the matter. This is consistent with our quantitative finding that no stable universal subset exists, and points to why legal issue relevance classification is fundamentally difficult: relevance judgment re-

Table 5: L1 Method Selection Stability Comparison

Method	Thres.	Always (5/5)	Stable (≥4/5)	Moderate (3/5)	Unstable (1-2/5)	Never (0/5)	F1 Score
L1 LR	0.3	13 (0.53%)	43 (1.75%)	71 (2.88%)	519 (21%)	1831 (74%)	73.60±2.12
	0.4	5 (0.20%)	20 (0.81%)	24 (0.97%)	326 (13%)	2094 (85%)	72.24±2.17
	0.5	2 (0.08%)	6 (0.24%)	18 (0.73%)	193 (7.8%)	2247 (91%)	72.94±5.31
	0.6	1 (0.04%)	3 (0.12%)	6 (0.24%)	106 (4.3%)	2349 (95%)	69.89±4.64
L1 SVC	0.3	386 (15.7%)	766 (31.1%)	475 (19.3%)	1222 (50%)	1 (0.04%)	76.94±1.46
	0.4	158 (6.4%)	328 (13.3%)	327 (13.3%)	1729 (70%)	80 (3.2%)	74.61±2.66
	0.5	64 (2.6%)	153 (6.2%)	179 (7.3%)	1516 (62%)	616 (25%)	74.41±2.03
	0.6	24 (1.0%)	61 (2.5%)	99 (4.0%)	928 (38%)	1376 (56%)	75.06±3.23

quires assembling a context-sensitive combination of factors that varies across cases.

(ii) Comprehensive coverage is the value, not prescription. Both practitioners independently identified the same utility in the question list: not as a checklist to follow step by step, but as a resource that ensures comprehensive coverage of considerations that might otherwise be overlooked. Lawyer #1 observed that questions they would not spontaneously consider still helped surface case-specific factors, while Lawyer #2 noted that the factors provided "relatively comprehensive coverage" of key considerations. This directly validates our design choice of retaining a full question pool with adaptive weighting rather than imposing L1 elimination, which discards precisely the domain-specific questions that contribute conditionally.

Answer to RQ3: *The stability and usability results reveal that - What is essential is not a fixed set of privileged reasoning questions, but comprehensive coverage with adaptive weighting. No universal essential question set exists, and this reflects how legal reasoning actually works: lawyers do not reason from a fixed checklist but assemble context-sensitive combinations of factors, which is also what makes relevance classification fundamentally difficult.*

5 Related Work

Feature Selection and Classification. Feature selection and classification methods include filter, wrapper, and embedded techniques (Gramegna and Giudici, 2022; Alsolami and Fukai, 2022), with recent advances using deep learning, reinforcement learning, and evolutionary approaches (Jia et al., 2024; Nguyen et al., 2024). LLM-Lasso uses large language models to guide feature selection and enhance robustness by integrating domain knowledge into Lasso regression (Zhang et al., 2025). Focus Instruction Tuning enables dynamic control over feature reliance in LLMs through natural-language prompts, improving robustness and fairness (Lamb et al., 2024). While traditional classifiers like SVMs and ensemble methods remain widely used (Bulut, 2022; Majdoubi et al., 2023), optimal selection still depends on empirical evaluation tailored to the dataset (Loyola-Fuentes et al., 2022; Montañana et al., 2024).

LLMs in the Legal Domain. Applying LLMs to legal tasks is challenging due to the complexity of legal knowledge. Studies indicate that current models often capture only surface-level concepts (Savelka et al., 2023), miss crucial legal rule details (Yuan et al., 2024), and struggle to identify important legal factors (Gray et al., 2024). These findings underscore the need for further development before LLMs can function autonomously in legal contexts.

6 Conclusion

We evaluate relevance assessment of legal issues by introducing LIC, the first expert-annotated dataset from 769 Malaysian Contract Act cases. To address the low precision problem of LLMs and the scarcity of training data, we propose LEPREC, a legal reasoning-inspired neuro-symbolic approach that transforms text-based reasoning into statistical classification over structured factors, followed by relevance prediction using a sparse linear model, which also supports statistical analysis between factors and issues. As a result, LEPREC achieves 30-40% improvement over end-to-end LLM approaches, with linear models substantially outperforming both deep learning methods and direct LLM judgments. Our stability and usability analysis reveals the conclusion from quantitative and qualitative perspectives that no universal essential question set exists in the practical legal reasoning process. Instead, effective performance emerges from comprehensive coverage with adaptive weighting. Our work lays the foundation for theoretically grounded, robust, and interpretable legal AI systems.

7 Limitations

The dataset comprises 769 cases with expert annotations, which naturally reflects the subjective understanding involved in legal issue relevance assessment. Different legal experts may reasonably interpret issue relevance differently based on their experience and perspective.

Our study focuses on Malaysian Contract Act cases from the Commonwealth legal system. While contract law represents a fundamental legal institution with universal reasoning principles (jurisdictional constraints, procedural context, factual centrality), and LEPREC's methodology is jurisdiction-agnostic (learning correlation patterns rather than encoding specific doctrines), empirical validation on additional jurisdictions would further establish cross-system robustness. The Commonwealth legal heritage shared across 50+ nations suggests strong transferability, but civil law systems may require investigation of whether reasoning pattern differences affect framework performance.

Our dataset is constructed from published judicial opinions, which present facts and issues as refined by judges; testing on party submissions (pleadings, briefs) would be valuable as they are more fact-rich and mirror real-world practice, though such materials are typically not publicly available and would require partnerships with law firms and extensive anonymization procedures.

Our framework relies on LLM-generated reasoning questions, and while our contextualized generation strategy produces sufficient coverage, exploring alternative question elicitation approaches could provide additional insights. We employ linear models for interpretability and efficiency, which assume linear combinations can capture relevance patterns. The learned weights across our comprehensive question set offer coefficient-level interpretability, though extracting high-level insights from detailed weight distributions requires careful analysis.

The annotation process involves senior lawyers and legal academics, reflecting the typical resource investment in building specialized legal AI datasets. While LEPREC demonstrates human expert-level performance, deployment in real-world legal practice would require additional validation to ensure the system does not introduce biases or disadvantage vulnerable populations.

8 Ethics Statement

We acknowledge and adhere to the ACL Code of Ethics throughout our research. All case data are from publicly available court records (Current Law Journal database) and have been carefully handled, with names, unique identifiers, and any potentially offensive language screened, anonymized, or redacted to protect privacy and avoid harmful content. The data collection protocol was approved by the Monash University Human Research Ethics Committee (MUHREC). Annotators were recruited via open advertisement targeting legal professionals with Commonwealth law backgrounds, with formal interviews conducted to ensure qualifications. All annotators provided informed consent after receiving detailed explanations of the annotation task, data usage, and their rights, and were compensated at rates meeting or exceeding local government-mandated standards, appropriate for their expertise level and jurisdiction. Payment terms, training procedures, and the right to withdraw were communicated clearly before participation. We employed AI assistants (GPT-4o: `gpt-4o-2024-05-13` and Claude 3.5 Sonnet: `claude-3-5-sonnet-20241022`) for fact and issue extraction, prompt refinement, and incremental generation, documenting all prompts, model versions, sampling parameters, and API settings in corresponding sections, while stressing that final outputs were reviewed and validated by human experts. To mitigate LLM-related risks (hallucination, bias, sensitive content), we applied targeted prompt engineering with few-shot exemplars, maintained human-in-the-loop review, and transparently documented usage. We acknowledge the environmental impact and encourage the community to balance performance gains with sustainability considerations.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Danial Alihosseini, Ehsan Montahaei, and Mahdieh So-

leymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98.

Ibrahim Alsolami and T. Fukai. 2022. An extension of fisher's criterion: Theoretical results with a neural network realization. *arXiv.org*.

Dor Bernsohn, Gil Semo, Yaron Vazana, Gila Hayat, Ben Hagag, Joel Niklaus, Rohit Saha, and Kyryl Truskovskyi. 2024. Legallens: Leveraging llms for legal violation identification in unstructured text. *arXiv preprint arXiv:2402.04335*.

Vahide Bulut. 2022. Classifying surface points based on developability using machine learning. *European Journal of Science and Technology*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Claude3.5. Claude3.5 official technical report.

CLJ Legal Network. 2024. CLJ Law – Current Law Journal Malaysia. Accessed: 2024-05-27.

Commonwealth Secretariat. n.d. Member countries.

Tao Feng, Lizhen Qu, and Gholamreza Haffari. 2023. Less is more: Mitigate spurious correlations for open-domain dialogue response generation models by causal discovery. *Transactions of the Association for Computational Linguistics*, 11:511–530.

Alex Gramegna and Paolo Giudici. 2022. Shapley feature selection. *FinTech*.

Morgan Gray, Jaromir Savelka, Wesley Oliver, and Kevin Ashley. 2024. Using llms to discover legal factors. In *Legal Knowledge and Information Systems*, pages 60–71. IOS Press.

Graham Greenleaf, Andrew Mowbray, and Philip Chung. 2013. Building a commons for the common law: the commonwealth legal information institute (commonlii) four years on. In *Legislative Drafting*, pages 117–124. Routledge.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Camilo Gutiérrez Patiño, Matthew Harman, Sarah Chamness Long, Jorge A. Morales, Ted Piccone, Alejandro Ponce, Natalia Rodríguez Cajamarca, Adriana Stephan, Kirssy González, and Jennifer VanRiper. 2019. Global insights on access to justice 2019. Technical report, World Justice Project.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Daniel P Jeong, Zachary C Lipton, and Pradeep Ravikumar. 2024. Llm-select: Feature selection with large language models. *arXiv preprint arXiv:2407.02694*.

Pengyue Jia, Yejing Wang, Zhaochen Du, Xiangyu Zhao, Yichao Wang, Bo Chen, Wanyu Wang, Huifeng Guo, and Ruiming Tang. 2024. Erase: Benchmarking feature selection methods for deep recommender systems. *Knowledge Discovery and Data Mining*.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Zhuang Li, and Adnan Trakic. 2024. Bridging law and data: Augmenting reasoning via a semi-structured dataset with irac methodology. *arXiv preprint arXiv:2406.13217*.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Zhuo, Patrick Emerton, and Genevieve Grant. 2023. Can chatgpt perform reasoning using the irac method in analyzing legal scenarios like a lawyer? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13900–13923.

Mikhail Fedorovich Kazantsev. 2022. The civilizational value of a contract. In *SHS Web of Conferences*, volume 134, page 00057. EDP Sciences.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models.

Tom A Lamb, Adam Davies, Alasdair Paren, Philip HS Torr, and Francesco Pinto. 2024. Focus on this, not that! steering llms with adaptive feature specification. *arXiv preprint arXiv:2410.22944*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. *arXiv preprint arXiv:2104.09420*.

J. Loyola-Fuentes, L. Pietrasanta, M. Marengo, and F. Coletti. 2022. Machine learning algorithms for flow pattern classification in pulsating heat pipes. *Energies*.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Oumaima Majdoubi, A. Benba, and A. Hammouch. 2023. Classification of parkinson's disease and other neurological disorders using voice features extraction and reduction techniques. *Informatyka Automatyka Pomiary w Gospodarce i Ochronie Środowiska*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Ricardo Montañana, Jos'e A. G'amez, and Jos'e M. Puerta. 2024. Odte - an ensemble of multi-class svm-based oblique decision trees. *Expert systems with applications*.

B. Nguyen, Bing Xue, and Mengjie Zhang. 2024. A constrained competitive swarm optimizer with an svm-based surrogate model for feature selection. *IEEE Transactions on Evolutionary Computation*.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525*.

Philipp Schroeder and Johan Lindholm. 2023. From one to many: Identifying issues in cjeu jurisprudence. *Journal of Law and Courts*, 11(1):163–186.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*.

Norman Otto Stockmeyer. 2021. Legal reasoning. *It's all about IRAC*.

Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *npj Digital Medicine*, 7(1):41.

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

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. 2024. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. In *Conference on Empirical Methods in Natural Language Processing*.

