Information-Theoretic Legal Issue Identification and Reward Modeling on Court Cases

Anonymous ACL submission

Abstract

Legal issue identification is a crucial first step in legal analysis, yet more than half of people worldwide struggle to meet their civil justice needs. While Large Language Models (LLMs) have shown promise in various application domains, their effectiveness in identifying legal issues from real-world court cases remains understudied. Previous evaluations have been limited to simplified scenarios or textbook examples, lacking the complexity of actual cases. To address this gap, we present LIC, a dataset of 769 real-world court cases pertinent to Contract Act Malaysia, with facts and legal issues extracted using GPT-4o and validated by top law students and junior lawyers. We propose a novel approach that generates and ranks legal issue candidates by incrementally incorporating case facts and propose a novel reward model based on mutual information (MI) for reranking. Our method uses a soft-threshold function to align MI with estimated relevance between issue candidates and facts during training. Experimental results demonstrate our methodology's superior performance compared to baselines on our test set. This work advances automated legal issue identification while providing a substantial dataset for future research in legal AI. Our dataset and the source code will be publicly available upon acceptance.

1 Introduction

According to the survey conducted by the World Justice Project with 100,000 survey participants in 101 countries, more than half are unable to meet their civil justice needs (Camilo Gutiérrez Patiño et al., 2019). *Identification of legal issues* in a given set of facts is a crucial first step in legal analysis, where an issue is a question or problem that requires the application of legal principles and law to determine the existence of rights, duties, or remedies (Stockmeyer, 2021; Kang et al., 2023). This requires expertise, nuanced reasoning, and a thorough understanding of complex legal scenarios.

The emergence of large language models (LLMs) has transformed fields that require deep domain expertise. However, to our knowledge, only two studies (Guha et al., 2023; Kang et al., 2024) evaluate LLMs and NLP models to identify legal issues. Guha et al. (2023) simplify the task as legal question-answering tasks, while Kang et al. (2024) evaluate LLMs in zero-shot or few-shot settings on approximately 50 scenarios sourced from textbooks or curated by law students, which lack the complexity of real-world court cases. Hence, three research questions (RQs) remain open: RQ1) How effectively can LLMs identify issues in real-world court cases? RQ2) How can training on legal cases further enhance their performance in this task?, and RQ3) To what extent can LLMs assess the quality of issues generated based on facts in court cases.

Despite its importance, progress in automating issue identification in court cases is hampered by a lack of legal datasets derived from real-world cases. Prior works (Guha et al., 2023; Kang et al., 2024) rely on legal scenarios that are not sourced from actual court cases, with their corpora containing fewer than 60 legal scenarios in total, which are suitable only for model testing. This scarcity of annotated data complicates the development and evaluation of task-specific models, posing a significant barrier to advancing research in this area.

To address the first two research questions, we curate a legal issue dataset on court cases, coined LIC, which comprises 769 real-world court cases pertaining to Contract Act Malaysia. We apply GPT-4o to extract facts and legal issues from these cases, which serve as silver ground-truth, and assess the quality of the extraction through evaluations by legal professionals. The results indicate that only a handful of extracted data do not meet the expectation. We also include human-annotated issues as the ground-truth for evaluation.

Different issues may provide different hints to users with varying legal background. A legal issue often pertains to only a subset of facts mentioned in a case, thus our corpus poses a new challenge for fine-tuning LLMs in that the relationships between facts and issues are unknown. Moreover, repeated sampling from LLMs can significantly increase the likelihood of obtaining correct responses (Brown et al., 2024) and irrelevant model inputs can do more harm than good (Feng et al., 2023).

To address these challenges, we propose generating a pool of legal issue candidates by incrementally incorporating individual facts from a case into an LLM and then ranking all candidates based on the mutual information (MI) (Ash, 2012). Due to the size of our dataset, we focus on reward modeling for ranking. During training, we propose to train the reward model by maximizing MI and aligning it with the estimated relevance between issue candidates and facts, achieved by applying a soft-threshold function (Feng et al., 2024) to the corresponding MI loss terms. Our extensive experiments indicate that:

- Our MI inspired model significantly outperforms competitive baselines and demonstrates the effectiveness of our sparsity-motivated loss function, MI estimation, and our incremental issue sampling strategy.

- Our method perform better on human-annotated ground-truth than LLM-annotated silver ones, which shows a strong alignment with the judgments of legal experts.

2 LIC Corpus

LIC corpus is built by collecting 769 real-world court cases in the area of Contracts Act Malaysia. The facts and issues in those cases are extracted automatically by GPT-4o and further validated by top law students and junior lawyers.

Court Case Collection. We construct the corpus pertaining to **illegality under Section 24 of the Contracts Act Malaysia** and the **formation of contracts**, due to their importance in Contract Law. The cases are selected from the *Current Law Journal* (CLJ)[1], 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.

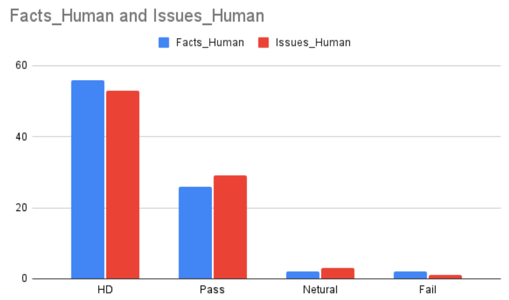

Figure 1: Evaluation results on fact and issue extraction.

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.

Issue and Fact 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 Sec. 4.1). Herein, the automatic 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., 2024a) and use the styles recommended in (Lin et al., 2023) for prompt design. The resulting prompts are illustrated in Sec. A.4 in Appendix. Specifically, we apply GPT-4o 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 Sec. A.12 of Appendix.

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

Structured ratings and detailed comments are provided for each element to assess.

As illustrated by Fig. 1, 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.

3 Issue Identification via Ranking

We formulate issue identification as a generation-augmented ranking problem because (i) issues generated by LLMs provide diverse hints to end users with varying legal backgrounds, (ii) the LLMs we evaluate are more reliable at comparing the quality of issues than at making decisions on issue selection (Paul et al., 2023; Tang et al., 2023), and (iii) the size of LIC is large enough for reward modeling but insufficient for fine-tuning LLMs on issue generation. Therefore, we focus on devising a reward model inspired by MI to evaluate issue candidates, which are sampled from an LLM by incrementally adding individual facts sorted in temporal order.

3.1 Issue Candidate Generation

Inspired by (Feng et al., 2023), it is desirable for LLMs to take as input only sufficient and necessary information. The irrelevant input data is the cause of spurious correlations. Therefore, we generate issue candidates by adopting an incremental strategy.

