Bridging Law and Data: Augmenting Reasoning via a Semi-Structured Dataset with IRAC methodology

Anonymous ACL submission

Abstract

Despite the advancements of Large Language Models (LLMs), their effectiveness in legal reasoning is limited due to unique legal terminologies and the need for highly specialized knowledge. These limitations can be addressed with high-quality data for complex legal reasoning. To this end, this paper introduces a benchmark, LegalSemi, annotated with IRAC (Issue, Rule, Application, Conclusion) for legal scenario analysis, developed by legal experts. It includes 54 legal scenarios annotated with full IRAC analysis and an associated structured knowledge graph (SKG). Our analysis reveals that Mistral-7b, a state-of-the-art LLM, is particularly adept at identifying legal concepts, while GPT-3.5 shows superior performance in analysis and conclusion tasks. Notably, standard LLMs face challenges in rule retrieval, an issue significantly mitigated by integrating SKG, which enhances the accuracy by 48%. LegalSemi serves as an innovative and valuable benchmark for complex legal reasoning, with the potential for broader applications across various legal domains.

1 Introduction

Access to justice is a broad social problem. Two-thirds of people in the United States experienced at least one legal issue in the past four years, with less than half of those problems having been completely resolved [1]. In India, more than 10,490 legal cases in Supreme Court of India have been pending for more than a decade (Madhana and Subhashree, 2022). IRAC framework (Metzler, 2002), standing for issue, rule, application, and conclusion, is the problem solving framework widely used by legal professionals to determine legal questions, facilitating legal reasoning to extract and transform facts in a legal scenario into a legal conclusion.

AI models, in particular large language models (LLMs), demonstrate great potentials to improve access to justice (Krasadakis et al., 2024). However, it is still challenging for LLMs to perform IRAC analysis on legal scenarios. The recent study (Kang et al., 2023) shows that ChatGPT fails to solve legal problems with IRAC completely correct on any of the evaluated legal scenarios. In a large proportion of the scenarios, ChatGPT managed to draw correct conclusions but produced wrong intermediate reasoning steps. In majority of the scenarios, ChatGPT was not able to cite correct legal rules during legal analysis. In real-world, it is crucial for legal professionals to understand every single reasoning step that leads to the final conclusion. We conjecture that i) LLMs, e.g. ChatGPT, do not fully understand the underlying legal knowledge; ii) errors in IRAC analysis may attribute to the well-known hallucination problem of LLMs (Rawte et al., 2023).

Recent advances show that it is possible to mitigate the hallucination problem of LLMs by leveraging structured knowledge graphs (SKGs) (Pan et al., 2024). SKGs can enhance LLMs in terms of interpretability and faithfulness by providing external knowledge (Kim et al., 2024). If legal knowledge is stored in SKGs, it is also easy to keep it up-to-date, in accordance with the revisions of legislation. Unfortunately, existing IRAC datasets do not contain any SKGs for legal knowledge.

To address the problems above, we carefully curate **LegalSemi**, a dataset comprising legal scenarios relevant to the "Formation of Contract" in Malaysian Contract Law, accompanied by rich structured IRAC analysis carried out by top law students. We extract structured semantic information from a law textbook and a legislation in a semi-automatic manner to build an SKG. In the SKG, a node represents either a legal concept, a court case, a legal rule, the interpretation of a legal rule or a concept in lay language, or relevant meta

[1] https://iaals.du.edu/publications/justice-needs-and-satisfaction-united-states-america

information, while an edge between two nodes denotes their relation. The rigorous layout in the textbook and the legislation facilitates rule-based extraction of semantic relations between legal concepts as well as their relations to legal rules and interpretations. We demonstrate the usefulness of the SKG for LLMs through extensive experiments and obtain the following key findings:

- Following (Kang et al., 2023), we apply an LLM to decompose a legal question into a set of simpler questions, followed by rule retrieval and performing legal analysis to answer each decomposed question. Incorporating legal concepts from the SKG, we improve the quality of decomposed question generation by 6%. Those improved decomposed questions lead to a significant 21% of improvement in application and a 13% enhancement in drawing conclusions.

- By enhancing an LLM with the structured legal knowledge in the SKG, we achieve a 60% increase in recall and a 12% improvement in the F1 score at top-5 results of rule retrieval. The improvement of rule retrieval further enhance the legal analysis in application by 48%. We found that legal concepts greatly help in bridging the semantic gaps between facts in scenarios and rules in the legislation. The interpretations in lay language further reduce language gaps between scenarios in lay language and statutes in legalese.