Erica Zhang, Ryunosuke Goto, Naomi Sagan, Jurik Mutter, Nick Phillips, Ash Alizadeh, Kangwook Lee, Jose Blanchet, Mert Pilanci, and Robert Tibshirani. 2025. Llm-lasso: A robust framework for domain-informed feature selection and regularization. *arXiv preprint arXiv:2502.10648*.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Appendix

A.1 Prompt Template of Facts and Issue Extraction

We use the following prompt for fact extraction.

```
You are a legal expert tasked with analyzing a
    court case.
Your goal is to extract the case name, summarize
     key legally
significant facts, and explain the court's final
     decision (held).

Instructions:
1. **Case Name**: Extract the full official case
     name.
   Example: Smith v. Jones [2020] 2 MLJ 35.
2. **Facts**: Identify the facts directly
    related to the legal
   issues. Focus on those that establish the
    dispute, actions,
   and agreements.
3. **Held (Conclusion)**: Provide the court's
    final decision,
   including penalties, remedies, or significant
     conclusions.

Output Format:
{
    "case_name": "Extracted case name",
    "facts": [
        "Fact 1...",
        "Fact 2..."
    ],
    "held": "Holding or judgment of the court."
}

Case Text:
{case_text}
```

Listing 1: Prompt for Fact and Held Extraction

The prompt below is used for issue extraction.

```
You are a legal expert analyzing a court case.
Your goal is to identify legal issues, apply
    relevant rules
to the facts, and provide legal conclusions.

Instructions:
1. Identify each legal issue in the case by
    framing a question
   starting with "Whether...".
2. For each issue, apply the relevant rules to
    the facts using
   an "if...then" structure.
3. Provide a clear answer (Yes/No or another
    legal conclusion)
   for each issue, based on legal reasoning.
4. Multiple applications may be required if more
     than one rule
   applies or if multi-step reasoning is
    necessary.

Output Format:
{
    "issues": [
        {
            "issue": "Whether issue 1...",
            "application": [
                "If [specific fact]... then [
    application of legal rule]...",
                "If [specific fact]... then [
    application of another legal rule]..."
            ],
            "answer": "Yes/No or detailed legal
    conclusion for issue 1..."
        },
        {
            "issue": "Whether issue 2...",
            "application": [
                "If [specific fact]... then [
    application of legal rule]..."
            ],
            "answer": "Yes/No or detailed legal
    conclusion for issue 2..."
        }
    ]
}

Example:
- Issue: "Whether the contract is enforceable
    under Section 24 of the Contracts Act."
- Application:
    - "If the contract is based on illegal
    consideration, then under Section 24, the
    contract is void."
    - "If no illegal consideration exists, then
    under the same section, the contract remains
     valid."
- Answer: "No, the contract is void due to
    illegal consideration."

Facts:
{facts}

Rules:
{rules}

Original Case Text:
{case_text}
```

Listing 2: Prompt for Issue Identification and Application

A.2 Dataset Quality Details

CLJ is a leading Malaysian legal publication providing case law reports, legal commentaries, and statutory updates, serving as a key reference for legal practitioners and researchers. Using predefined filtering criteria. We prioritize Federal and High Court judgments due to their higher citation reputation. Each case was sourced in its original PDF format, preserving the judicial text as delivered.

Starting with an initial set of 243 cases, we expand the dataset by tracing cited cases within each judgment. This citation-based expansion yields approximately 20 related cases per primary case, ultimately increasing the dataset to 769 cases. Spanning judgments from the 1990s to the present, the dataset encapsulates a diverse range of legal scenarios and judicial writing styles.

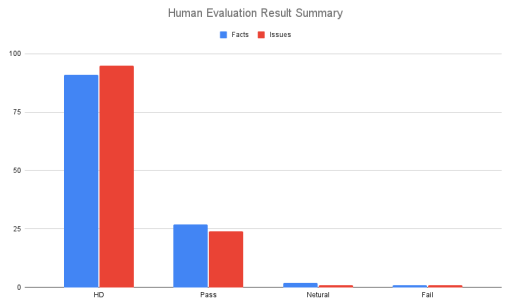

Figure 2: Evaluation results on fact and issue extraction.

Legal cases are lengthy, and it is expensive to manually extract facts and issues from those cases. Therefore, we reduce human effort by applying GPT-4o to extract facts and issues, and by annotating a set of generated issues manually by law students as the ground-truth (see prompt template in Appendix A.1). Herein, the automatically extracted issues are referred to as silver ground-truth. For both fact and issue extraction, we follow the best practice on prompting (Wang et al., 2024) and use the styles recommended in (Lin et al., 2023) for prompt design. Specifically, we apply GPT4o with those prompts to extract legally significant facts and legal issues from the PDF files of collected court cases. In the end, we got 5,690 issues and 7,397 facts. *Data Quality:* To check the quality of extracted content, we engage a team of four annotators, including junior lawyers and law students with strong academic records (B+ or higher in relevant legal subjects). Annotators evaluate outputs from randomly sampled cases by comparing them against original content, validating key elements manually. Using predefined criteria, they assign ratings ranging from "High Distinction (HD)", for highly accurate and detailed outputs, to "Fail", for outputs with significant omissions or irrelevance. The detailed guidelines are provided in Appendix A.11. Structured ratings and detailed comments are provided for each element to assess. As illustrated by Fig. 2, 65.1% of the model outputs are rated as HD and 30.2% as Pass, with facts achieving the highest annotator agreement. Only a small fraction of the facts and issues are categorized as Fail. While HD-rated outputs are ideal, Pass-rated outputs also hold significant value for tasks requiring basic reasoning or tolerating minor inaccuracies. Common errors in Pass-rated outputs include incomplete or poorly sequenced facts, insufficiently framed or misaligned issues. However, even within these outputs, relevant and accurate content often remains, which can be highly beneficial for model training processes. This makes Pass-rated outputs a valuable resource for enhancing dataset diversity and providing foundational reasoning.

A.3 Human Annotation

Explanation Annotator Recruitment and Payment. We recruited annotators through open advertisements and formal calls within our university community, ensuring a diverse pool of participants. Prior to commencing annotation, we obtained approval from our institutional ethics board. All annotators were compensated at or above local government–mandated rates, with payment schedules and amounts clearly communicated in advance. We also provided training sessions, ongoing support, and the opportunity to withdraw at any time without penalty.

Relevance Checking. Your task is, for each issue, to decide whether it is "Relevant" or "Irrelevant" to the scenario.

- "**Relevant**": The issue is not only related to the scenario, but is directly tied to the main dispute or core facts of the case. It goes beyond merely stating a basic legal principle or background fact.
- "**Irrelevant**": The issue is either unrelated to the scenario or is only a fundamental/basic statement that does not bear on the case's primary controversy.

A.4 Dataset Curation Algorithm

See the algorithm list in Algo.2.

A.5 Case Analysis and Usability Study

To validate LEPREC's interpretability from both a computational and a practitioner perspective, we present two complementary analyses based on a real court case. First, we present a case study and illustrate the reasoning process from LEPREC and legal practitioners. Second, we conduct a human usability study with legal practitioners from two distinct legal systems to assess whether the procedure of LEPREC is useful across jurisdictions. We present the court case following the structure in LIC, as:

Table 6: Example of the Annotation Answer Sheet. We will both give the extracted facts and ground truth legal issues to help the annotators better capture the core dispute of the case. Then, for each generated legal issue, the annotators need to select "Relevant" or "Irrelevant" and write several justifications for their decision.

Note	Relevance	Legal Issues
Asean Security Paper Mills Sdn Bhd v. Mitsui Sumitomo Insurance (Malaysia) Bhd [2008] 6 CLJ		
Facts: 1. Asean Security Paper Mills Sdn Bhd (the respondent/appellant) made a claim on an insurance policy after a fire destroyed their warehouse containing security paper. 2. The insurance companies, including Mitsui Sumitomo Insurance (the applicant/respondent), denied the claim on grounds of fraud, alleging that the fire was caused by arson rather than spontaneous combustion. 3. The High Court found that the fire was due to spontaneous combustion; this decision was overturned by the Court of Appeal, which concluded that the fire was an act of arson. 4. The respondent/appellant appealed to the Federal Court, leading to the reinstatement of the High Court's original decision. **Ground Truth Legal Issues:** - Whether the Federal Court should exercise its review jurisdiction under Rule 137 of the Rules of the Federal Court 1995 to prevent injustice in this case. - Whether the findings of fact made by the High Court and reinstated by the Federal Court can be subject to review. - Whether the application for review by Mitsui Sumitomo Insurance was an appropriate use of inherent jurisdiction.		
	`Irrelevant`	Whether the fire damage to the warehouse containing security paper is covered under the terms of the insurance policy, including any applicable exclusions or conditions.
	`Irrelevant`	Whether Asean Security Paper Mills Sdn Bhd complied with all obligations under the insurance contract when making the claim, such as timely notification and preservation of evidence.

Court Case Name: Asean Security Paper Mills Sdn Bhd v. Mitsui Sumitomo Insurance [2008] 6 CLJ.

Fact List:

- The appellant, Encony Development Sdn Bhd, executed a sale and purchase agreement (SPA) with the respondents on 2 September 2010 for a condominium unit, whereby the respondents paid an initial deposit and subsequently the balance ten percent deposit.
- The statutory SPA included clauses that made timely payment of installments essential, allowing the appellant to terminate the agreement for non-payment.
- The respondents failed to make progress payments after receiving requests from the appellant and were subsequently issued a notice of default on 12 November 2010, followed by a lawful termination of the SPA on 10 December 2010.
- The respondents challenged the termination in the High Court, arguing that there were binding representations made prior to the execution of the SPA which created a collateral contract.

A.5.1 Case Study: Alignment with Expert Legal Reasoning

We present a detailed walkthrough of LEPREC's factor-based analysis in a Federal Court jurisdiction case, comparing it with annotators' answers to different reasoning questions from LIC construction. Note that the annotators didn't review all the reasoning questions, and we use this case study to illustrate the alignment of LEPREC and practical legal practitioners' reasoning process. The investigated issues are:

- (**ground truth: Irrelevant**) is: *"Whether the fire damage to the warehouse containing security paper is covered under the terms of the insurance policy, including any applicable exclusions or conditions."*

Expert Reasoning. Legal professionals naturally decompose this judgment into a set of reasoning questions:

Algorithm 2: Dataset Curation Pipeline

Input: Raw court case PDFs from CLJ
Output: Labeled set LIC_L and unlabeled set LIC_U

```
// Step 1: Fact and Issue Extraction
```

Apply GPT-4o to extract facts $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and silver truth issues $\hat{\mathcal{Y}}$ from each case PDF;

```
// Step 2: Incremental Issue Generation
```

for $i = 1$ **to** m **do**
  Generate additional issue candidates $\hat{\mathcal{Y}}_i$ using facts $[\mathbf{x}_1, \ldots, \mathbf{x}_i]$;
  Update $\hat{\mathcal{Y}} = \hat{\mathcal{Y}} \cup \hat{\mathcal{Y}}_i$;

```
// Step 3: Expert Annotation
```

For each (fact set, candidate issue) pair $\langle \mathbf{X}, \hat{Y}_j \rangle$, collect binary relevance label $y_j \in \{\text{Relevant}, \text{Irrelevant}\}$ from senior legal experts;

Pairs with expert labels form LIC_L; GPT-extracted silver truth pairs form LIC_U;

- **Q1** *Is this issue within the Federal Court's jurisdictional scope?* $\rightarrow$ **No.** The Federal Court hears appeals on questions of law, not fresh coverage disputes.
- **Q2** *Was this specific dispute actually before this court?* $\rightarrow$ **No.** The dispute concerns whether the Court of Appeal correctly reversed the High Court's factual finding, not policy coverage.
- **Q3** *Does this issue involve fresh factual determinations?* $\rightarrow$ **Yes.** Interpreting policy terms would require fresh factual analysis the Federal Court does not conduct.
- **Q4** *Does this issue relate to the established facts?* $\rightarrow$ **Yes.** The issue connects to fire damage facts, providing factual context but not legal relevance.
- **Q5** *Is this a derivative or secondary issue?* $\rightarrow$ **Yes.** Policy coverage only becomes relevant after determining the cause of fire, which is the primary appellate question.

Expert conclusion: "This issue is irrelevant because it falls outside the Federal Court's jurisdictional mandate and does not address the core legal question under appeal. While factually connected, it is a derivative concern not before the court."

LEPREC Analysis. Table 7 shows the reasoning questions activated by LEPREC's learned model for this issue-fact pair. The linear combination produces a strong negative score (-0.417), correctly predicting **Irrelevant**.

Table 7: LEPREC's active reasoning questions for the insurance coverage issue.