Given a list of facts $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$, we incrementally generate a set of issue candidates $\hat{\mathcal{Y}} = \{\mathcal{Y}_1, ..., \mathcal{Y}_n\}$ using an LLM by taking the following steps:

1. Apply the LLM based on $\mathbf{x}_1$ to generate a list of issue candidates $\hat{\mathcal{Y}}_1$. Then $\hat{\mathcal{Y}} = \hat{\mathcal{Y}}_1$.
2. Apply the LLM based on $[\mathbf{x}_1, \mathbf{x}_2]$ to generate a list of issues $\hat{\mathcal{Y}}_2$. Then $\hat{\mathcal{Y}} = \hat{\mathcal{Y}} \cup \hat{\mathcal{Y}}_2$.
3. Repeat the above step by incrementally adding another fact to generate issue candidates to extend the set $\hat{\mathcal{Y}}$ until all facts are used.

Under different depths of given context information, LLMs are required to focus on different levels of information, which facilitates the discovery of new issue candidates. The prompt in this generation process is generated by GPT4o and further polished with Claude (See Appendix.A.10.1).

3.2 Issue Candidate Scoring

The generated issue candidates in the incremental generation process may contain irrelevant ones. Thus, considering precisely justifying and criticizing the quality of issue candidates, we introduce a sparsity-motivated reward model for predicting scores to reflect the quality of each issue candidate.

Given a set of issue candidates $\hat{\mathcal{Y}} = \{\mathcal{Y}_1, \ldots, \mathcal{Y}_n\}$, we predict a score for each candidate by estimating its relevance to the given facts. Mutual information (MI), an information-theoretic concept widely applied across various domains (Sordoni et al., 2021; Mukherjee et al., 2020), quantifies the dependence between two random variables (Effenberger, 2013; Jain and Murthy, 2014). Instead of directly predicting a score, we introduce the concept of mutual information estimation, where ***a higher mutual information value signifies a stronger relationship between the issue candidate and the given facts***. This approach effectively frames the scoring process as a measure of relatedness. Specifically, the scoring function $s(\mathbf{X}; \mathcal{Y}_j)$ is defined as:

$$s(\mathbf{X}; \mathbf{c}_j) = I(\mathbf{X}; \mathcal{Y}_{\mathrm{j}}) \quad (1)$$

where the score of j-th issue candidate given facts $\mathbf{X}$ is calculated based on the mutual information between the facts and issue. Considering that the given facts are formulated in a pointed manner, we compute the mutual information based on the chain rule (Sparavigna, 2015), as follows:

$$I(\mathbf{X}; \mathcal{Y}_j) = I(\mathbf{x}_1; \mathcal{Y}_j) + \sum_{i=2}^{m} I(\mathbf{x}_i; \mathcal{Y}_j | \mathbf{x}_{i-1}, ..., \mathbf{x}_1) \quad (2)$$

where the mutual information consists of two parts, i) the sum of (non-conditional) mutual information on the first facts $I(\mathbf{x}_1; \mathcal{Y}_j)$ and ii) conditional mutual information on the following facts $\sum_{i=2}^{m} I(\mathbf{x}_I; \mathcal{Y}_j | \mathbf{x}_{i-1}, ..., \mathbf{x}_1)$ with the candidate.

Sparsity of Mutual Information. We observe that some early generated issues may reappear when we feed the following facts in the incremental generation process. This situation suggests that i) ***not all issue candidates relate to an issue*** and ii) ***the increasing gradient of (conditional) mutual information varies within the incremental issue generation process***. More specifically, if $\{\mathbf{x}_1, ..., \mathbf{x}_{i-1}\}$ already provides sufficient and necessary information for $\mathcal{Y}_j$, $I(\mathbf{x}_i; \mathcal{Y}_j | \mathbf{x}_{i-1}, ..., \mathbf{x}_1)$ should be close to zero, where conditional independence is applied to $p(\mathcal{Y}_j, \mathbf{x}_i | \mathbf{x}_{i-1} ..., \mathbf{x}_1)$. In another word, if an issue $\mathcal{Y}_j$ depends only on the fact sequence up to $\mathbf{x}_t$. The MI terms from t to m, almost independent of the issue candidate, should be close to zero (See details in Appendix. A.3). To enforce such a constraint, we apply the soft threshold function to MI terms following (Xu and Cheung, 2019; Feng et al., 2024).

Mutual Information Approximation Given the input facts $\mathbf{X}$ and issue candidates $\mathcal{Y}_j$, we use semantic entropy to estimate the mutual information following Kuhn et al., 2023. The semantic entropy estimation can examine the uncertainty in the mean space, which is defined as:

$$\begin{aligned} & I(\mathbf{x}_i; \mathcal{Y}_j | \mathbf{x}_{i-1}, ..., \mathbf{x}_1) \\ & = H(\mathcal{Y}_j | \mathbf{x}_{i-1}, ..., \mathbf{x}_1) - H(\mathcal{Y}_j | \mathbf{x}_i, \mathbf{x}_{i-1}, ..., \mathbf{x}_1) \\ & \approx -\frac{1}{|\mathbf{y}'|} \log p(\mathbf{y}' | \mathbf{x}_{i-1}, ..., \mathbf{x}_1) \\ & + \frac{1}{|\mathbf{y}'|} \log p(\mathbf{y}' | \mathbf{x}_i, \mathbf{x}_{i-1}, ..., \mathbf{x}_1) \end{aligned} \tag{3}$$

$$\begin{aligned} & I(\mathbf{x}_1; \mathcal{Y}_j) \\ & = -H(\mathcal{Y}_j | \mathbf{x}_1) + H(\mathcal{Y}_j) \\ & \approx -\frac{1}{|\mathbf{y}'|} \log p(\mathbf{y}' | \mathbf{x}_1) + \frac{1}{|\mathbf{y}'|} \log p(\mathbf{y}') \end{aligned} \tag{4}$$

The probability relies on the confidence estimated by an LLM. As $\mathbf{y}'$ has variable length, we adopted $-\frac{1}{|\mathbf{y}'|} \log p(\mathbf{y}' | \mathbf{x}_{i-1}, ..., \mathbf{x}_1)$ to compute the perplexity of $\mathbf{y}'$ conditioned on different input, following a similar idea (Meng et al., 2024). Additionally, in Equation. 4, we found that the application of $\frac{1}{|\mathbf{y}'|} \log p(\mathbf{y}')$ will negatively affect the learning process, which is further discussed in ablation study (Sec. 4.3). Thus, we only apply $\frac{1}{|\mathbf{y}'|} \log p(\mathbf{y}')$ in inference process.