2 Dataset

LegalSemi is constructed based on Contract Law Malaysia. The dataset is valuable because the legal documents pertinent to this law are less likely to be memorized by the existing LLMs. Besides, contracts are important legal documents that are common in everyday life. As it is time-consuming for law students to annotate legal scenarios, we focus on *formation of contracts*, which is one of the most important subareas of the Contract Law, and it includes rich and representative scenarios for IRAC analysis.

To build the dataset, we start by collecting legal scenarios pertaining to *formation of contracts*, followed by annotating the scenarios with legal concepts that could eventually support complete IRAC analysis. In order to address the limitations of LLMs, including i) wrong references to statutes and precedents, ii) gaps between everyday language and legalese, and iii) weak legal reasoning capability, we construct a structured knowledge graph (SKG) to support neuro-symbolic approaches.

2.1 Scenarios Collection

To ensure diversity of scenarios and coverage of legal concepts pertinent to *formation of contracts*, we gather scenarios based on the law textbook "Law for Business" (Trakic et al., 2022) used by law students when studying contract law.

In particular, we choose five main topics: *offer and acceptance*, *consideration*, *certainty*, *capacity*, and *intention to create legal relations*. The corresponding chapters in the text book are Chapter 4 "Formation of Contract: Proposal and Acceptance", Chapter 5 "Consideration", Chapter 6 "Promissory Estoppel", and Chapter 7 "Intention to Create Legal Relationships and Capacity". The section headings of these chapters represent the corresponding subtopics, such as *proposal*, *acceptance*, and *minors* etc.. There are 55 unique subtopics in total.

Based on the main topics and subtopics, we recruit four second-year law students and two junior lawyers to gather scenarios in two ways. First, we collect 24 scenarios from tutorial questions, books, and past exam questions. Second, for the remaining subtopics, we utilize ChatGPT to suggest candidate scenarios with the prompt : *" You are a legal professional, based on the example scenarios, main topic, and subtopics, create a new scenario around avg_length"*. The average length is calculated based on the human-authored scenarios. This parameter is used to guide ChatGPT to generate scenarios with a length that matches those curated by humans. As the result, the main topics are evenly distributed among all the scenarios, and each subtopic is covered by at least one scenario. To ensure the quality of the scenarios, we ask two of the six law students to evaluate the quality of the scenario candidates using the following questions, as shown in Fig. 1.

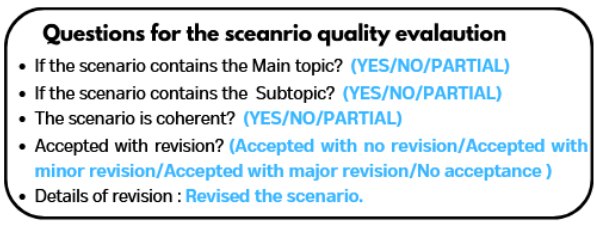

Figure 1: Scenario evaluation questions.

The result shows that only 16.67% of scenarios are not agreed upon by the annotators. The disagreement regarding the subtopics stands at 7.41%. Moreover, 94.44% of the scenarios are coherent, which is a testament of the overall good quality of the scenarios. We also found that 66.7% of scenarios are accepted without any revision. Given the specific constraints of the scenarios, the quality of the scenarios created by both the annotators and ChatGPT is good, as evidenced by the results. For scenarios requiring revision, the best performing annotator is tasked to modify them before any further data annotation.

2.2 Data Annotators and Annotation Tool

From a pool of applicants, we carefully selected six data annotators. This diverse team comprises four second-year law students from three distinct Malaysian universities and two junior lawyers. Compensation was set at MYR30 per hour, reflecting the complexity and importance of their tasks. Each annotator typically requires approximately three hours to complete the annotation of a single scenario using the IRAC method. These annotators are required to have achieved at least a B grade in related law subjects. Furthermore, following a comprehensive briefing session, they were mandated to pass a specialized pre-test before being recruited.

To facilitate this intricate annotation process, we developed an online data annotation platform, grounded in the principles of IRAC methodology. It is designed for universal accessibility, requiring only an internet connection. It features a 'Review' function, allowing annotators to refine and adjust their inputs as necessary. Data output is organized into a structured .json and ./txt format, significantly enhancing efficiency and streamlining the data processing workflow for subsequent analysis [2].