Feature	Weight	Reasoning Question	Answer
f_{776}	-0.211	Is this issue within Federal Court's jurisdictional scope?	No
f_{1534}	-0.080	Was this dispute actually before this court?	No
f_{751}	-0.080	Does the issue involve fresh factual determinations?	Yes
f_{440}	$+0.122$	Does the issue relate to established facts?	Yes
f_{2384}	-0.065	Is this a derivative/secondary issue?	Yes
f_{1820}	-0.058	Does this require interpretation of contractual terms?	Yes
f_{892}	-0.045	Would resolving this require examining insurance industry standards?	Yes
f_{1247}	$+0.039$	Does the issue reference specific statutory provisions?	No
f_{563}	-0.042	Is this about quantum or extent of damages?	No

$$\begin{aligned}\text{Score} &= (-0.211) + (-0.080) + (-0.080) \\ &\quad + (+0.122) + (-0.065) + (-0.058) \\ &\quad + (-0.045) + (+0.039) + (-0.042) + \ldots \\ &= -0.417 \text{ (strong negative)}\end{aligned}$$

$\Rightarrow$ **Prediction: Irrelevant** ✓*Correct*

Alignment with Expert Reasoning. Three observations confirm alignment between LEPREC's learned weights and expert legal reasoning. *(i) Jurisdictional reasoning dominates*: The highest-weighted feature (f_{776}, -0.211) directly corresponds to the expert's primary consideration, whether the issue falls within the Federal Court's jurisdictional scope. *(ii) Derivative reasoning is captured*: The negative weight on f_{2384} (derivative issue) mirrors the expert's conclusion that policy coverage is secondary to the cause-of-fire question. *(iii) Factual connection is appropriately discounted*: The positive weight on f_{440} (relates to established facts) reflects that the issue is factually connected, but this positive signal is outweighed by the jurisdictional and procedural negatives, consistent with expert reasoning that factual relatedness alone does not establish relevance.

A.5.2 Human Usability Study

We further evaluate whether LEPREC's reasoning questions are useful to legal practitioners in practice, and whether this utility extends beyond the Malaysian legal context.

Investigated Legal Issue Candidates.

1. Whether the alleged pre-SPA representations created a binding collateral contract that could override the terms of the statutory SPA.
2. Whether the clauses in the statutory SPA making timely payment of instalments essential were valid and enforceable.
3. Whether the respondents breached a fundamental term of the SPA by failing to make required progress payments.
4. Whether the notice of default issued on 12 November 2010 was sufficient and in compliance with the terms of the SPA.
5. Whether the appellant's termination of the SPA was lawful given the respondents' failure to make progress payments.

Reasoning Questions. LEPREC selected 50 reasoning questions from the learned linear model: the top 25 positively weighted questions (indicative of relevance) and the top 25 negatively weighted questions (indicative of irrelevance), listed in Table 16. Positively weighted questions target core dispute and outcome-determinativeness, while negatively weighted questions flag peripheral or domain-mismatched considerations, mirroring the interpretability mechanism illustrated in the case study above.

Participants and Procedure. We invited two legal practitioners with substantial experience in contract law: a lawyer with over 10 years of experience in civil and commercial litigation from China (Lawyer #1), and a law professor from Australia (Lawyer #2). Both practitioners are from countries outside Malaysia, thereby providing evidence for the generalisability of LIC and LEPREC beyond the Malaysian legal context. Due to time constraints, participants reviewed the reasoning questions holistically across the five issue candidates based on the given facts of the court case and provided feedback via a structured follow-up interview on three questions:

- ***Reasoning Alignment*** *- To what extent does each factor reflect a reasoning step you would naturally consider when judging whether a legal issue is relevant to the case facts?.*
- ***Discriminative Validity*** *- Given the facts and different issues, does the factor's answer difference between the different issues align with your own judgment about why one issue is relevant and the other is not?*
- ***Practical Utility*** *- Do you find these factors useful for you as a practitioner in the legal domain?*

Practitioner #1 (China). *Reasoning Alignment (Q1):* The practitioner rated overall alignment as high, noting that reasoning around the core dispute was "almost entirely consistent" with their own natural reasoning. Coding the interview transcript reveals three key themes. *(i) Core dispute alignment*: Questions targeting the main point of contention (Q3), the rights and obligations of the parties (Q13), and the validity of specific contractual clauses (Q45, Q50) were identified as directly reflective of natural legal reasoning, paralleling the jurisdictional and derivative reasoning identified in the case study above. *(ii) Domain mismatch*: Q33 (*"Is the issue related to a specific professional standard?"*) was flagged as misplaced for a contractual dispute, as the analysis turns on legal provisions and contractual obligations rather than professional conduct norms, consistent with its negative weight in the model. *(iii) Complementary coverage*: The practitioner noted that questions they would not spontaneously consider did not detract from the list; rather, they ensured comprehensive coverage of case-specific considerations. The practitioner further observed that a well-designed reasoning question should be both general enough to apply across cases of the same type and specific enough to accommodate individual circumstances, a balance they found the list maintained well.
Discriminative Validity (Q2): The practitioner rated discriminative validity as "Yes," confirming that the answer patterns of the reasoning questions were consistent with their own relevance judgments, describing the discrimination as "professionally competent." A concrete example was provided: for Issue 4 (validity of the notice of default), any analysis necessarily involves interpreting specific contractual terms, which aligned precisely with the answer patterns of Q45 and Q50. No reasoning question was found to produce a distinction contrary to the practitioner's judgment.
Practical Utility (Q3): The practitioner found the reasoning questions practically useful, noting that they "address real, concrete considerations rather than abstract principles" and help practitioners "think more comprehensively and identify the core dispute more directly." Q3 was highlighted as particularly effective for Issue 2 (validity of the installment payment clause), focusing attention pre-

cisely on the question the court needed to resolve. The practitioner further noted that the list serves a dual function: general questions confirm reasoning the practitioner would already undertake, while case-specific questions serve as useful prompts, reminding practitioners to consider aspects they might not have thought to examine independently.

Practitioner #2 (Australia). *Reasoning Alignment (Q1):* The practitioner rated alignment as 2 out of 5 ("to some extent"), noting that in practice they "would not normally ask all of these" and would not have a fixed list of questions for any given consideration. The practitioner observed that the actual reasoning process is "much more intuitive than considering a list of particular questions in turn," and that which factors are relevant depends on the nature of the matter. Importantly, however, the practitioner acknowledged that "the factors that were listed were relatively comprehensive in terms of coverage," which is precisely the design intent of LEPREC. The lower rating, therefore, reflects the prescriptive format rather than a rejection of the underlying reasoning dimensions.

Discriminative Validity (Q2): The practitioner rated discriminative validity as "Partially," noting that with such a brief set of facts, it is difficult to identify the appropriate level of relevance without knowing the full procedural context. Specifically, the practitioner observed that it was not immediately apparent why the matter went to the High Court, and that whether this involved first instance or appellate proceedings would materially affect the relevance of several issues. This reflects the same challenge our framework addresses: relevance judgment is inherently context-sensitive, and the difficulty of assessing it from facts alone is a key motivation for structured reasoning question coverage.

Practical Utility (Q3): The practitioner found the questions useful as a "catch all" list that provides "useful general coverage of many if not all of the key questions," echoing Lawyer #1's observation about comprehensive coverage. However, the practitioner noted they are not useful as a prescriptive checklist, as many questions would have no specific relevance to a given issue. The practitioner further observed that the list may be of value to junior practitioners learning what to consider, while cautioning that going through a large number of irrelevant questions risks losing concentration. This aligns with our quantitative finding that domain-specific questions contribute conditionally rather than universally, reinforcing the case for adaptive weighting over fixed selection.

Discussion. Together, the case study and usability study provide complementary evidence for LEPREC's interpretability. The case study demonstrates that the learned weights reproduce expert legal reasoning at the feature level: jurisdictional and procedural questions dominate, factual relatedness is appropriately discounted, and derivative issues are correctly penalized. The usability study extends this to practitioner utility beyond the Malaysian legal context. Despite differences in rating and background, both practitioners independently converged on the same core observation: the value of LEPREC's reasoning questions lies in comprehensive coverage rather than fixed prescription, and legal reasoning in practice does not follow a universal checklist. This convergence validates both our framework design and the broader claim that contract law reasoning exhibits universal principles that generalize across jurisdictions (cf. RQ3, Sec . 4).

A.6 Discussion on Deployment Readiness

While LEPREC demonstrates substantial improvements over state-of-the-art LLM baselines, achieving 79.70-80.19% F1 compared to 55-62% for GPT-4o and Claude, we acknowledge that the current system represents research-stage performance rather than a production-ready deployment. To properly evaluate this performance gap, we must first contextualize LEPREC's results against realistic human expert baselines. Our inter-annotator agreement analysis reveals that three senior legal scholars with over 15 years of experience each achieved Fleiss' $\kappa = 0.659$ ("Substantial Agreement"), with pairwise Cohen's κ scores ranging from 0.584 to 0.746. Using majority voting as the gold standard consensus, this indicates that a single expert achieves approximately 65-75% agreement with consensus judgments. LEPREC's 80% F1 score actually matches or exceeds typical individual human expert performance against consensus, demonstrating that the system has achieved human expert-level accuracy on this inherently subjective task. This subjectivity is fundamental to legal issue relevance assessment. Experts with identical training and rubrics often reach different legitimate conclusions on the same cases, reflecting interpre-

tive differences rather than errors. This suggests that the realistic performance ceiling for this task is bounded by legitimate expert disagreement, likely around 75-85% F1, and LEPREC's performance places it within this realistic ceiling range. Thus, we argue that the gap to production deployment is therefore not primarily about improving accuracy beyond human expert levels, but rather about meeting the operational, robustness, and trust requirements of real-world legal practice, such as adding post-checking with real legal professionals after prediction.

A.7 Diversity Discussion

Table 1 shows that the *incremental* generator beats the one–pass *baseline* on both quality and diversity.

First, for quality, Fréchet BERT Distance drops from 1311 to 1177 and BERT Embedding Distance from 1354 to 1227. Because smaller distances mean the candidates sit nearer to the ground-truth issues in embedding space, these reductions (roughly 10%) confirm that the staged prompts recover more of the key semantics that human experts mark.

On the diversity side, three signals move in the right direction. Self-EMBD rises from 211.8 to 225.1, so the candidates spread out further in representation space. Raw Self-BLEU nearly doubles at the 3–, 4–, and 5–gram levels (11.85/14.52/16.64 versus 24.90/30.36/34.95). If one prefers to report $1-$Self-BLEU, the gain is the same magnitude, just flipped in sign. Distinct-N climbs by about 40%, adding more than three thousand new 3-, 4-, 5-grams to the pool.

In short, revealing facts to the LLM one step at a time both tightens semantic coverage and uncovers a broader set of legally plausible formulations—a combination that downstream filters and reviewers can exploit.

A.8 Data Example: Case ID - IFSG681

A.8.1 Case Facts (Scenario)

- The developer, Sri Damansara Sdn Bhd, collected a booking fee of RM10,000 on 6 January 2012 prior to signing the sale and purchase agreement (SPA) dated 28 June 2012, which contravened Housing Development regulations.

- When the vacant possession was delivered on 22 December 2016, the purchasers filed a claim for damages for late delivery, calculating it from the booking fee payment date to the delivery date, which the Tribunal upheld.

- The developer argued that the calculation should start from the SPA date and questioned the validity of the Tribunal's decision and the method of calculating the purchase price considering a credit note provided.

A.8.2 Ground Truth Issues

1. Whether the late delivery claim ought to be calculated from the expiry of 42 months from the date of the SPA of 28 June 2012 until the date of delivery of vacant possession.

2. Whether the purchase price for the calculation of the late delivery claim should be based on the price as stated in the SPA or the reduced price via the device of a credit note issued by the developer.

A.8.3 Issue Candidates

Scenario#1:

1. The developer, Sri Damansara Sdn Bhd, collected a booking fee of RM10,000 on 6 January 2012 prior to signing the sale and purchase agreement (SPA) dated 28 June 2012, which contravened Housing Development regulations.

Generated Issue Candidates Given Scenario#1:

1. Whether the collection of a booking fee by Sri Damansara Sdn Bhd prior to signing the sale and purchase agreement (SPA) is in violation of Housing Development regulations.

2. Whether the developer's action of collecting a booking fee on 6 January 2012, nearly six months before the SPA was signed on 28 June 2012, is legal and permissible under applicable housing laws.

Scenario#2:

1. The developer, Sri Damansara Sdn Bhd, collected a booking fee of RM10,000 on 6 January 2012 prior to signing the sale and purchase agreement (SPA) dated 28 June 2012, which contravened Housing Development regulations.