3.3 Training with Soft-Threshold Function.

We introduce a sparsity-motivated training loss defined below to learn the reward model, because not all MI terms in Eq. (2) indicate valid relationships between issues and the corresponding facts. Thus, the soft-threshold function $g(\cdot)$ defined below is chosen to automatically select the MI terms for parameter updating.

$$\begin{aligned} & \min -\{\sum_{i=2}^{m} g(|I(\mathbf{x}_i; \mathbf{c}_j | \mathbf{x}_{i-1}, ..., \mathbf{x}_1)| - \tau) \\ & - g(|I(\mathbf{x}_1; \mathbf{c}_j)| - \tau)\} + \exp(-\tau) \end{aligned} \tag{5}$$

where

$$\frac{d}{dt} g(v) = \begin{cases} 2 - 4|v|, & -0.4 \leq v \leq 0.4 \\ 0.4, & 0.4 \leq |v| \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The sparsity is achieved when the MI terms with small absolute values yield zero so that they are not used to update model parameters.

Forward Data Sampling. As a result of the incremental issue-generation process, the number of issue candidates increases significantly compared to one-time issue-generation. Similar issues might appear when we provide more context information. Issues generated in later iterations, if similar to earlier ones, serve as forward versions of their counterparts from previous iterations, which can provide a more comprehensive and deeper understanding of some concerns. Thus, the sampling process, named forward data sampling, is employed as follows:

1. The silver ground truth issues and generated issue candidates are mixed into an issue pool.
2. Sentence Transformer (Reimers and Gurevych, 2019), a sentence embedding model, is introduced for estimating the similarity among the issues.
3. Non-duplicated issues are sampled based on the similarity matrix. For each pair of similar issues, the issue generated in the **later** incremental iteration is selected.[2]

4 Experiments

The architecture of our legal judgment system comprises three stages: (i) incremental issue generation based on an LLM, (ii) issue rewarding, and (iii) issue ranking. As outlined in the previous sections, the key process ensuring the quality of legal judgments is stage (ii), which employs mutual information (MI) estimation.

Accordingly, to answer the RQs in Sec. 1, in the experiment, we first present the main results comparing with baselines and ablation study comparing with alternative methods on two evaluations, as

[2] If a silver ground truth issue is similar to a generated issue, the generated issue is selected, as it is more conducive to machine understanding when produced by a LLM.

human- and LLM-annotated testset, then give our answer to RQ1, RQ2, and RQ3 in Appendix. A.1.

4.1 Experimental Setup

Construction of Ground Truth While issue identification strongly relies on the understanding of the given case, the explanation and understanding of legal cases can be very subjective (Strębska, 2013; Klaasen, 2017). In the real scenario, the solution of legal cases can be expressed in various ways (Kong et al., 2019; Rotolo and Sartor, 2023), which indicates that the ground truth in our legal corpus is not the only solution to understanding the cases. Thus, to best monitor the actual legal analysis process, we conducted a human annotation for our testset and ran our experiments on the annotated testset.

Human Annotation. With respect to the above, we assigned three human annotators to annotate the sampled testset. Each annotator was asked to assess the following question:

> "*Given legal **Facts**, **Issue A**, and **Issue B**, please determine whether **Issue B** is **Similar** to or a **Paraphrase** of **Issue A** within the context of the given **Facts**.*"

wherein each of the legal cases in our testset consists of i) background facts, ii) ground truth issues, and iii) generated issue candidates. As we consider the ground truth as one of the solutions for the given case, we ask the annotators to judge if the generated issue candidates perform in a similar meaning with the ground truth issues. If the annotator judged the two issues to be similar—meaning that the generated issue closely resembled or paraphrased the ground truth issue—it was marked as a positive issue. In total, 200 issue pairs are sampled from the testset, then annotated by three annotators hired from Prolific[3] for 9 euro per hour by selecting CommonWealth Member Countries [4], with Fleiss' Kappa yielding a value of 0.6707, where, considering the difficulty of legal issue judgment, the Fleiss' Kappa value in our human annotation can achieve substantial agreement. We named the human-annotated results as `Test-I` in the following description.

LLM Annotation. Similarly, considering the advanced legal reasoning ability of Claude, we employ Claude as an annotator to compare the ground truth issues and generate issue candidates that follow the process of Human Annotation. We adopted the prompt generated by GPT4o and polishing by Claude, which are detailed in Sec. A.11. We use the LLMs annotated testset as the second testset, named `Test-II`.

Datasets. We formulate the dataset in the form of the IRAC (Burton, 2017; Kang et al., 2023), where each legal case consists of (i) facts (scenario), (ii) issues, (iii) rules, (iv) analysis, and (v) conclusion. Here, we only use the facts and issues parts. We generate the issue following our incremental generation strategy in section 3. We perform the incremental generation on 769 legal cases, which consist of 3,352 points of facts, 2,566 ground truth issues (written by human annotators), and 11,065 generated issues. We use 669 cases as our trainset, 50 cases for validation, 200 issue pairs in `Test-I`, and 50 cases in `Test-II`.

Evaluation Metrics While there are two labels in the annotated issues, as positive or negative issues, we employ the reward models in the context of the corresponding facts to predict scores for all generated issue candidates. By further sorting the scores in descending order, the evaluation for the reward models can be converted to a ranking task, where the positive issues should be ranked in higher positions. Thus, following the previous work in the recommendation system, keyword identification, and RAG studies (Martinc et al., 2020; Liu et al., 2023; Yuan et al., 2024b), we adopted the measures for ranking task, including Precision@K, Recall@K, F1-Score@K, nDCG@K, and MAP@K. We opted for the 4 settings of K in our experiments as $[1, 5, 10, 20]$ to comprehensively reflect the results.

Baseline Models We consider the reward model and LLMs as judges to serve as the baseline.

Reward Models i) Prometheus is an open-source language model specially designed for evaluation (Kim et al., 2024). We employed the relative rewarding mode of Prometheus 2 [5] as a strong baseline in our experiment (named "Prometheus" in result table). ii) generative verifier (Zhang et al., 2024) is a simple but effective reward model, where the reward task is further simplified to the next token prediction on "yes" and "no" tokens. We fine-tuned the generative verifier on our trainset by treating the ground truth issues and their similar issues as positive samples and the other issues

[3] https://www.prolific.com/

[4] https://thecommonwealth.org/our-member-countries

[5] https://github.com/prometheus-eval/prometheus-eval

Table 1: Main Results on Test-I (Human-Annotated)