2.3 Annotation of IRAC Analysis

Annotation details The following are the details of the annotation steps. Figure 2 shows the example of the annotation of each step.

Step 1: Legal concepts identification. Drawing from a predefined list of legal concepts, annotators are tasked with highlighting relevant legal concepts within the presented scenarios. They are primarily guided to reference the index of a designated legal textbook, for example: 'advertisements:invitation to treat','acceptance:proposal conditions','offeree'. This approach ensures that the identification of legal concepts is grounded in authoritative legal sources, providing a robust foundation for further analysis. However, given the dynamic nature of legal terminology, the procedure does support flexibility where common legal concepts, such as 'offeror', if not listed in the index, the annotators have the discretion to incorporate these terms.

Step 2: Issue and decompose questions. The annotators need to input the main issues for the given scenarios. The main questions should be based on 'Was there a valid contract between A and B?', while the decomposed questions should be the sub-issues based on the scenarios. Figure 2 shows example issues and decomposed questions.

Step 3: Rules. The annotators need to select the relevant sections from a drop-down list containing all the sections from the Contract Act 1950. For example, 'Section 2a'. In total, we have 280 sections listed in the database. Additionally, in the text box, they must input related court cases with page numbers. For instance, *Eckhardt Marine GMBH v Sheriff, High Court of Malaya, Seremban & Ors [2001] 4 MLJ 4 (CA) [3/4]*. These input contents will be displayed as buttons, which they can reuse in the analysis.

Step 4: Analysis. In the analysis, annotators are required to analyze the given scenario in point form. They are encouraged to use IF...THEN.... conditional statements for the analysis. One example of the analysis: *"1. IF Vanessa's advertisement is an invitation to treat, then {she receives a call from a customer, Niko, to reserve that vinyl. {[Niko's reply to the invitation to treat]} is an offer {Section 2a}{Preston Corp Sdn Bhd v Edward Leong [1982] 2 MLJ 22 (FC)[2/4]}. 2. IF {Fine, I will reserve the vinyl for you until Wednesday 8pm. If I don't hear from you by then, I will sell the vinyl for someone else {[Vanessa's reply to the offer]} is an absolute and unqualified acceptance, then there is a valid acceptance {Section 7a}."*. The { } indicates the legal concepts which the annotators highlight in the previous step. They need to reuse the legal concepts, sections and court cases, wherever applicable.

Step 5: Conclusion. The last step of IRAC is conclusion, which answers the main questions. According to common legal practice, it is intended to present the full sentence of the conclusion. Therefore, we provide a text box for inputting the conclusion. For instance, *"There is no valid contract between Emma and Danny."*