2. When the vacant possession was delivered on 22 December 2016, the purchasers filed a claim for damages for late delivery, calculating it from the booking fee payment date to the delivery date, which the Tribunal upheld.

Generated Issue Candidates Given Scenario#2:

1. Whether the developer's collection of a booking fee prior to signing the sale and purchase agreement (SPA) was in contravention of Housing Development regulations.

2. Whether the calculation of late delivery damages should commence from the date of the booking fee payment or from the date of the signed sale and purchase agreement (SPA).

Scenario#3:

1. The developer, Sri Damansara Sdn Bhd, collected a booking fee of RM10,000 on 6 January 2012 prior to signing the sale and purchase agreement (SPA) dated 28 June 2012, which contravened Housing Development regulations.

2. When the vacant possession was delivered on 22 December 2016, the purchasers filed a claim for damages for late delivery, calculating it from the booking fee payment date to the delivery date, which the Tribunal upheld.

3. The developer argued that the calculation should start from the SPA date and questioned the validity of the Tribunal's decision and the method of calculating the purchase price considering a credit note provided.

Generated Issue Candidates Given Scenario#3:

1. Whether the developer's collection of a booking fee prior to signing the Sale and Purchase Agreement (SPA) was in contravention of Housing Development regulations.

2. Whether the calculation of damages for late delivery should start from the date of the booking fee payment or the date of the SPA.

3. Whether the Tribunal's decision to uphold the purchasers' claim for damages based on the booking fee payment date is valid.

4. Whether the method of calculating the purchase price should consider the credit note provided by the developer.

A.9 Example of Sparsity of Mutual Information

Legal Facts:

1. Fact#1. The Appellants, Tioh Chee Seng and Hew Fui Li, purchased a residential unit in Ayuman Suites under a sale and purchase agreement dated 21 October 2015.

2. Fact#2. The 1st Respondent, Talent Team Sdn. Bhd., is the developer responsible for constructing Ayuman Suites, while the 2nd Respondent, a legal firm, is alleged to be the stakeholder of certain sums related to the purchase.

3. Fact#3. The Appellants claimed late delivery of the property, seeking liquidated ascertained damages (LAD) based on delays exceeding the stipulated timeframes for completion.

4. Fact#4. The main dispute centers on the calculation of the delay—specifically, whether it should be based on the booking fee payment date or the date of signing the sales and purchase agreement.

5. Fact#5. The Sessions Court ruled that the LAD calculation should commence from the signing of the sales and purchase agreement.

Identified Legal Issues (Selected):

1. Issue #1 (Identified when providing Facts 1–3): Whether the 2nd Respondent, as the alleged stakeholder, has any liability regarding the sums related to the purchase.

2. Issue #2 (Identified when providing Facts 1–4): Whether the legal firm, acting as the alleged stakeholder, bears any responsibility in the dispute over the late delivery and calculation of LAD.

3. Issue #3 (Identified when providing Facts 1–5): Whether the 2nd Respondent (legal firm) has any liability as the alleged stakeholder of certain sums related to the purchase.

In this case, Issues #1, #2, and #3 all relate to the liability of the 2nd Respondent (Legal Firm). Once Issue #1 is identified based on Facts 1–3, the subsequent iterations—despite introducing additional facts—continue to surface the same or similar issues. This suggests that Facts #4 and #5 contribute little to the identification of this legal issue.

A.10 Prompt Template for Incremental Issue Generation

```
    Scenario: \{scenario\}

This scenario describes a legal case. Based on
    the details provided, please identify the
    most relevant legal issues.

Guidelines:
1.      Do not alter or deviate from the meaning
     presented in the scenario.
2.      Format each legal issue as "Whether ...",
     for example: "Whether the alleged agreement
     between the plaintiff and defendant is
    enforceable considering the Statute of
    Frauds."
3.      Provide your response strictly in JSON
    format as shown below:

{["YOUR FIRST LEGAL ISSUE","YOUR SECOND LEGAL
    ISSUE", ...}
```

Listing 3: Prompt for Incremental Issue Generation

A.11 Evaluation Guideline for Human

Facts Evaluation High Distinction (HD):

- Facts are presented clearly and concisely in a structured point form.
- Closely aligned with statutory language and terminology.
- No irrelevant details, and all essential elements are thoroughly included.

Pass:

- Facts are mostly accurate and clear, though some minor details may be missing or imprecise.
- Minor elements could be better structured or clarified.

Not Pass:

- Facts are incomplete, unclear, or contain irrelevant information that detracts from the analysis.
- Key details are missing, leading to a lack of proper context.

Neutral:

- Facts are presented and generally acceptable, but lack the depth or clarity needed for proper evaluation.
- Facts may not align clearly with the case or legal standards, preventing detailed assessment.

Issues Evaluation High Distinction (HD):

- All relevant legal issues are clearly identified in a structured manner, typically starting with *"Whether..."*.
- Issues are aligned with the facts and the applicable rules, demonstrating a comprehensive understanding.

Pass:

- Most key legal issues are identified, but some may be phrased imprecisely or omitted.
- Overall, the issues are reasonable, but there may be minor gaps in alignment with facts and rules.

Not Pass:

- Significant legal issues are missing or misidentified, demonstrating a poor understanding of the case.
- Issues are formulated incorrectly or too broadly.

Neutral:

- Issues are present, but lack clarity, structure, or alignment with the case, making it difficult to assess their relevance.

A.12 Baseline Settings

We run all the experiments on NVIDIA A100 GPUs. For the hyperparameters of the LLMs, e.g., Temperature, Top-p, etc, we use the default settings for all commercial models, including Claude and GPT-4o. The random seeds are set to 42.

We adopt the default settings for all generative models. Note that for Qwen3-14B, we disable the thinking function and only look into their `no think` mode. For the regression methods, we use GridCV to tune the hyperparameters, including λ, C, etc. For the deep learning methods, we tune the learning rate and epoch based on the validation set.

A.13 Prompt Template for Preliminary Results

```
You are a meticulous judge deciding whether the
    issue below is "Relevant" or
"Irrelevant" under the exact definitions
    provided.

Definitions (use exactly):
```

```
- "Relevant":  The issue is not only related to
    the scenario, but is directly
  tied to the main dispute or core facts of the
    case. It goes beyond merely
  stating a basic legal principle or background
    fact.
- "Irrelevant": The issue is either unrelated to
     the scenario or is only a
  fundamental/basic statement that does not bear
     on the case's primary
  controversy.

Think silently and step-by-step.  Follow these
    internal reasoning steps:
  1. Identify the main dispute and the core
    facts from the Scenario Facts.
  2. Compare the candidate Issue to the dispute:
     does it address that dispute
    directly, or is it merely background/
    unrelated?
  3. Decide if the Issue adds substantive
    analysis beyond a basic legal truism.
  4. Conclude "Relevant" or "Irrelevant" based
    on the definitions.
- Do NOT reveal your reasoning or the steps
    above.
- After finishing your internal analysis, output
     **exactly one word**-
  either "Relevant" or "Irrelevant"-and nothing
    else.

### Scenario Facts
{facts}

### Issue
{issue}

### Instruction
First reason internally following the steps.
Then output one word: Relevant OR Irrelevant

Your response:
```

Listing 4: Prompt for Preliminary Results

A.14 Full List of RQ2

B RQ3: Selection Stability Analysis

Methodology *Bootstrap Stability Selection Protocol:* For both L1 Logistic Regression and L1 SVC, we employ bootstrap-based stability selection within nested 5-fold cross-validation.

Outer loop (5-fold CV): Split data into 5 folds. For each fold k: (i) Use 4 folds for training, 1 fold for testing; (ii) Perform hyperparameter tuning via inner 5-fold CV on training data; (iii) Select optimal regularization parameter C_k; (iv) Run stability selection procedure; (v) Evaluate on held-out test fold.

Stability selection within each training fold: Perform 100 bootstrap iterations. In each iteration i: randomly subsample 60% of training data (without replacement), fit L1 model using optimal C_k on subsample, and record which features have non-zero coefficients: $S_i = \{j : w_j \neq 0\}$. Then compute selection frequency for each feature: $\text{freq}(j) = \frac{1}{100}\sum_{i=1}^{100} \mathbb{1}[j \in S_i]$. Apply threshold $\tau \in \{0.3, 0.4, 0.5, 0.6\}$: feature j is "stable" in fold k if $\text{freq}(j) \geq \tau$.

Cross-fold consistency analysis: For each feature j, count in how many folds it was deemed stable: $\text{stability}(j) = \sum_{k=1}^{5} \mathbb{1}[\text{freq}_k(j) \geq \tau]$. We categorize features as: always selected ($\text{stability}(j) = 5$), highly stable ($\text{stability}(j) = 4$), moderately stable ($\text{stability}(j) = 3$), unstable ($\text{stability}(j) \in \{1, 2\}$), and never selected ($\text{stability}(j) = 0$).

Comparison with Standard L2 Linear Models: To assess whether L1-selected features are genuinely important, we fit Standard (L2-regularized) Logistic Regression on the full dataset using optimal C determined by cross-validation. This provides reference coefficients β_j^{L2} and p-values p_j for all 2464 features. We then compare L1-selected features' L1 coefficients versus Standard LR coefficients. Our hypothesis is that if L1 identifies truly important features, $|\beta_j^{\text{L1}}|$ should correlate positively with $|\beta_j^{\text{L2}}|$. However, we observe near-zero or negative correlations, proving disagreement.

L1 Logistic Regression Detailed Results *Complete Selection Statistics:* Table 11 presents comprehensive statistics across all four threshold configurations.

Always-Selected Features Analysis: Table 12 examines all features selected in all 5 folds for each threshold configuration, comparing their L1 coefficients with Standard LR weights.

Critical observation: Every single always-selected feature across all thresholds has Standard LR coefficient $|\beta|<0.07$ and p-value >0.999, indicating no statistical significance. These features are consistently selected by L1 not because they are important, but because they are arbitrary representatives picked first from correlated question clusters during L1's optimization path. Figure 7 through Figure 10 visualize the selection patterns across thresholds.

Per-Fold Performance Details: Table 13 shows detailed per-fold statistics revealing the relationship between feature counts, hyperparameters, and performance.

Key observations emerge from this analysis. Fold 2 consistently has the lowest C (0.17), selecting fewest features (8 to 35), often with worst

Table 8: Results of Deep Learning Methods

Methods	F1	Acc.	Prec.	Rec.
FFN_{oss}	$56.82_{\pm 3.96}$	$57.90_{\pm 4.79}$	$60.90_{\pm 3.02}$	$61.96_{\pm 3.18}$
CNN_{oss}	$39.35_{\pm 4.36}$	$59.92_{\pm 13.08}$	$44.21_{\pm 9.52}$	$49.40_{\pm 2.40}$
$Tranf._{oss}$	$61.72_{\pm 2.83}$	$63.05_{\pm 3.83}$	$65.22_{\pm 1.64}$	$66.85_{\pm 2.22}$
$LSTM_{oss}$	$47.89_{\pm 4.70}$	$64.15_{\pm 4.70}$	$51.08_{\pm 10.29}$	$51.45_{\pm 1.85}$
$ResNet_{oss}$	$60.47_{\pm 4.59}$	$65.57_{\pm 4.90}$	$60.80_{\pm 5.09}$	$60.74_{\pm 4.31}$
FFN_{Phi}	$60.79_{\pm 5.32}$	$66.73_{\pm 9.21}$	$65.60_{\pm 6.67}$	$61.96_{\pm 2.92}$
CNN_{Phi}	$40.47_{\pm 9.20}$	$56.17_{\pm 16.12}$	$47.73_{\pm 23.15}$	$50.11_{\pm 2.64}$
$Tranf._{Phi}$	$70.53_{\pm 3.74}$	$75.33_{\pm 5.80}$	$75.06_{\pm 4.81}$	$71.13_{\pm 2.46}$
$LSTM_{Phi}$	$45.36_{\pm 8.84}$	$59.24_{\pm 13.10}$	$52.97_{\pm 3.18}$	$51.78_{\pm 2.14}$
$ResNet_{Phi}$	$67.33_{\pm 3.61}$	$72.48_{\pm 4.19}$	$68.84_{\pm 4.89}$	$67.33_{\pm 3.40}$
FFN_{Qwen}	$75.65_{\pm 2.57}$	$79.29_{\pm 2.43}$	$76.10_{\pm 2.84}$	$75.57_{\pm 2.47}$
CNN_{Qwen}	$43.33_{\pm 3.08}$	$69.78_{\pm 0.96}$	$54.78_{\pm 24.83}$	$51.22_{\pm 1.51}$
$Tranf._{Qwen}$	$75.44_{\pm 1.82}$	$80.14_{\pm 2.02}$	$78.22_{\pm 3.64}$	$74.39_{\pm 2.10}$
$LSTM_{Qwen}$	$63.48_{\pm 2.94}$	$65.90_{\pm 3.79}$	$64.04_{\pm 1.93}$	$65.81_{\pm 1.84}$
$ResNet_{Qwen}$	$67.21_{\pm 3.76}$	$72.30_{\pm 2.93}$	$67.42_{\pm 3.66}$	$67.06_{\pm 3.81}$

performance. Folds 1, 4, and 5 have highest C (100), selecting most features (34 to 362), with varied performance. Despite 10× feature count variation within thresholds, F1 varies only 6 to 20 points. Threshold 0.5 shows extreme variance: best fold (F1=0.823) and some of worst (F1=0.671). No consistent relationship exists between number of features and performance within or across thresholds.