Methods	MAP	@K=1					@K=5					@K=10					@K=20				
		P	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**
Claude	38.52	12.50	4.17	6.25	12.50	12.50	22.50	32.50	26.00	25.41	62.50	28.75	84.82	41.08	50.12	**100.00**	**19.38**	**100.00**	31.40	57.34	**100.00**
GPT4o	42.76	12.50	4.17	6.25	12.50	12.50	27.50	42.92	32.70	32.72	62.50	26.25	81.25	38.14	50.99	87.50	**19.38**	**100.00**	**31.40**	59.97	**100.00**
Prometheus	52.57	12.50	4.17	6.25	12.50	12.50	40.00	65.95	46.85	49.65	100.00	30.00	86.90	42.65	60.55	**100.00**	**19.38**	**100.00**	**31.40**	66.92	**100.00**
UR3	63.05	50.00	22.62	28.13	50.00	50.00	42.50	63.27	47.44	58.48	100.00	**36.25**	91.67	49.48	72.73	**100.00**	**19.38**	**100.00**	**31.40**	75.86	**100.00**
Gen.Veri.$_{\text{-CoT}}$	*56.64*	*37.50*	*17.08*	*20.24*	*37.50*	*37.50*	*40.00*	*60.03*	*44.78*	*53.57*	***100.00***	*32.50*	*90.48*	*45.59*	*67.87*	***100.00***	***19.38***	***100.00***	***31.40***	*72.42*	***100.00***
Gen.Veri.$_{\text{+CoT}}$	*70.44*	*75.00*	*29.29*	*37.95*	*75.00*	*75.00*	*42.50*	*61.07*	*46.55*	*65.18*	***100.00***	***36.25***	*91.96*	***49.57***	*79.67*	***100.00***	***19.38***	***100.00***	***31.40***	*82.65*	***100.00***
Gen.Veri.$_{\text{-CoT}}$	52.47	37.50	8.75	13.99	37.50	37.50	40.00	57.95	44.34	49.12	87.50	31.25	88.69	44.12	62.98	**100.00**	**19.38**	**100.00**	**31.40**	68.49	**100.00**
Gen.Veri.$_{\text{+CoT}}$	62.41	37.50	19.17	22.92	37.50	37.50	47.50	69.11	52.61	62.83	**100.00**	31.25	84.23	43.50	69.64	**100.00**	**19.38**	**100.00**	**31.40**	76.26	**100.00**
Ours	*57.87*	*50.00*	*15.00*	*22.32*	*50.00*	*50.00*	*45.00*	*59.40*	*49.31*	*55.85*	*87.50*	*33.75*	***92.56***	*47.15*	*70.04*	***100.00***	***19.38***	***100.00***	***31.40***	*73.66*	***100.00***
Ours	**72.08**	**87.50**	**33.45**	**44.20**	**87.50**	**87.50**	**52.50**	**74.61**	**57.85**	**74.67**	**100.00**	33.75	**92.56**	47.15	**81.43**	**100.00**	**19.38**	**100.00**	**31.40**	**85.09**	**100.00**

Table 2: Main Results on Test-II (LLM-Annotated)

Methods	MAP	@K=1					@K=5					@K=10					@K=20				
		P	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**
Claude	40.53	0.00	0.00	0.00	0.00	0.00	32.50	49.06	36.33	30.56	87.50	32.50	83.75	44.26	48.81	**100.00**	**23.75**	**100.00**	**36.44**	57.54	**100.00**
GPT4o	44.52	12.50	2.50	4.17	12.50	12.50	37.50	56.37	41.90	40.26	**100.00**	30.00	82.86	41.74	52.86	**100.00**	22.50	**100.00**	34.87	61.94	**100.00**
Prometheus	55.91	37.50	11.87	17.50	37.50	37.50	**45.00**	54.40	46.97	53.50	87.50	32.50	87.14	44.34	64.75	**100.00**	21.88	**100.00**	33.97	71.62	**100.00**
UR3	56.69	37.50	10.31	15.28	37.50	37.50	**45.00**	46.56	43.98	49.24	75.00	38.75	90.63	**50.59**	66.67	**100.00**	21.25	**100.00**	33.35	70.49	**100.00**
Gen.Veri.$_{\text{-CoT}}$	*47.39*	*12.50*	*6.25*	*8.33*	*12.50*	*12.50*	*40.00*	*46.56*	*41.53*	*42.26*	*87.50*	***33.75***	*89.58*	*46.63*	*60.19*	***100.00***	*20.63*	***100.00***	*32.97*	*65.32*	***100.00***
Gen.Veri.$_{\text{+CoT}}$	*49.92*	*37.50*	*9.90*	*15.28*	*37.50*	*37.50*	*40.00*	*62.19*	*44.75*	*50.30*	*87.50*	*33.75*	***96.88***	*46.29*	*64.51*	***100.00***	*18.13*	***100.00***	*29.09*	*66.20*	***100.00***
Gen.Veri.$_{\text{-CoT}}$	52.44	**50.00**	12.20	19.37	50.00	**50.00**	37.50	45.98	40.63	45.58	62.50	30.00	80.58	41.60	59.77	**100.00**	20.63	**100.00**	32.85	68.60	**100.00**
Gen.Veri.$_{\text{+CoT}}$	54.14	**50.00**	11.58	18.54	50.00	**50.00**	40.00	46.19	41.55	48.92	75.00	30.00	80.80	41.50	61.42	**100.00**	20.62	**100.00**	32.78	70.13	**100.00**
Ours	*56.20*	***50.00***	*11.16*	*18.12*	*50.00*	***50.00***	***45.00***	***68.38***	***50.73***	*57.17*	***100.00***	*35.00*	*91.96*	*48.02*	*68.21*	***100.00***	*20.63*	***100.00***	*32.85*	*72.15*	***100.00***
Ours	**65.12**	**50.00**	**21.88**	**27.32**	**50.00**	**50.00**	**45.00**	64.58	49.65	**63.75**	**100.00**	31.25	88.02	44.05	**73.22**	**100.00**	20.00	**100.00**	32.14	**79.16**	**100.00**

as negative samples in cross-entropy loss. Two types of prompt settings are used, including no CoT (named "-CoT" in result tables) and using CoT (named "+CoT" in result tables). Here, we report both untrained and trained versions of the generative verifier. Here, based on the results of a fine-tuned generative verifier on cross-entropy loss, we aim to show if the legal issue identification task can be simply regarded as a classification task. iii) UR3 (Yuan et al., 2024b) is a ranking model designed for RAG (Retrieval-Augmentation Generation), where UR3 can predict a score for each document. We treat the issue as a document to employ this method as our baseline (named "UR3" in result tables).

Large Language Models: Given the surprising reasoning abilities demonstrated by various large language models (LLMs) across different domains, we treat LLMs as strong baselines for comparison with our methods in both reward model designs. Specifically, following the idea of LLMs as a judge (Zheng et al., 2023), we selected GPT4o (Hurst et al., 2024) and Claude[6] to serve as the baselines in LLM as a judge method (named "GPT4o" and "Claude" in result tables).