[2]Website: https://legal-annotator.vercel.app/

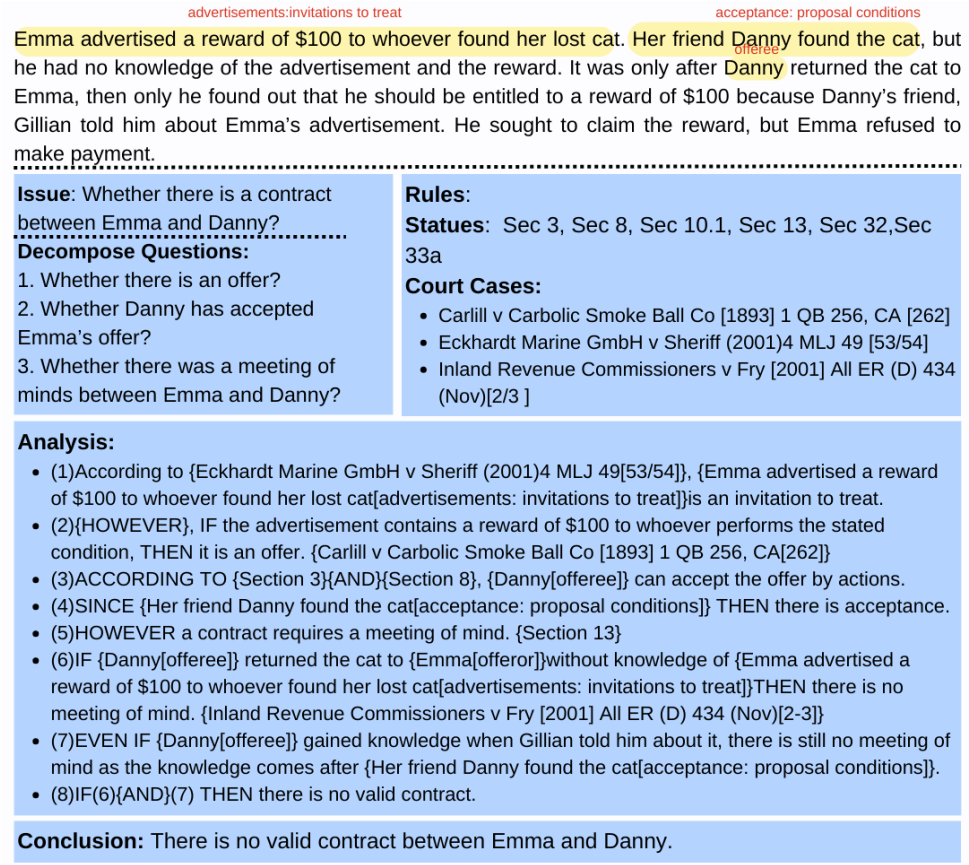

Figure 2: Example of annotation.

2.3.1 Data Quality Insurance

Drawing upon the principles of the IRAC methodology and its associated standards, we develop comprehensive annotation guidelines to streamline the evaluation process. The evaluation covers six distinct aspects: overall performance, issue identification, rule clarification, legal concept identification, analytical generation, and the conclusions. For each area, we establish three evaluation criteria: -1 for disagreement, 1 for agreement, and 0 for neutrality. Specific instructions are provided for each score in every area, as detailed in our guidelines.[3] For each scenario, another annotator performs verification by assigning scores based on these evaluation criteria. The inter-rater agreement rate surpasses 0.8 demonstrating a high level of consistency among evaluators. In instances of disagreement, we consult an expert to make the final decision and to implement necessary adjustments.

[3]Evaluation Guidelines: https://anonymous.4open.science/r/CLIRAC-B3FC/Evaluton%20Guidelines.pdf

2.4 Structured Knowledge Graphs

Structured Knowledge Graphs (SKGs) significantly enhance Large Language Models (LLMs) by providing organized, interconnected data representations. This methodical arrangement allows LLMs to make coherent and clear interpretations, aligning seamlessly with their ability to recognize data patterns and relationships. This is particularly beneficial in domains that demand precision, such as scientific research, financial analysis, and medical diagnostics (Sajid, 2023).

Legal text often resembles structured knowledge. For example, under the Contract Act 1970, *Section 2(a)* states: *"when one person signifies to another his willingness to do or to abstain from doing anything, with a view to obtaining the assent of that other to the act or abstinence, he is said to make a proposal;"* .This section is related to the legal concept *"offer"* and corresponds to paragraph P4-014 in the text book.

Given the nature of legal knowledge and the benefits of SKGs for LLMs, we design an SKG based on the legal knowledge from book paragraphs, legal concepts, laws, and court cases.

The Entity-Relationship Diagram (ERD) of the knowledge graph is shown in Fig 3.

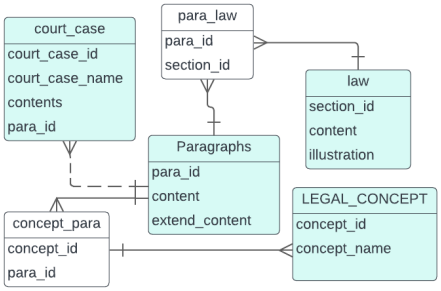

Figure 3: ERD diagram of the legal knowledge structure.

By using the SKG, we can access related information based on any given data point. For example, legal concept is "absolute," the output find the related paragraphs with associated laws or court cases. From the output we know that the related law is *Section 2(a)* and *Section 7(a)*"

From the SKG, we understand that even with just a part of the information, we can trace all related data. The paragraphs provide explanations for the associated legal concepts. Normally, legal concepts and laws are challenging to process or understand. However, the paragraphs from the book are closer to common English. One of the main advantage to use SKG it is help to lower the language gap between legal language and common English. In addition, compared to using other knowledge graphs, our SKG are specifically related to the current scenario and offer supporting information for the relevant paragraphs, court cases, or laws. In the follwoing section, we will dicsus more about the applciation of SKG.