Negative Correlation Analysis: The correlation between L1 and Standard LR coefficients becomes increasingly negative as thresholds tighten. At threshold 0.3, correlation is r = −0.026 (essentially zero). At threshold 0.4, r = −0.059 (weak negative). At threshold 0.5, r = −0.106 (moderate negative). At threshold 0.6, r = −0.249 (strong negative). This monotonic deterioration proves that as L1 is forced to select fewer features, it increasingly prioritizes features that Standard LR considers less important. By threshold 0.6, L1 and Standard LR are in substantial disagreement, demonstrating that aggressive sparsity causes L1 to make systematically poor selections.

L1 SVC Detailed Results *Complete Selection Statistics:* Table 14 presents comprehensive statistics for L1 SVC across threshold configurations.

Comparison with L1 LR reveals key differences. L1 SVC achieves 32× more stable selections (158 vs. 5 always-selected at threshold 0.4), but selects 7 to 23× more features per fold (average 792 vs. 113 features at threshold 0.4). It shows better performance: 0.746 vs. 0.722 F1 at threshold 0.4 (2.4 point gain). L1 SVC exhibits less hyperparameter sensitivity with C varying only 3.6× vs. 599× for L1 LR, and smaller feature count variation with 2.4 to 6.5× range vs. 8 to 10× for L1 LR.

Always-Selected Features Analysis: Unlike L1 LR, L1 SVC's always-selected features include genuinely high-weighted features. However, the set changes dramatically with threshold. At threshold 0.3, 386 always-selected features (15.7% of total) emerge. The top 10 include f776 (β=−0.211), f622 (β=0.205), f1914 (β=0.179), f753 (β=0.172), and f119 (β=0.170). These are spread across many legal reasoning categories and represent comprehensive retention rather than sparse selection. At threshold 0.6, only 24 always-selected features (1.0% of total) remain. The top 10 include f776 (β=−0.211), f1914 (β=0.179), f753 (β=0.172), f1952 (β=0.158), and f751 (β=−0.152). These overlap with threshold 0.3 top features but represent a dramatically reduced set, still more stable than L1 LR's single always-selected feature. Figure 3 through Figure 6 illustrate these patterns.

Per-Fold Performance Details: Table 15 shows L1 SVC's per-fold statistics.

A key observation emerges: Fold 2 with C=100 selects 2369 features (96% of all features) and achieves best performance (F1=0.785). Other folds with C=27.83 select approximately 600 features with lower but stable performance (F1=0.702 to 0.753). This demonstrates L1 SVC's tendency toward dense solutions when C is high, contradicting the sparsity objective of L1 regularization.

Table 9: Results of Feature Selection Methods

Methods	F1	Acc.	Prec.	Rec.
$\text{L1Reg}_{\text{Phi}}$	$80.01_{\pm 3.61}$	$83.34_{\pm 3.06}$	$81.13_{\pm 3.93}$	$79.32_{\pm 3.45}$
$\text{L1Reg}_{\text{oss}}$	$65.69_{\pm 3.74}$	$72.05_{\pm 3.29}$	$66.91_{\pm 4.30}$	$65.07_{\pm 3.44}$
$\text{L1Reg}_{\text{Qwen}}$	$73.80_{\pm 2.25}$	$78.03_{\pm 2.04}$	$74.47_{\pm 2.57}$	$73.39_{\pm 2.22}$
$\text{L1SVC}_{\text{Phi}}$	$77.60_{\pm 2.58}$	$80.89_{\pm 1.97}$	$77.68_{\pm 2.23}$	$77.62_{\pm 2.93}$
$\text{L1SVC}_{\text{Qwen}}$	$70.61_{\pm 2.59}$	$74.83_{\pm 2.51}$	$70.68_{\pm 2.77}$	$70.61_{\pm 2.39}$
$\text{L1SVC}_{\text{oss}}$	$64.38_{\pm 1.57}$	$68.85_{\pm 1.47}$	$64.20_{\pm 1.57}$	$64.78_{\pm 1.69}$
LMSel.	$39.10_{\pm 0.49}$	$64.23_{\pm 1.34}$	$33.73_{\pm 0.22}$	$46.52_{\pm 0.99}$
LMSel._{L}	$45.63_{\pm 1.79}$	$50.25_{\pm 1.72}$	$46.27_{\pm 1.76}$	$45.84_{\pm 1.92}$
LMSel.^{P} + Regression (Continuous)				
Logistic	$74.86_{\pm 2.85}$	$77.44_{\pm 2.71}$	$74.24_{\pm 2.71}$	$76.32_{\pm 2.85}$
SVC	$73.98_{\pm 3.48}$	$76.85_{\pm 3.06}$	$73.41_{\pm 3.31}$	$75.14_{\pm 3.66}$
LDA	$71.45_{\pm 3.35}$	$77.19_{\pm 2.59}$	$73.75_{\pm 3.52}$	$70.36_{\pm 3.19}$
Ridge	$73.95_{\pm 2.31}$	$76.60_{\pm 2.09}$	$73.31_{\pm 2.17}$	$75.41_{\pm 2.46}$
LMSel.^{P} + Regression (Binary)				
Logistic	$57.09_{\pm 1.55}$	$64.90_{\pm 1.43}$	$57.79_{\pm 1.41}$	$57.06_{\pm 1.48}$
SVC	$57.78_{\pm 1.57}$	$66.33_{\pm 0.82}$	$58.90_{\pm 1.20}$	$57.64_{\pm 1.52}$
LDA	$53.23_{\pm 2.31}$	$65.74_{\pm 2.19}$	$56.38_{\pm 3.18}$	$54.07_{\pm 1.77}$
Ridge	$56.15_{\pm 0.43}$	$64.39_{\pm 1.88}$	$57.08_{\pm 0.91}$	$56.17_{\pm 0.38}$

Cross-Method Comparison *Selection Overlap Analysis:* We analyze which features are selected by both L1 LR and L1 SVC at threshold 0.4. L1 LR has 5 always-selected features, while L1 SVC has 158. The overlap (always-selected in both) contains only 3 features, representing 60% of L1 LR's selections but just 1.9% of L1 SVC's. When considering features selected in at least 3 folds, L1 LR has 29 such features while L1 SVC has 655. The overlap in this category contains 11 features, representing 38% of L1 LR but only 1.7% of L1 SVC. This low overlap (38 to 60%) proves that different L1-based methods identify different "important" features, further demonstrating selection instability. If a true essential question set existed, both methods should converge to it.

Why Does L1 SVC Select More Features: The key difference stems from the optimization objectives. L1 Logistic Regression minimizes

$$\min_{\mathbf{w},b} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i; \mathbf{w}, b) + \lambda \|\mathbf{w}\|_1 \right\}$$

The log-loss is smooth and convex, with L1 regularization pushing many weights to exactly zero when C is small. In contrast, L1 SVC minimizes

$$\min_{\mathbf{w},b} \left\{ \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) + \lambda \|\mathbf{w}\|_1 \right\}$$

The hinge loss has a non-smooth "hinge" at margin boundaries. For our high-dimensional, correlated feature space, SVC's dual formulation tends to utilize many support vectors, each requiring non-zero weights for its associated features. This dual structure makes L1 SVC less aggressive at zeroing weights than L1 LR. When C is large (weak regularization), L1 SVC approaches unregularized SVC, which uses many features to maximize margin. This explains Fold 2's selection of 2369/2464 features.

Interpretation and Implications Three fundamental reasons explain the absence of a universal essential question set.

Why No Stable Essential Set Exists: Question generation process creates context-specific, not universal, questions. Our reasoning questions were generated for specific case-issue pairs (approximately 300 pairs, approximately 8 questions each = 2464 total). These questions are contextualized to particular legal scenarios rather than being a universal reasoning bank. For example, "Is the issue central to the insurer's stated reason for denying the claim?" only applies to insurance cases, "Does the employee's conduct fall under the protected activity definition?" only applies to employment cases, and "Was proper notice given according to the lease terms?" only applies to property/contract cases. A question that is critical for one case type

is noise for others. This domain-specificity means no fixed subset works universally; effective reasoning requires context-dependent weighting. *Massive redundancy from correlated questions.* With 2464 questions, many are semantically similar due to rephrasing of the same concept ("Is this a legal vs. factual question?" versus "Does answering require applying law?"), different levels of abstraction ("Is this jurisdictional?" versus "Does this court have authority?" versus "Is venue proper?"), and domain-specific variants of general questions. This redundancy means many different sparse subsets capture similar information, explaining why L1 achieves 0.70 to 0.75 F1 with completely different feature sets across folds. Any reasonable subset of approximately 50 to 200 questions suffices because they collectively cover the key reasoning dimensions, but which specific 50 to 200 questions doesn't matter much.

Linear models enable implicit contextualization.: Standard linear models achieve 80% F1 (versus L1's 70 to 75%) by retaining all questions with learned weights. The linear combination $\mathbf{w}^\top \mathbf{f} = \sum_{j=1}^{h} w_j f_j$ enables domain-specific questions to contribute conditionally. When feature f_j (domain-specific) co-occurs with features indicating that domain (e.g., insurance keywords in case facts), their combined contribution is large. When domain indicators are absent, f_j's contribution is attenuated by negative contributions from other features. This implicit contextualization through weighted linear combinations is more sophisticated than L1's binary keep/discard decisions, explaining the 5 to 10 point performance gap.

Implications for Legal AI Systems: The findings have practical implications for building legal AI systems.

Favor retention over selection. Systems should retain comprehensive question sets with learned weights rather than attempting to identify "essential" subsets. The 5 to 10 point F1 gap between Standard linear models (80%) and L1 methods (70 to 75%) quantifies the cost of feature elimination in correlation-aware classification tasks.

Contextualization is key. Legal reasoning inherently requires context-dependent weighting. Domain-specific questions that appear noisy globally (low marginal contribution) are critical in specialized contexts. Systems must preserve these questions and adjust their influence based on case characteristics.

Interpretability versus performance tradeoffs. L1 provides sparse, seemingly interpretable models (e.g., "only 50 questions matter"), but this interpretability is illusory. Selected questions are unstable (vary across folds), different methods select different questions, selected questions have no special importance (near-zero Standard LR weights), and performance suffers due to elimination of contextual questions. True interpretability requires analyzing Standard linear model weights to understand which reasoning aspects receive highest weights across all questions, not which arbitrary subset L1 happens to select.

Summary: The comprehensive stability analysis across L1 LR and L1 SVC demonstrates five key findings. First, no stable universal essential question set exists for legal issue relevance. Second, L1-based selection achieves competitive performance with unstable, method-dependent feature subsets. Third, multiple sparse/dense combinations work equivalently due to massive feature redundancy. Fourth, retention with adaptive weighting (Standard models: 80% F1) outperforms elimination (L1: 70 to 75% F1). Fifth, contextualization through linear combinations is superior to binary feature selection. These findings conclusively validate Challenge 2 and justify our correlation-aware retention-based approach over sparsity-seeking feature selection methods.

Feature Stability Visualizations We illustrate the feature selection patterns through stability overview figures for both methods across all threshold configurations.