4.2 Results and Discussions

Comparison with Baselines. Table. 1 and Table. 2 respectively show the results of the Test-I (human-annotated ground truth) and Test-II (LLM-annotated ground truth), where our method achieves dominant leading in 19 out of 21 measures in Test-I and 14 out of 21, which shows the effectiveness of our rewarding capability compared with other baselines. Besides, among Test-I and Test-II, our results on human-annotated evaluation are generally better than the performance on LLM-annotated evaluation, which indicates that our method has higher alignment with human judgment rather than machine judgments.

MI based Inference. *Italics* rows are the untrained version of the generative verifier, and our method and underscore results are the best in the untrained version. Even if our strategies are mostly designed for training, we still achieve i) comparable performance with untrained generative verifiers in two CoT settings, where we beat the generative verifier on five measures against nine measures better from generative verifier in Test-I, and we have 19 better or equal performance compared with the generative verifier in Test-I. CoT is an informative additional input for generative LLMs, known as a systematic process where an individual or model explicitly breaks down a problem into smaller, manageable steps, leading to a final solution (Zhang et al., 2023). With the help of our design of mutual information inference, our untrained version can achieve overall better performance in two test sets even without the CoT decoration, which demon-

[6] https://www.anthropic.com/news/claude-3-5-sonnet

strates the effectiveness of the designing of MI estimation.

Sparsity-Motivated Training. The training strategies in our method are emphasized in previous sections. The sparsity-motivated training does not directly optimize some absolute labels. We observed that the trained generative verifier performed no gain but even worse than its untrained counterpart. This outcome suggests that the legal identification task, as a novel and complex task, cannot simply be treated as a traditional reward or ranking task that relies on positive and negative inputs (Sybrandt et al., 2020; Cai et al., 2022), which needs more appropriate objective designing on a higher dimension. Here, our sparsity-motivated training enables smoother learning objectives, optimizing the mutual information between the issue and independent facts. The results have shown that after training with sparsity objectives, our model gained great performance improvement, with 100% improvement or equal among all the measures.

4.3 Ablation Study

Ablation Study Settings. In the ablation study, we ensured consistency across all training and testing presets and hyperparameters, modifying only a single setting at a time to determine the most effective combination. Our method consists of three key components: i) mutual information estimation, ii) the soft-threshold function, and iii) forward data sampling. Based on these, we designed eight alternative methods for evaluation. **-CMI** refers to the first set, where we only use perplexity to replace the conditional mutual information estimation in soft-threshold. **+H($\mathcal{Y}$)** is the alternative mentioned in Sec. 3.2, wherein our method we cancel $H(\mathcal{Y})$ in the training and only apply it in inference process. **-CMI +H($\mathcal{Y}$)** is the combination of the first two alternatives, which aims to show if the negative effects come from the cross-influence between CMI and **+H($\mathcal{Y}$)**. **-SoftThres.** belongs to the second set, which removes the soft-threshold function in training. **Full Data**, **-Rein.Sam.**, **Rand.Sam.**, and **Reve.Sam.** are corresponding to the third set. **Full Data** removes similarity checking and sampling process and enables full input of issues, which may contain more but duplicated issues in the trainset. **-Rein.Sam.** is similar to **Full Data** but has a constraint with the same amount of issues with the results of forward data sampling. **Rand.Sam.** maintains similarity checking but randomly selects one issue from each similar issue pair. For **Reve.Sam.**, we intensively select the earlier issues rather than the later ones.

Overall Ablation Results. Table 3 presents the results of the ablation study based on human-annotated evaluation. Our method achieves the highest performance in 15 out of 21 measures on `Test-I` and 12 out of 21 measures on `Test-II`. The results across both evaluations are generally consistent, with our method excelling at @K=1, @K=5, and @K=20, though it does not completely lead at @K=10.

Comparison of MI Variants. Comparing our method to alternatives in the first set (mutual information estimation), we observe the following: (i) Incorporating $H(\mathcal{Y})$ during training significantly impacts performance on Test-I. (ii) Removing CMI negatively affects performance at @K=1, @K=5, and @K=20. (iii) The -CMI +H($\mathcal{Y}$) variant generally outperforms +H($\mathcal{Y}$), likely because the model learns patterns along with mutual information. When +H($\mathcal{Y}$) is introduced without contextual information, it may confuse the model. (iv) The performance improvements from the alternatives to our method demonstrate that our semantic entropy-based approach is the most effective approximation for mutual information estimation.

Effectiveness of Soft-Thresholding. The results of -Soft-Threshold. are unstable and particularly worse than our method at @K=1, @K=5, and @K=20. The soft-threshold function is designed to selectively optimize the mutual information term for stable performance, and the results confirm its effectiveness.

Sampling Methods. For the alternative methods related to forward data sampling, all perform below our method at @K=1, @K=5, and @K=20. Specifically: i) Although Full Data includes a larger training set, it does not achieve superior performance, highlighting the necessity of similarity checking and the sampling process. ii) -Rein.Sam. performs similarly to Full Data, indicating that overlapping and similar information does not contribute meaningfully to training. iii) Rand.Sam. shows only slight improvements on Test-I, suggesting that similarity checking alone does not yield consistently strong results. iv) Reve.Sam. is the best-performing alternative in the data sampling group, but its performance remains inconsistent. In our

Table 3: Ablation Results on `Test-I` (Human-Annotated)

Methods	**MAP**	**@K=1**					**@K=5**					**@K=10**					**@K=20**				
		P	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**
Ours -	**72.08**	**87.50**	**33.45**	**44.20**	**87.50**	**87.50**	**52.50**	**74.61**	**57.85**	**74.67**	**100.00**	33.75	92.56	47.15	**81.43**	**100.00**	**19.38**	**100.00**	**31.40**	**85.09**	**100.00**
-CMI	63.47	50.00	14.70	21.88	50.00	50.00	50.00	64.73	54.40	62.04	87.50	33.75	92.56	47.15	72.92	**100.00**	19.38	**100.00**	31.40	76.42	**100.00**
$+H(\mathcal{Y})$	59.72	37.50	12.92	18.75	37.50	37.50	45.00	59.40	49.31	56.23	87.50	31.25	88.99	44.21	68.54	**100.00**	19.38	**100.00**	31.40	73.92	**100.00**
-CMI $+H(\mathcal{Y})$	63.47	50.00	14.70	21.88	50.00	50.00	50.00	64.73	54.40	62.04	87.50	33.75	92.56	47.15	72.92	**100.00**	19.38	**100.00**	31.40	76.42	**100.00**
-SoftThres.	62.30	50.00	21.25	26.49	50.00	50.00	40.00	62.50	45.96	56.48	87.50	32.50	90.77	45.68	70.61	**100.00**	19.38	**100.00**	31.40	75.19	**100.00**
Full Data	53.85	37.50	12.92	18.75	37.50	37.50	35.00	46.79	38.88	44.51	87.50	33.75	92.26	47.06	66.07	**100.00**	19.38	**100.00**	31.40	69.86	**100.00**
-Rein.Sam.	54.75	37.50	12.92	18.75	37.50	37.50	37.50	48.57	40.96	47.18	87.50	33.75	92.56	47.15	67.16	**100.00**	19.38	**100.00**	31.40	70.87	**100.00**
Rand.Sam	56.14	37.50	12.92	18.75	37.50	37.50	40.00	50.36	43.05	49.52	87.50	33.75	92.56	47.15	67.88	**100.00**	19.38	**100.00**	31.40	71.61	**100.00**
Reve.Sam	54.54	37.50	12.92	18.75	37.50	37.50	35.00	46.49	38.69	44.80	87.50	**35.00**	**94.35**	**48.62**	67.43	**100.00**	19.38	**100.00**	31.40	70.31	**100.00**