2.5 Data Statistics

Data supporting Legal AI, particularly in fostering reasoning capabilities, is indeed rare and the task of annotating for reasoning is challenging. In our comparative analysis presented in Table 1, we evaluate our dataset **LegalSemi** against other notable works in this domain.

Among these, SIRAC (Kang et al., 2023) emerges as the most comparable dataset to ours. However, **LegalSemi** surpasses SIRAC in several key aspects: greater number of scenarios, longer average scenario lengths, legal concepts and linked with an external knowledge graph. These enhancements not only add complexity but also depth to the dataset.

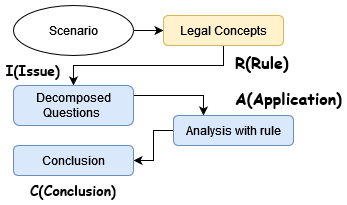

Figure 4: Flow of the legal reasoning

SARA (Holzenberger and Van Durme, 2021), another dataset in the comparison, does not employ the IRAC analysis, which is a critical framework for structured legal reasoning. In addition, they have very limited rules (7 rules) applied. Legal Bench (Guha et al., 2022), while valuable, are constrained by their use of fixed decomposed questions for the reasoning process. This approach may restrict the flexibility and adaptability of the reasoning process to diverse legal scenarios.

3 Experiments and Results

In this section, we provide the baseline experiments for the legal reasoning tasks. Figure 4 shows the flow of legal reasoning using the IRAC methodology. To perform comprehensive reasoning, we separate the tasks based on legal concepts, decomposed questions, analysis with rules, and conclusions. We present a comprehensive analysis of the experiments conducted and the results obtained from LegalSemi. Our primary objective is to enhance the legal reasoning framework based on the insights derived from our collected data.

3.1 Incorporation of Legal Concepts for Generation of Decomposed Questions

A legal concept refers to the entity highlighted by legal professionals. These concepts typically linked to the key issues in given legal scenarios. However, compared to common entities, legal concepts are more challenging to extract from the scenario text. For example, in the legal context, an advertisement to sell a book is considered an "invitation to treat," whereas in a common sense context, an advertisement is more directly associated with the product and its purchase. As experiments for all different LLMs, we found that Mistral-7b currently stands out for its efficiency in legal concept identification with accuracy of 92%.

Kang et al. (2023) demonstrate that decomposed questions aid the overall reasoning process. However, generating precise decomposed questions re-

	Num Scenario	Avg Scenario Len	IRAC	Legal Concept	Avg_ DecomQ	Paris of legalQA	Rules	Analysis Avg Len	External Knowledge
SIRAC (Kang et al., 2023)	40	585	yes	0	3.35	21	58	7.05	No
Legal_Semi	54	1048	yes	297	3.85	262	90	13.4	Yes
sara_entailment (Holzenberger and Van Durme, 2021)	277	99.25	no	60	NA	375	9	No	No
sara_numeric (Holzenberger and Van Durme, 2021)	100	255.5	no	60	NA	100	9	No	No
LEGAL BENCH legal_reasoning_causality (Guha et al., 2022)	59	1153.5	yes	0	3	59	No	No	No
legal BENCH contract_qa (Guha et al., 2022)	88	264.25	yes	0	7	88	No	No	No

Table 1: The statistics of the relevant datasets.

Model / Methods	Direct	Direct& RAP	Direct & Legal Concept	Direct & Legal Concepts& RAP	WIN/TIE/LOSS
GPT3.5	0.79	0.75	0.86	0.85	2078/267/215
Llama2-70b-chat	0.31	0.32	0.36	0.49	602/105/103
Mistral-7b-instruct	0.28	0.23	0.23	0.25	764/1200/852
Gemini	0.69	0.69	0.59	0.67	1429/434/337

Table 2: Evaluation result for the decompose questions.

mains a significant challenge. LegalBench (Guha et al., 2022) employs a fixed list of decomposed questions, which is not optimal for the legal field due to the unique nature of each scenario. Therefore, we set up experiments use legal concepts as a guide, helping the LLMs generate the correct decomposed questions.

Result and discussion. Table 2 showcases the automated outcomes for all considered models, revealing GPT-3.5's consistent strong performance with an accuracy of 85%, marking the highest accuracy achieved. Notably, Gemini exhibits performance levels closely resembling those of GPT-3.5, distinguishing itself among the models.