Figure 3: Stability Overview of L1 SVC (Threshold 0.3)

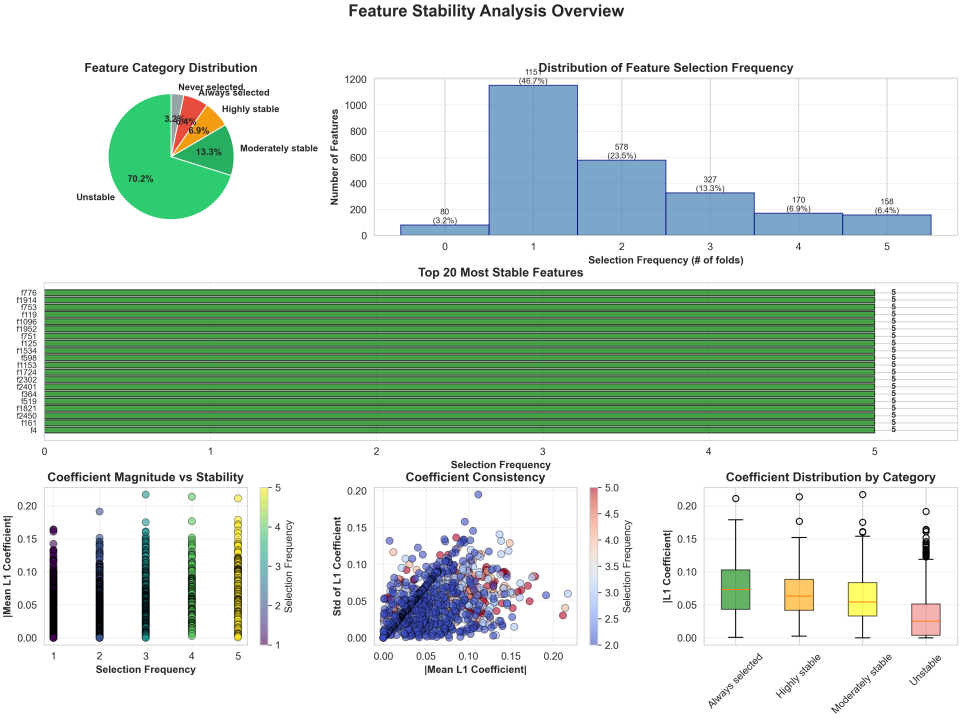

Figure 4: Stability Overview of L1 SVC (Threshold 0.4)

Figure 5: Stability Overview of L1 SVC (Threshold 0.5)

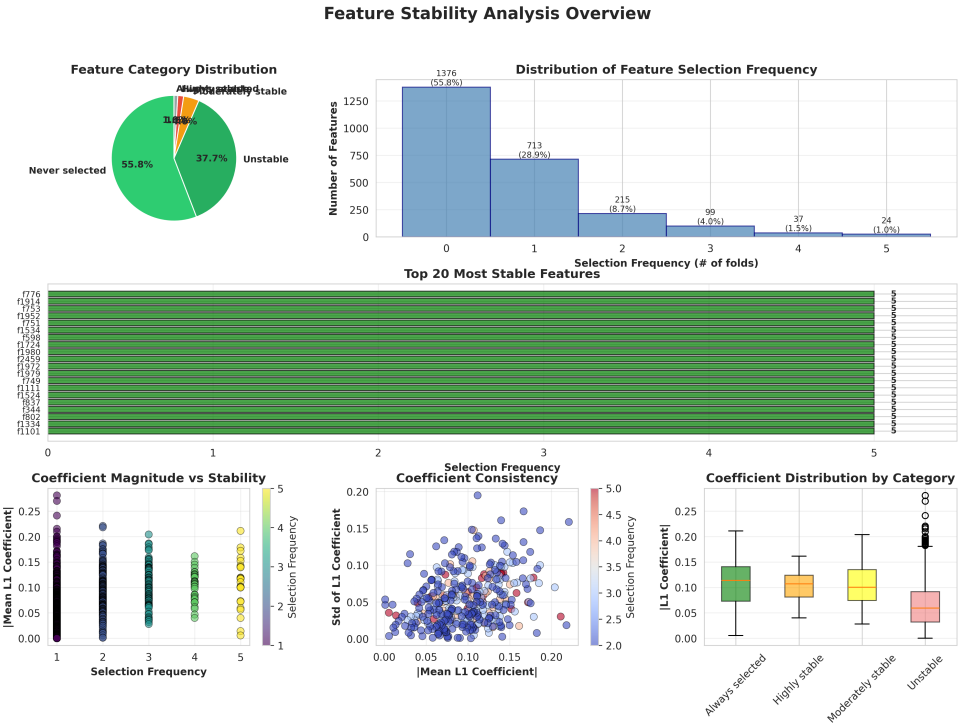

Figure 6: Stability Overview of L1 SVC (Threshold 0.6)

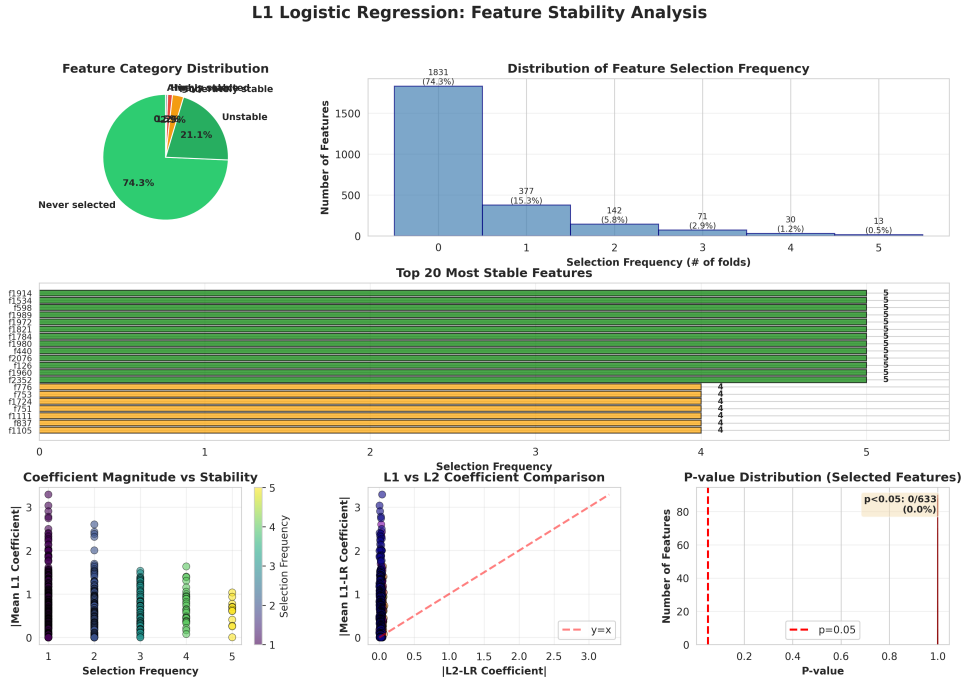

Figure 7: Stability Overview of L1 Regression (Threshold 0.3)

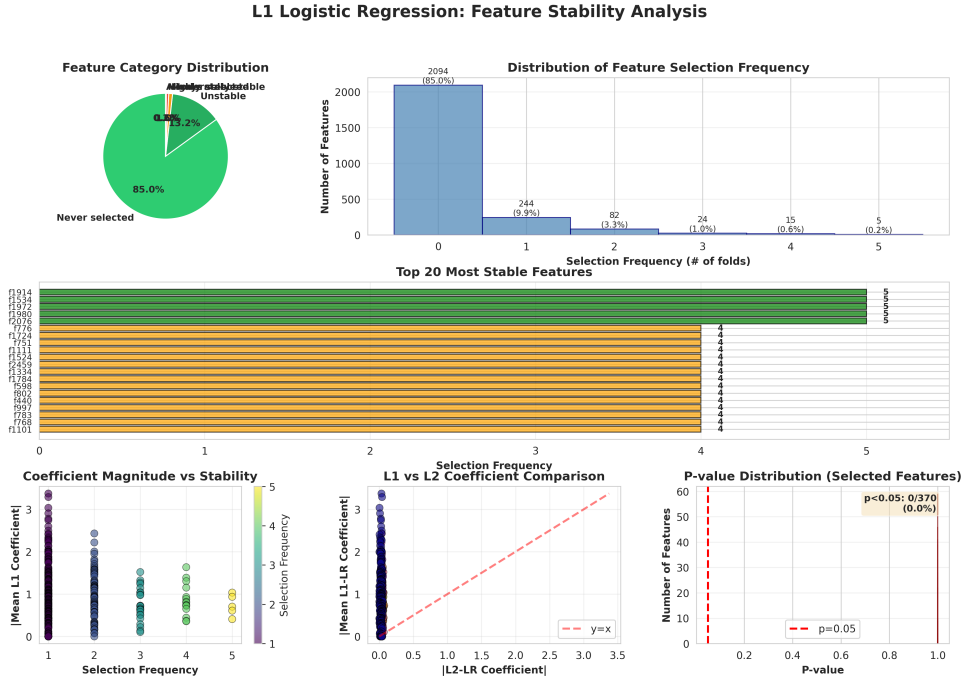

Figure 8: Stability Overview of L1 Regression (Threshold 0.4)

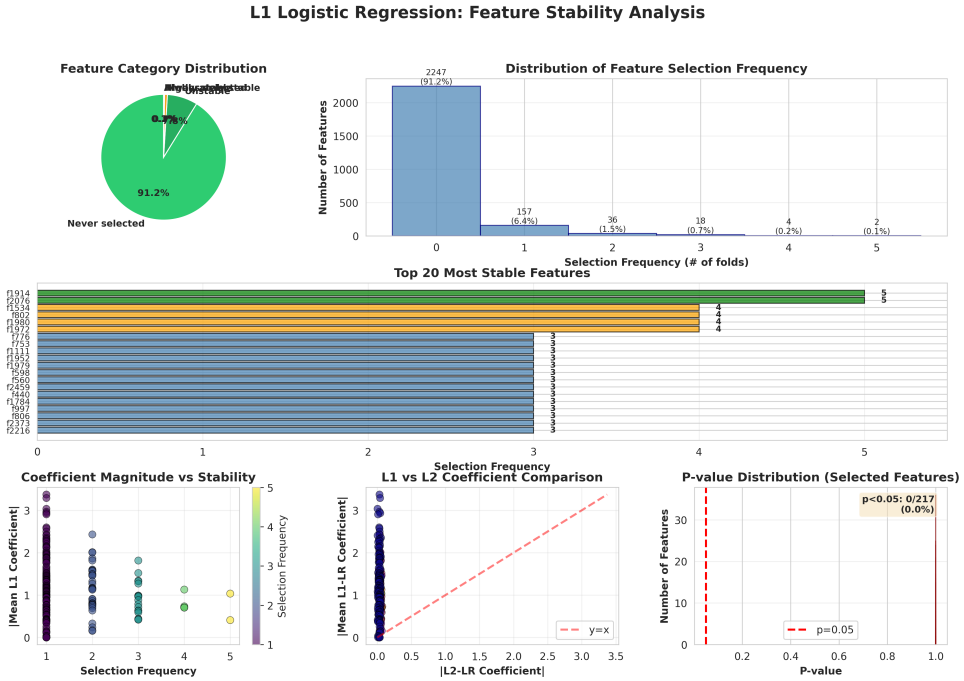

Figure 9: Stability Overview of L1 Regression (Threshold 0.5)

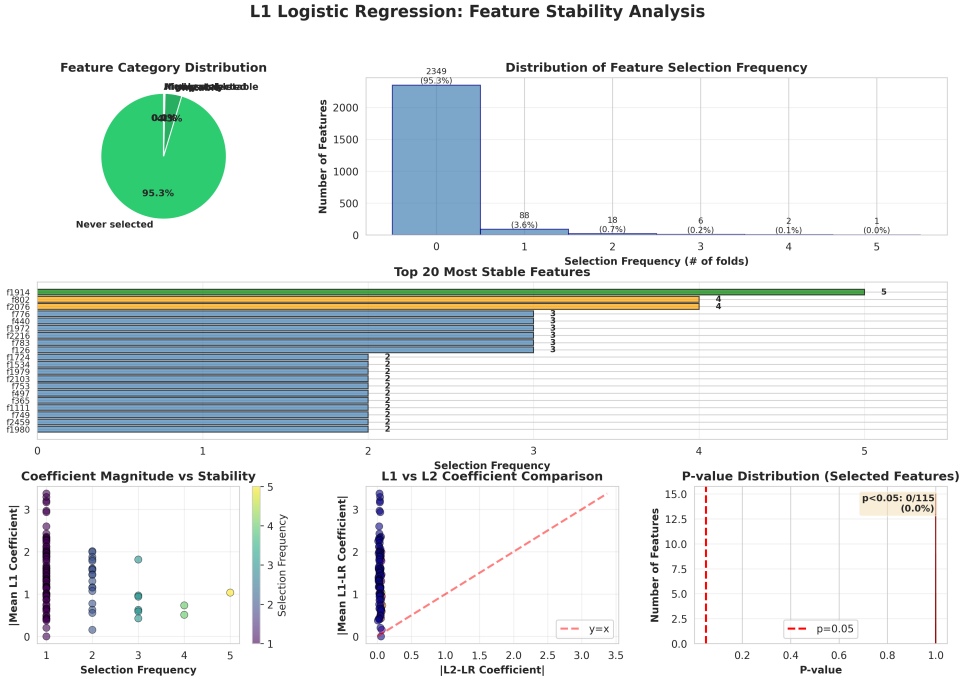

Figure 10: Stability Overview of L1 Regression (Threshold 0.6)