Table 4: Ablation Results on `Test-II` (LLM-Annotated)

Methods	**MAP**	**@K=1**					**@K=5**					**@K=10**					**@K=20**				
		P	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**	**P**	**R**	**F1**	**nDCG**	**HR**
Ours -	**65.12**	50.00	**21.88**	**27.32**	50.00	50.00	45.00	**64.58**	**49.65**	**63.75**	**100.00**	31.25	88.02	44.05	**73.22**	**100.00**	20.00	**100.00**	32.14	**79.16**	**100.00**
-CMI	48.08	37.50	8.04	13.12	37.50	37.50	35.00	45.24	37.73	42.92	75.00	27.50	87.05	38.90	58.47	**100.00**	18.75	**100.00**	29.87	64.97	**100.00**
$+H(\mathcal{Y})$	50.14	37.50	8.33	13.57	37.50	37.50	40.00	63.54	46.50	48.85	87.50	28.75	82.29	40.77	58.34	**100.00**	19.37	**100.00**	31.23	66.21	**100.00**
-CMI $+H(\mathcal{Y})$	51.90	50.00	12.20	19.37	50.00	50.00	37.50	45.24	39.96	46.11	75.00	28.75	87.05	40.50	61.65	**100.00**	19.37	**100.00**	30.86	68.16	**100.00**
-SoftThres.	55.39	25.00	6.67	10.42	25.00	25.00	45.00	57.01	49.18	50.91	87.50	32.50	84.31	44.67	62.85	**100.00**	20.63	**100.00**	32.82	69.85	**100.00**
Full Input	51.93	37.50	7.19	11.94	37.50	37.50	40.00	42.47	39.83	43.23	75.00	**37.50**	**92.71**	**49.68**	63.43	**100.00**	21.88	**100.00**	34.05	67.25	**100.00**
-Rein.Sam.	50.22	37.50	7.01	11.67	37.50	37.50	37.50	46.21	40.01	42.27	87.50	31.25	81.45	42.52	57.51	**100.00**	21.25	**100.00**	33.39	66.13	**100.00**
Rand.Sam.	49.81	37.50	7.41	12.29	37.50	37.50	37.50	41.92	38.66	40.18	75.00	33.75	90.18	46.36	60.52	**100.00**	20.62	**100.00**	32.78	65.61	**100.00**
Reve.Sam.	57.21	**62.50**	11.56	19.22	**62.50**	**62.50**	**47.50**	51.77	47.26	54.89	87.50	**37.50**	90.31	49.27	67.41	**100.00**	**23.13**	**100.00**	**35.53**	72.76	**100.00**

settings, selecting earlier generated issues tends to favor ground-truth issues. While ground-truth issues provide sufficient learning information, their format may not be well-suited for machine understanding, aligning with our discussion in Sec. 3.3.

5 Related Work

Reward Model. Recent advances in reward modeling have significantly improved preference learning, with LLMs producing quality preference labels more efficiently than human annotation (Zhou et al., 2024; Dubois et al., 2023). Multiple models now evaluate distinct attributes such as coherency and actuality (Gao et al., 2024; Wu et al., 2023). Notably, the CLoud reward model employs natural language critiques to enhance accuracy (Ankner et al., 2024), and adversarial regularization addresses out-of-distribution issues (Yang et al., 2024).

Mutual Information Estimation. Mutual information has become an effective tool for LLM regularization and performance assessment in tasks like question answering and causal graph discovery (Gendron et al., 2024; Wang et al., 2024b; Darvariu et al., 2024). Recent research also focuses on developing benchmarking methods (Xu et al., 2024) and exploring task calibration (Li et al., 2024). Additionally, mutual information has provided insights into the relationship between LLM feature spaces and LoRA distributions (Zhang et al., 2025).

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 (avelka et al., 2023), miss crucial legal rule details (Yuan et al., 2024a), 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

This paper introduces LIC, a dataset of 769 real-world court cases from Contract Act Malaysia, and presents a novel approach to legal issue identification using mutual information-based reward modeling with the sparsity-motivated training process. Our methodology, combining incremental fact incorporation and soft-threshold function application, significantly outperforms existing baselines, particularly on human-annotated evaluations. The ablation studies validate the effectiveness of our three key components: mutual information estimation, soft-threshold function, and forward data sampling. This work not only advances automated legal issue identification but also provides a substantial dataset for future legal AI research. Our contributions represent a meaningful step toward improving access to legal services, with practical implications for addressing global civil justice needs.

7 Limitation

Our approach to legal issue identification using LLMs has several limitations. First, while our dataset LIC represents a significant step forward, it is limited to cases from Contract Act Malaysia. This geographical and domain-specific focus may affect the generalizability of our findings to other jurisdictions or areas of law. Additionally, although validated by legal professionals, the silver ground-truth extracted by GPT-4o may still contain inherent biases or inconsistencies. A second limitation concerns the computational complexity of our method. The incremental fact incorporation process generates multiple issue candidates for each fact, potentially leading to significant computational overhead as the number of facts increases. While our forward sampling strategy helps mitigate this challenge, the trade-off between comprehensive coverage and computational efficiency remains a consideration. Finally, our mutual information-based approach, while effective, relies on approximations of semantic entropy that may not fully capture the nuanced relationships between legal facts and issues. The performance of our method could be affected by the quality of these approximations and the underlying language model's ability to understand complex legal contexts.

8 Ethics statements

We acknowledge and adhere to the ACL Code of Ethics throughout our research. Our work on automated legal issue identification raises several important ethical considerations that we have carefully addressed. The development of our dataset involved collaboration with law students and junior lawyers for validation. We ensured fair compensation for their expertise and maintained transparency about the intended use of their contributions. All case data used in our research is from publicly available court records, and we have taken care to handle this information responsibly. We recognize that automated legal analysis tools could impact access to justice and legal decision-making. While our work aims to improve access to legal services, we emphasize that our system is designed to assist, not replace, legal professionals. Users should be aware that the system's outputs are suggestions rather than definitive legal advice, and critical decisions should involve qualified legal practitioners. Furthermore, we acknowledge potential biases in both our dataset and model outputs. These could stem from historical biases in legal systems, regional variations in law interpretation, or limitations in language model training. We encourage users of our framework to consider these factors when applying our methodology in real-world contexts.