Table 10: Results of Regression Methods

Methods	F1	Acc.	Prec.	Rec.
LR_{Phi}	$79.70_{\pm 2.93}$	$82.49_{\pm 2.41}$	$79.58_{\pm 2.89}$	$80.05_{\pm 3.23}$
$\underline{LR_{Phi}}$	$64.70_{\pm 3.18}$	$68.86_{\pm 2.81}$	$64.56_{\pm 3.01}$	$65.41_{\pm 3.51}$
SVC_{Phi}	$80.19_{\pm 2.83}$	$82.66_{\pm 2.38}$	$79.67_{\pm 2.70}$	$81.01_{\pm 3.13}$
$\underline{SVC_{Phi}}$	$62.30_{\pm 3.56}$	$67.59_{\pm 2.30}$	$62.39_{\pm 3.18}$	$62.70_{\pm 4.07}$
$Ridge_{Phi}$	$80.10_{\pm 2.86}$	$82.91_{\pm 2.41}$	$80.06_{\pm 2.89}$	$80.28_{\pm 3.05}$
$\underline{Ridge_{Phi}}$	$61.63_{\pm 2.79}$	$65.99_{\pm 1.87}$	$61.56_{\pm 2.65}$	$62.44_{\pm 3.47}$
LDA_{Phi}	$79.56_{\pm 4.01}$	$83.50_{\pm 2.69}$	$81.77_{\pm 2.97}$	$78.39_{\pm 4.30}$
$\underline{LDA_{Phi}}$	$63.83_{\pm 3.09}$	$72.73_{\pm 1.73}$	$67.73_{\pm 2.41}$	$63.11_{\pm 2.79}$
SGD_{Phi}	$73.90_{\pm 9.33}$	$77.45_{\pm 8.89}$	$74.68_{\pm 9.97}$	$73.85_{\pm 9.07}$
$\underline{SGD_{Phi}}$	$62.95_{\pm 1.99}$	$67.09_{\pm 2.54}$	$62.80_{\pm 2.00}$	$63.58_{\pm 1.57}$
RF_{Phi}	$66.30_{\pm 3.01}$	$75.76_{\pm 2.49}$	$74.10_{\pm 4.82}$	$64.99_{\pm 2.54}$
$\underline{RF_{Phi}}$	$61.56_{\pm 4.09}$	$72.14_{\pm 2.16}$	$66.97_{\pm 4.58}$	$61.16_{\pm 3.28}$
KNN_{Phi}	$66.94_{\pm 3.24}$	$73.57_{\pm 3.05}$	$69.04_{\pm 4.09}$	$66.10_{\pm 3.02}$
$\underline{KNN_{Phi}}$	$62.46_{\pm 3.15}$	$71.63_{\pm 2.75}$	$66.25_{\pm 4.37}$	$61.78_{\pm 2.77}$
DT_{Phi}	$62.52_{\pm 2.13}$	$68.18_{\pm 1.54}$	$62.65_{\pm 1.99}$	$62.51_{\pm 2.36}$
$\underline{DT_{Phi}}$	$60.42_{\pm 4.01}$	$65.24_{\pm 3.80}$	$60.50_{\pm 3.76}$	$60.98_{\pm 4.18}$
NB_{Phi}	$49.61_{\pm 3.17}$	$51.00_{\pm 4.18}$	$53.17_{\pm 2.32}$	$53.52_{\pm 2.62}$
$\underline{NB_{Phi}}$	$39.63_{\pm 4.76}$	$61.12_{\pm 13.55}$	$46.36_{\pm 6.86}$	$49.69_{\pm 0.55}$
LR_{Qwen}	$77.13_{\pm 1.35}$	$80.55_{\pm 1.13}$	$77.37_{\pm 1.42}$	$77.01_{\pm 1.51}$
$\underline{LR_{Qwen}}$	$76.26_{\pm 3.03}$	$79.71_{\pm 3.07}$	$76.60_{\pm 3.63}$	$76.17_{\pm 2.47}$
SVC_{Qwen}	$74.80_{\pm 1.47}$	$79.04_{\pm 0.95}$	$75.73_{\pm 1.17}$	$74.27_{\pm 1.89}$
$\underline{SVC_{Qwen}}$	$74.14_{\pm 3.77}$	$78.11_{\pm 3.69}$	$74.76_{\pm 4.46}$	$73.81_{\pm 3.23}$
$Ridge_{Qwen}$	$74.67_{\pm 4.01}$	$78.36_{\pm 3.95}$	$75.06_{\pm 4.66}$	$74.52_{\pm 3.39}$
$\underline{Ridge_{Qwen}}$	$73.17_{\pm 3.32}$	$77.18_{\pm 3.34}$	$73.63_{\pm 3.88}$	$72.99_{\pm 2.84}$
LDA_{Qwen}	$76.46_{\pm 3.00}$	$79.96_{\pm 2.85}$	$76.79_{\pm 3.57}$	$76.27_{\pm 2.59}$
$\underline{LDA_{Qwen}}$	$76.75_{\pm 3.01}$	$80.30_{\pm 2.87}$	$77.26_{\pm 3.69}$	$76.45_{\pm 2.62}$
SGD_{Qwen}	$75.24_{\pm 2.94}$	$79.37_{\pm 3.08}$	$76.65_{\pm 4.39}$	$74.58_{\pm 2.28}$
$\underline{SGD_{Qwen}}$	$73.13_{\pm 4.49}$	$78.61_{\pm 2.94}$	$75.57_{\pm 3.71}$	$72.06_{\pm 4.44}$
RF_{Qwen}	$74.45_{\pm 2.94}$	$79.63_{\pm 2.74}$	$77.52_{\pm 4.56}$	$73.04_{\pm 2.42}$
$\underline{RF_{Qwen}}$	$73.55_{\pm 2.53}$	$78.78_{\pm 2.90}$	$76.97_{\pm 5.40}$	$72.36_{\pm 2.07}$
KNN_{Qwen}	$74.53_{\pm 2.06}$	$79.12_{\pm 1.81}$	$76.06_{\pm 2.61}$	$73.66_{\pm 2.01}$
$\underline{KNN_{Qwen}}$	$74.35_{\pm 1.58}$	$79.12_{\pm 1.45}$	$76.16_{\pm 2.16}$	$73.36_{\pm 1.52}$
DT_{Qwen}	$67.81_{\pm 4.88}$	$71.79_{\pm 5.57}$	$68.72_{\pm 4.76}$	$68.67_{\pm 4.08}$
$\underline{DT_{Qwen}}$	$67.83_{\pm 3.37}$	$72.30_{\pm 3.03}$	$68.14_{\pm 3.54}$	$68.20_{\pm 3.60}$
NB_{Qwen}	$65.53_{\pm 4.02}$	$69.61_{\pm 3.86}$	$65.49_{\pm 3.86}$	$66.24_{\pm 3.93}$
$\underline{NB_{Qwen}}$	$65.50_{\pm 3.83}$	$69.44_{\pm 3.99}$	$65.49_{\pm 3.67}$	$66.27_{\pm 3.56}$
LR_{oss}	$68.33_{\pm 1.54}$	$72.56_{\pm 1.50}$	$68.16_{\pm 1.66}$	$68.59_{\pm 1.41}$
$\underline{LR_{oss}}$	$65.06_{\pm 2.24}$	$69.87_{\pm 1.51}$	$65.10_{\pm 2.03}$	$65.39_{\pm 2.71}$
SVC_{oss}	$66.38_{\pm 2.89}$	$71.80_{\pm 2.54}$	$66.79_{\pm 3.06}$	$66.09_{\pm 2.79}$
$\underline{SVC_{oss}}$	$61.36_{\pm 2.14}$	$66.83_{\pm 1.71}$	$61.45_{\pm 2.11}$	$61.54_{\pm 2.30}$
$Ridge_{oss}$	$67.08_{\pm 2.33}$	$71.38_{\pm 1.94}$	$67.00_{\pm 2.20}$	$67.53_{\pm 2.69}$
$\underline{Ridge_{oss}}$	$62.23_{\pm 1.53}$	$66.58_{\pm 0.84}$	$62.11_{\pm 1.38}$	$62.93_{\pm 1.98}$
LDA_{oss}	$68.15_{\pm 1.77}$	$73.32_{\pm 2.81}$	$69.39_{\pm 3.20}$	$67.95_{\pm 1.25}$
$\underline{LDA_{oss}}$	$65.83_{\pm 2.17}$	$70.71_{\pm 2.08}$	$65.90_{\pm 2.28}$	$65.91_{\pm 2.14}$
SGD_{oss}	$65.10_{\pm 3.68}$	$71.30_{\pm 4.12}$	$66.49_{\pm 5.20}$	$64.53_{\pm 3.12}$
$\underline{SGD_{oss}}$	$63.40_{\pm 1.37}$	$67.59_{\pm 2.22}$	$63.52_{\pm 1.73}$	$64.17_{\pm 1.49}$
RF_{oss}	$63.28_{\pm 3.80}$	$73.32_{\pm 2.13}$	$69.17_{\pm 3.86}$	$62.62_{\pm 3.20}$
$\underline{RF_{oss}}$	$62.26_{\pm 2.59}$	$72.56_{\pm 1.18}$	$67.69_{\pm 1.97}$	$61.69_{\pm 2.27}$
KNN_{oss}	$65.03_{\pm 1.51}$	$72.81_{\pm 1.26}$	$67.92_{\pm 2.06}$	$64.21_{\pm 1.51}$
$\underline{KNN_{oss}}$	$63.01_{\pm 1.38}$	$70.79_{\pm 0.94}$	$65.11_{\pm 1.13}$	$62.52_{\pm 1.52}$
DT_{oss}	$58.19_{\pm 2.68}$	$61.88_{\pm 3.74}$	$58.97_{\pm 2.67}$	$59.71_{\pm 2.83}$
$\underline{DT_{oss}}$	$61.29_{\pm 2.84}$	$65.24_{\pm 2.50}$	$61.06_{\pm 2.74}$	$62.08_{\pm 3.03}$
NB_{oss}	$42.13_{\pm 3.73}$	$43.01_{\pm 3.32}$	$57.39_{\pm 3.48}$	$55.50_{\pm 2.65}$
$\underline{NB_{oss}}$	$41.73_{\pm 3.60}$	$42.34_{\pm 3.13}$	$54.47_{\pm 2.76}$	$53.74_{\pm 2.22}$

Table 11: Complete L1 LR Selection Statistics

Metric	**0.3**	**0.4**	**0.5**	**0.6**
Selection Consistency Across Folds				
Always (5/5 folds)	13 (0.53%)	5 (0.20%)	2 (0.08%)	1 (0.04%)
Highly stable (4/5)	30 (1.22%)	15 (0.61%)	4 (0.16%)	2 (0.08%)
Moderately stable (3/5)	71 (2.88%)	24 (0.97%)	18 (0.73%)	6 (0.24%)
Unstable (1-2/5)	519 (21.1%)	326 (13.2%)	193 (7.8%)	106 (4.3%)
Never selected	1831 (74.3%)	2094 (85.0%)	2247 (91.2%)	2349 (95.3%)
Performance Metrics				
Mean F1	0.736±0.021	0.722±0.022	0.729±0.053	0.699±0.046
Mean Accuracy	0.775±0.010	0.758±0.020	0.727±0.033	0.726±0.046
Per-Fold Feature Count Variation				
Mean # features	211.8	113.2	61.6	31.0
Std # features	129.2	73.3	43.3	21.2
Min features	35	21	11	8
Max features	362	192	112	65
Range (max/min)	10.3×	9.1×	10.2×	8.1×
Hyperparameter Statistics				
C range	0.17-100	0.17-100	0.17-100	0.17-100
C ratio (max/min)	599×	599×	599×	599×
Correlation with Standard LR				
L1-L2 coefficient corr.	-0.026	-0.059	-0.106	-0.249