References

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D. Chang, and Prithviraj Ammanabrolu. 2024. Critique-out-loud reward models. *ArXiv*, abs/2408.11791.

Robert B Ash. 2012. *Information theory*. Courier Corporation.

Jaromír avelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4). *ArXiv*, abs/2306.09525.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Kelley Burton. 2017. " think like a lawyer" using a legal reasoning grid and criterion-referenced assessment rubric on irac (issue, rule, application, conclusion). *Journal of Learning Design*, 10(2):57–68.

Yinqiong Cai, J. Guo, Yixing Fan, Qingyao Ai, Ruqing Zhang, and Xueqi Cheng. 2022. Hard negatives or false negatives: Correcting pooling bias in training neural ranking models. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Sarah Chamness Long Jorge A. Morales Ted Piccone Alejandro Ponce Natalia Rodríguez Cajamarca Camilo Gutiérrez Patiño, Matthew Harman, Kirssy González Adriana Stephan, Sarah Chamness Long Jennifer VanRiper, Camilo Gutiérrez Patiño, Rachel Martin Jorge A. Morales Alejandro Ponce, Alicia Evangelides, Erin Campbell Alicia Evangelides Emily Gray Amy Gryskiewicz Camilo Gutiérrez Patiño Ayyub Ibrahim Priya Khosla Sarah Chamness Long Rachel Martin Jorge A. Morales Alejandro Ponce Natalia Rodríguez Cajamarca Leslie Solís Adriana Stephan, Lindsey Bock, Gabriel Hearn-Desautels Francesca Tinucci Adriana Stephan, Kirssy González, and Jennifer VanRiper. 2019. Global Insights on Access to Justice 2019. Technical report, World Justice Project.

Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. 2024. Large language models are effective priors for causal graph discovery. *ArXiv*, abs/2405.13551.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *ArXiv*, abs/2305.14387.

Felix Effenberger. 2013. A primer on information theory, with applications to neuroscience. *ArXiv*, abs/1304.2333.

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.

Tao Feng, Lizhen Qu, Zhuang Li, Haolan Zhan, Yuncheng Hua, and Gholamreza Haffari. 2024. Imo: Greedy layer-wise sparse representation learning for out-of-distribution text classification with pre-trained models. In *ACL*.

Bofei Gao, Feifan Song, Yibo Miao, Zefan Cai, Zhe Yang, Liang Chen, Helan Hu, Runxin Xu, Qingxiu Dong, Ce Zheng, Shanghaoran Quan, Wen Xiao, Ge Zhang, Daoguang Zan, Keming Lu, Bowen Yu, Dayiheng Liu, Zeyu Cui, Jian Yang, Lei Sha, Houfeng Wang, Zhifang Sui, Peiyi Wang, Tianyu Liu, and Baobao Chang. 2024. Towards a unified view of preference learning for large language models: A survey. *ArXiv*, abs/2409.02795.

Gaël Gendron, Bao Trung Nguyen, Alex Yuxuan Peng, Michael Witbrock, and Gillian Dobbie. 2024. Can large language models learn independent causal mechanisms? In *Conference on Empirical Methods in Natural Language Processing*.

Morgan A. Gray, Jaromír avelka, Wesley M. Oliver, and Kevin D. Ashley. 2024. Using llms to discover legal factors. In *International Conference on Legal Knowledge and Information Systems*.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Namita Jain and C. A. Murthy. 2014. A new estimate of mutual information based measure of dependence between two variables: properties and fast implementation. *International Journal of Machine Learning and Cybernetics*, 7:857 – 875.

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.

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.

Abraham Klaasen. 2017. Constitutional interpretation in the so-called 'hard cases': Revisiting s v makwanyane.

Camillia Kong, John Coggon, Michael Dunn, and Penny Cooper. 2019. Judging values and participation in mental capacity law. *Laws*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Bolian Li, Yifan Wang, Ananth Y. Grama, and Ruqi Zhang. 2024. Cascade reward sampling for efficient decoding-time alignment. *ArXiv*, abs/2406.16306.

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*.

Bing Liu, Pengyu Xu, Sijin Lu, Shijing Wang, Hongjian Sun, and Liping Jing. 2023. Sequential tag recommendation. *ArXiv*, abs/2310.05423.

Matej Martinc, Bla$zkrlj, andSenjaPollak$.2020. $Tnt-kid : Transformer-basedneuraltaggerforkeywordidentification$. *Natural Language* 409 − −448.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. 2020. Ccmi : Classifier based conditional mutual information estimation. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 1083–1093. PMLR.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Antonino Rotolo and Giovanni Sartor. 2023. Argumentation and explanation in the law. *Frontiers in Artificial Intelligence*, 6.

Alessandro Sordoni, Nouha Dziri, Hannes Schulz, Geoff Gordon, Philip Bachman, and Remi Tachet Des Combes. 2021. Decomposed mutual information estimation for contrastive representation learning. In *International Conference on Machine Learning*, pages 9859–9869. PMLR.

Amelia Carolina Sparavigna. 2015. Mutual information and nonadditive entropies: The case of tsallis entropy. *International journal of sciences*, 4:1–4.

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

Katarzyna Strębska. 2013. Usality relations in legal judgments on the example of the european court of human rigths. *Computer Languages, Systems & Structures*, 15:69–82.

Justin Sybrandt, Ilya Tyagin, M. Shtutman, and Ilya Safro. 2020. Agatha: Automatic graph mining and transformer based hypothesis generation approach. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. *arXiv preprint arXiv:2305.17256*.

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

Ziyu Wang, Hao Li, Di Huang, and Amir M. Rahmani. 2024b. Healthq: Unveiling questioning capabilities of llm chains in healthcare conversations. *ArXiv*, abs/2409.19487.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *ArXiv*, abs/2306.01693.

Shengwei Xu, Yuxuan Lu, Grant Schoenebeck, and Yuqing Kong. 2024. Benchmarking llms' judgments with no gold standard. *ArXiv*, abs/2411.07127.

Zhe Xu and Ray C. C. Cheung. 2019. Accurate and compact convolutional neural networks with trained binarization. In *British Machine Vision Conference*.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024. Regularizing hidden states enables learning generalizable reward model for llms. *ArXiv*, abs/2406.10216.

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. 2024a. 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*.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Jun Zhao, and Kang Liu. 2024b. Improving zero-shot LLM re-ranker with risk minimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17967–17983, Miami, Florida, USA. Association for Computational Linguistics.

Jing Zhang, Hui Gao, Peng Zhang, Shuzhen Sun, Chang Yang, and Yuexian Hou. 2025. The scaling law for lora base on mutual information upper bound.

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*.

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark B. Gerstein, Rui Wang, Gongshen Liu, and Hai Zhao. 2023. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *ArXiv*, abs/2311.11797.

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.

Hang Zhou, Chenglong Wang, Yimin Hu, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2024. Prior constraints-based reward model training for aligning large language models. *ArXiv*, abs/2404.00978.

A Appendix

A.1 RQs

> **Answer for RQ1**
>
> *The main results indicate that mutual information estimation demonstrates surprisingly robust effectiveness of issue identification in direct inference compared with the other untrained baselines.*

> **Answer for RQ2**
>
> *Sparsity-Motivated training for identifying legal issues can significantly improve the model's ability to accurately recognize legal issues and achieve high alignment with humans.*

> **Answer for RQ3**
>
> Our experiments demonstrate that LLMs can effectively assess the quality of generated legal issues when enhanced by mutual information-based approaches. This is evidenced by our method's strong alignment with human expert judgment and the success of our untrained model in evaluating fact-issue relationships. However, the assessment capabilities are most effective when structured through specialized frameworks like our sparsity-motivated training, rather than traditional reward or ranking approaches.

A.2 Hyperparameter Settings

With respect to the design of our identification process, our method allows the engagement of various generative models. Here, we adopted Llama 3.2 3B [7] with LoRA (Hu et al., 2021) as the backbone for our method and all applicable baselines. During the fine-tuning, the learning rate is 1e-5. We use a linear learning rate scheduler that dynamically decreases the learning- ing rate after a warm-up period. All experiments are conducted on NVIDIA A100 GPU.

[7] https://huggingface.co/meta-llama/Llama-3.2-3B

A.3 Formula Detail of Sparsity of Mutual Information

$$
\begin{aligned}
& I(\mathbf{x}_i; \mathcal{Y}_j | \mathbf{x}_{i-1}, ..., \mathbf{x}_1) \qquad (7)\\
&= p(\mathcal{Y}_j, \mathbf{x}_i, ..., \mathbf{x}_1) \log \frac{p(\mathcal{Y}_j, \mathbf{x}_i | \mathbf{x}_{i-1} ..., \mathbf{x}_1)}{p(\mathcal{Y}_j | \mathbf{x}_{i-1} ..., \mathbf{x}_1) p(\mathbf{x}_i | \mathbf{x}_{i-1} ..., \mathbf{x}_1)}\\
&= p(\mathcal{Y}_j, \mathbf{x}_i, ..., \mathbf{x}_1) \log \frac{p(\mathcal{Y}_j | \mathbf{x}_{i-1} ..., \mathbf{x}_1) p(\mathbf{x}_i | \mathbf{x}_{i-1} ..., \mathbf{x}_1)}{p(\mathcal{Y}_j | \mathbf{x}_{i-1} ..., \mathbf{x}_1) p(\mathbf{x}_i | \mathbf{x}_{i-1} ..., \mathbf{x}_1)}\\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
I(\mathbf{X}; \mathcal{Y}_j) &= \sum_{i=2}^{m} I(\mathbf{x}_i; \mathcal{Y}_j | \mathbf{x}_{i-1}, ..., \mathbf{x}_1) + I(\mathbf{x}_1; \mathcal{Y}_j) \qquad (8)\\
&= \sum_{i=2}^{t} I(\mathbf{x}_i; \mathcal{Y}_j | \mathbf{x}_{i-1}, ..., \mathbf{x}_1) + I(\mathbf{x}_1; \mathcal{Y}_j)
\end{aligned}
$$

A.4 Fact 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 courts
    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.5 Human Annotation Designing

We create a Google Form for annotators in this task. For each case, the ground scenarios (Facts), issue A (Ground Truth Issue), and issue B (Generated Issue) are given.

A.5.1 General Annotation Description

There are 20 cases in the form. Each case has two questions. Please carefully read the cases and issues to give your response. You are provided with two legal issues and their associated factual scenarios. (Issue A is in the description, and Issue B is respectively depicted in each row.) Your task is to determine if the two issues are similar or paraphrased to each other based on given facts.

A.5.2 Case Example

Scenario: The applicant, Tay Yong Kwang, served three months and two days as pupillage with a practising advocate and solicitor, and not less than 18 months with a legal officer, which counts as not less than three months with an advocate and solicitor under the Legal Profession Act. Tay attended a Postgraduate Practical Course in law conducted by the Board of Legal Education outside normal office hours from 13 April 1981 to 30 June 1981. Tay petitioned the Court for admission as an advocate and solicitor of the Supreme Court and sought clarification on the construction of section 11(5) of the Legal Profession Act regarding whether time spent attending prescribed courses outside normal office hours should be counted towards his pupillage. **Issue A**: Whether the time spent attending a prescribed course outside normal office hours should count towards the pupillage period as per section 11(5) of the Legal Profession Act. **Issue B** (A separate row in Multiple-Choice Grid): Whether Tay Yong Kwang's combined experience of three months and two days as pupillage with a practising advocate and solicitor, and not less than 18 months with a legal officer, satisfies the requirements for admission as an advocate and solicitor under the Legal Profession Act

Given legal Facts, Issue A, and Issue B, please determine whether Issue B is Similar to or a Paraphrase of Issue A within the context of the given Facts.

The annotators are required to select from one of the options to answer the question:

- Yes, the issue in this row is similar to issue A (given in the description)

- No, the issue in this row is not similar to issue A (given in the description)

A.6 Data Example: Case ID - IFSG681

A.7 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 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.9 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.10 Prompt Designing

A.10.1 Incremental Issue Generation

Prompt 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", ... }

A.11 Pairwise Issue Comparison

Prompt for Pairwise Issue Comparison

Given a legal scenario and two potential legal issues, determine which issue is more relevant or significant.
Given Information: - Given Scenario: {facts} - Issue A: {issue_a} - Issue B: {issue_b}
Returns: - str: Either "Issue A is better" or "Issue B is better", based on legal analysis.
Instructions:

1. Analyze the scenario carefully, identifying key facts and legal principles.
2. Evaluate Issue A and Issue B based on their relevance, strength, and legal impact.
3. Compare both issues, considering legal precedent, logic, and significance.
4. Decide which issue is more relevant or important in resolving the scenario.
5. Return the decision as either "Issue A is better" or "Issue B is better".

Expected Output (Your response should select from one of the following answers): - "Issue A is better" (if contract breach is legally stronger) - OR "Issue B is better" (if tenant rights violation is more significant)
Your Response:

A.12 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.