Table 12: L1 LR Always-Selected Features: Comparison with Standard LR

Threshold	**Feature**	**L1 β**	**L1 SD**	**Std LR β**	**p-value**
0.6	f1914	1.034	0.902	0.061	0.9997
0.5	f1914	1.034	0.902	0.061	0.9997
	f2076	0.409	0.537	0.051	0.9997
0.4	f1914	1.034	0.902	0.061	0.9997
	f1534	-0.936	0.985	-0.061	0.9998
	f1972	0.703	0.455	0.065	0.9998
	f1980	-0.610	0.640	-0.059	0.9998
	f2076	0.409	0.537	0.051	0.9997
0.3	f1914	1.034	0.902	0.061	0.9997
	f1534	-0.936	0.985	-0.061	0.9998
	f598	0.775	0.531	0.044	0.9998
	f1989	0.704	0.654	0.041	0.9998
	f1972	0.703	0.455	0.065	0.9998
	f1821	-0.692	0.744	-0.053	0.9998
	f1784	0.640	0.491	0.057	0.9998
	f1980	-0.610	0.640	-0.059	0.9998
	f440	0.608	0.518	0.061	0.9998
	f2076	0.409	0.537	0.051	0.9997
	f126	0.268	0.310	0.054	0.9998
	f1960	-0.255	0.436	-0.048	0.9998
	f2352	-0.004	0.009	0.011	0.9999

Table 13: L1 LR Per-Fold Performance Breakdown

Threshold	Fold	F1	Acc	# Feat	C	Freq
0.3	1	0.708	0.761	362	100.00	0.155±0.147
	2	0.756	0.777	35	0.17	0.023±0.072
	3	0.763	0.790	85	2.15	0.059±0.095
	4	0.734	0.781	289	100.00	0.144±0.138
	5	0.718	0.768	288	100.00	0.144±0.135
0.4	1	0.705	0.752	192	100.00	0.155±0.147
	2	0.717	0.744	21	0.17	0.023±0.072
	3	0.740	0.765	44	2.15	0.059±0.095
	4	0.696	0.734	169	100.00	0.144±0.138
	5	0.755	0.793	139	100.00	0.144±0.135
0.5	1	0.724	0.761	112	100.00	0.155±0.147
	2	0.671	0.698	13	0.17	0.023±0.072
	3	0.688	0.710	19	2.15	0.059±0.095
	4	0.741	0.768	83	100.00	0.144±0.138
	5	0.823	0.848	82	100.00	0.144±0.135
0.6	1	0.743	0.773	65	100.00	0.155±0.147
	2	0.623	0.655	8	0.17	0.023±0.072
	3	0.713	0.735	14	2.15	0.059±0.095
	4	0.676	0.705	34	100.00	0.144±0.138
	5	0.749	0.781	35	100.00	0.144±0.135

Table 14: Complete L1 SVC Selection Statistics

Metric	0.3	0.4	0.5	0.6
Selection Consistency Across Folds				
Always (5/5 folds)	386 (15.7%)	158 (6.4%)	64 (2.6%)	24 (1.0%)
Highly stable (4/5)	380 (15.4%)	170 (6.9%)	89 (3.6%)	37 (1.5%)
Moderately stable (3/5)	475 (19.3%)	327 (13.3%)	179 (7.3%)	99 (4.0%)
Unstable (1-2/5)	1222 (49.6%)	1729 (70.2%)	1516 (61.5%)	928 (37.7%)
Never selected	1 (0.04%)	80 (3.2%)	616 (25.0%)	1376 (55.8%)
Performance Metrics				
Mean F1	0.769±0.015	0.746±0.027	0.744±0.020	0.751±0.032
Mean Accuracy	0.805±0.011	0.785±0.018	0.781±0.015	0.786±0.024
Per-Fold Feature Count Variation				
Mean # features	1340.6	791.6	564.0	341.6
Std # features	846.3	746.4	649.9	369.0
Min features	1031	578	311	149
Max features	2463	2369	1772	974
Range (max/min)	2.4×	4.1×	5.7×	6.5×
Hyperparameter Statistics				
C range	27.8-100	27.8-100	27.8-100	27.8-100
C ratio (max/min)	3.6×	3.6×	3.6×	3.6×

Table 15: L1 SVC Per-Fold Performance Breakdown

Threshold	Fold	F1	Acc	# Feat	C	Freq
0.4	1	0.702	0.756	610	27.83	0.311±0.172
	2	0.785	0.811	2369	100.00	0.579±0.128
	3	0.743	0.777	597	27.83	0.307±0.168
	4	0.753	0.789	604	27.83	0.309±0.163
	5	0.749	0.789	578	27.83	0.306±0.161

Table 16: Top 25 Positively and Top 25 Negatively Weighted Reasoning Questions selected from the LEPREC linear model for the usability study.

#	Weight	Reasoning Question	Rationale
Top 25 Positively Weighted Questions (indicative of relevance)			
1	+	Has the court explicitly stated that the issue would not affect its determination on the merits of the case?	When a court explicitly states that an issue does not impact its decision on the merits, it often indicates the issue's irrelevance to the core dispute.
2	+	Does the issue address the core legal question raised by the parties' actions?	The relevance of an issue is often determined by its relationship to the fundamental legal questions arising from the parties' actions and claims.
3	+	Does the issue address the main point of contention in the dispute as described in the facts?	This confirms whether the issue captures the essence of the legal controversy, which is crucial for establishing its relevance to the case.
4	+	Does the issue require an examination of the specific terms or conditions of the agreement in light of the alleged illegality?	This determines if the issue necessitates a detailed analysis of the agreement's terms in relation to illegal activities, often central to determining contract validity.
5	+	Did the court indicate that this issue would not affect the determination on the merits of the case?	When a court explicitly states that an issue does not impact the case's outcome, it often suggests the issue is not central to the main dispute.
6	+	Is the issue central to determining the legality or propriety of the proceedings described in the facts?	This establishes whether the issue is fundamental to assessing the overall validity of the legal process.
7	+	Does the issue address the key elements that led to the court's finding in the scenario?	Evaluating whether the issue relates to the court's reasoning helps determine its significance to the case's outcome.
8	+	Is the issue focused on a specific legal conclusion rather than a general legal principle?	Specific legal conclusions tied to the case are more likely to be directly relevant than general legal principles.
9	+	Is the issue related to the legality or enforceability of an agreement mentioned in the facts?	Assessing the legal validity of an agreement is often at the core of contractual disputes.
10	+	Does the issue address a key point of contention that was ruled upon by the courts mentioned in the facts?	This checks whether the issue aligns with the core legal matters judicially considered.
11	+	Does the issue relate to the legality or enforceability of the agreement described in the facts?	This checks whether the issue touches on the legal validity of the agreement, often a key factor in determining relevance.
12	+	Does the issue address the core dispute that led to legal proceedings?	This determines whether the issue is at the heart of the conflict that resulted in the case.
13	+	Does the issue pertain to a matter that could affect the rights or obligations of the parties involved?	This assesses whether the issue has practical implications for the parties, crucial for establishing relevance.
14	+	Is the issue related to the primary action or counteraction taken by the parties in the scenario?	This establishes whether the issue is connected to the main moves or decisions made by the parties.
15	+	Does the issue directly address the primary legal dispute in the scenario?	The relevance of an issue depends on whether it addresses the core legal dispute rather than peripheral matters.
16	+	Does the issue involve a decision made by the highest court mentioned in the scenario?	Whether the highest court's decision is involved often represents the final word on the matter.
17	+	Is the issue focused on a specific legal remedy or action that could be taken by the court?	This assesses whether the issue involves a concrete legal action, often indicating direct relevance.
18	+	Is the issue related to a potential legal violation or illegal act mentioned in the facts?	Issues involving matters of legality are often central to determining the outcome of a case.
19	+	Is the issue related to a specific legal claim or defense mentioned in the facts?	Issues pertaining to specific legal claims or defenses raised by the parties are often highly relevant.

Continued on next page.

Table 16 continued.

#	Weight	Reasoning Question	Rationale
20	+	Is the issue directly connected to the outcome or judgment described in the scenario?	This assesses whether the issue is linked to the consequences or rulings that resulted from the actions in question.
21	+	Does the issue pertain to a matter that is no longer in contention in the current legal proceeding?	Issues no longer actively disputed may be less relevant to the current stage of proceedings.
22	+	Is the issue related to a specific clause or condition that is being contested by the parties?	An issue focused on a specific, contested clause is more likely to be central to the dispute.
23	+	Does resolving this issue potentially affect the rights and liabilities of all parties involved?	An issue impacting the rights and liabilities of all parties is likely at the heart of the legal controversy.
24	+	Is the issue framed in a way that directly challenges or supports the main finding described in the scenario?	Aligning the issue with the central finding reinforces its relevance to the core matter.
25	+	Does resolving this issue directly impact the outcome of the disciplinary or legal proceedings?	Issues with a direct bearing on the case's outcome are typically highly relevant.
Top 25 Negatively Weighted Questions (indicative of irrelevance)			
26	−	Has the issue already been conclusively decided by a higher court in the scenario?	The finality of a higher court's decision can affect the relevance of an issue in ongoing proceedings.
27	−	Does resolving this issue definitively settle the legal dispute between the parties?	An issue's ability to conclusively resolve the dispute is a strong indicator of its relevance.
28	−	Is there a direct financial consequence mentioned in relation to the issue?	Financial implications often indicate that an issue is material rather than incidental.
29	−	Does the issue relate to the legal process or procedural aspects of the case?	Procedural matters can be significant, but may not address the substantive coverage.
30	−	Is the issue specifically mentioned in the court's reasoning for its decision?	Courts' explicit reasoning often indicates which issues were central to their decisions.
31	−	Is the issue related to the insurance companies' reason for denying the claim?	The primary controversy revolves around the denial reason, not the documentation process.
32	−	Is the issue related to the financial obligations outlined in the agreement?	Financial obligations are frequently at the core of contractual disputes.
33	−	Is the issue related to a specific professional standard or ethical requirement mentioned in the facts?	Connecting the issue to professional standards helps establish relevance in professional misconduct cases.
34	−	Does the issue relate to the interpretation of evidence presented in the case?	Issues involving evidence interpretation are typically relevant to the central matters of a dispute.
35	−	Is the issue connected to the economic impact of the terms described in the scenario?	This explores whether the issue has direct financial implications related to the facts.
36	−	Is the issue rendered moot by subsequent developments in the case?	An issue may become irrelevant if superseded or rendered moot by later rulings.
37	−	Is the issue about interpreting a specific clause in the agreement?	Understanding whether the issue involves contract interpretation is crucial for assessing relevance.
38	−	Does the issue concern the handling of funds in a professional capacity?	This focuses on whether the issue involves financial responsibilities, often central to professional conduct cases.
39	−	Is the issue related to the interpretation of a specific contractual term?	Contractual interpretation issues are often core to legal disputes.
40	−	Is the issue a fundamental legal principle that applies to all similar cases?	Fundamental legal principles may not always be directly relevant to the specific controversy.

Continued on next page.

Table 16 continued.

#	Weight	Reasoning Question	Rationale
41	–	Does the issue involve a challenge to the enforceability of a contractual provision?	Challenges to enforceability are typically central to contract disputes.
42	–	Is the issue connected to the economic impact of the contract terms on the parties?	Economic consequences of contract terms often form the heart of contractual disputes.
43	–	Is the issue tied to a specific request for relief mentioned in the facts?	Connection to requested relief suggests the issue is central to resolving the dispute.
44	–	Does the issue involve the interpretation or application of a relevant law or regulation?	Assessing whether the issue involves legal interpretation helps determine its significance.
45	–	Is the issue about interpreting a specific clause in the contract?	Whether the issue focuses on contract interpretation is crucial for assessing its relevance.
46	–	Is the issue tied to the legal standard or rule applied in the case?	Linking the issue to the applicable legal standard helps establish its significance.
47	–	Does the issue directly address the enforcement of a contractual agreement mentioned in the facts?	This establishes whether the issue is directly related to the core contractual dispute.
48	–	Is the issue related to the main financial transaction or arrangement in dispute?	Connecting the issue to the primary financial matter helps establish its centrality.
49	–	Does resolving this issue alone determine the outcome of the legal dispute?	An issue's ability to singularly resolve the case is a key factor in assessing legal relevance.
50	–	Is the issue focused on interpreting specific terms of the agreement in question?	Issues centered on interpreting specific contractual terms are typically more relevant than abstract principles.