With the usage of legal concepts, the result increases by 7% for GPT-3.5. The strength of GPT-3.5 lies in its proficient generation of effectively decomposed questions, contributing to its superior performance. Integrating legal concepts enhances accuracy across models, except for Mistral-7b. This suggests GPT-3.5's existing capabilities are already optimized, making additional prompts less impactful. Despite enhancements in other models, GPT-3.5 maintains its lead, underscoring its robustness in this task. Gemini, despite slight differences from other models, performs commendably and offers cost advantages being freely accessible.

3.2 Rule Retrieval

Rule retrieval presents a significant challenge for Large Language Models (LLMs). From our preliminary experiments, we observe that LLMs face challenges in accurately identifying the **Rule** component in legal cases. To address this issue, we have explored different methods aiming at enhancing the rule retrieval with or without the SKG.

3.2.1 Experimental Setup

In our experiments, we vary between different types of queries, different types of documents for building the search indexes, with or without reranking, detailed below.

Queries. We consider two different types of queries: scenario texts or legal concepts. Herein, a scenario is mapped to an TF-IDF query vector for a traditional IR engine. Legal concepts are sent as SQL queries. Legal concepts are estimated by using LLMs. Among the LLMs being evaluated, Mistral-7b-instruct (Jiang et al., 2023) is chosen because it achieves an accuracy of 92%.

Indexing. To test the effectiveness of our indexing strategies, we deploy the queries from previous steps. We use TF-IDF as indexing system that builds indexes for the fast search of documents given their feature vectors. We compare building the index with either legal rules or interpretations in lay language. As the interpretations extracted from the textbook have a low coverage, we also apply ChatGPT to generate interpretations for the uncovered legal rules.

Similarity Measures. As textual queries are embedded into TF-IDF vectors, we apply Euclidean distance as the similarity measure to compare textual queries with indexed documents.

Reranking methods. When legal concepts are issued as queries, we rerank retrieved legal rules directly using TF-IDF vectors or using associated interpretations. Herein, we also consider using the interpretations generated by ChatGPT.

Evaluation metrics We consider precision, recall, and F1 scores at top-k retrieved results, where k =5, 10, and 50, respectively.

3.2.2 Results and Discussions

Language Gap between Scenarios and Law. We compare first different document types for indexing when using scenarios as queries, without reranking. As shown in the upper part of Table 3, precision at varying top-k are below 3% and the highest recall is 12.5% when using scenarios as textual queries and legal rules as the index. With interpretations as the index, the precision is improved but the recalls drop significantly.

One of the main reasons is the language gap between law and scenarios. Scenarios are expressed in plain English, whereas the law uses legalese. For example: *Sec 2a: when one person signifies to another his willingness to do or to abstain from doing anything, with a view to obtaining the assent of that other to the act or abstinence, he is said to make a proposal;*. "Signifies" here refers to a proposal that could be made in any form: orally, in writing, through conduct, or a combination of these methods. The legal definition which differs compared to common English. Although interpretations are conveyed using lay language, the coverage of associated legal rules is fairly low because the textbook includes a limited number of examples.

Legal Concepts for Mitigating Semantic Gaps. When issuing legal concepts as SQL queries, followed by reranking legal rules based on either associated interpretations or the rules directly, we observe a surge of recall and a significant improvement of precision. It suggests that legal concepts help mitigate semantic gaps between scenarios and legal rules. Interestingly, it's worth noting that when comparing interpretations of law sourced from textbooks versus those from GPT-3.5, the former tends to yield better results. The highest recall rate is achieved when indexing applies textbook interpretations, reaching 35.3% in the top 50 results.

3.3 Application

Legal reasoning poses one of the most significant challenges for current language models. While people often utilize knowledge graphs and multi-hop reasoning to address complex issues, these methods do not perform well in legal reasoning tasks due to the complex reasoning steps needed for legal scenarios. Professionals typically employ the IRAC methodology to conduct the reasoning process. They begin by identifying the issues and rules, followed by analysis. Kang et al. (2023) show that decomposed questions improve the quality of the analysis. LLMs are more accurate when we ask more specific and simpler questions. It remains to investigate whether LLMs benefit from legal rules and their interpretation for legal analysis.

Experiment setup We compare different inputs for LLMs to generate legal analysis for Application: i) a scenario and its main question, ii) decomposed questions based on detected legal concepts, and iii) decomposed questions and the ground truth rules. The prompts used in the experiments are outlined in the Appendix.

Results. Table 4 shows the rule application results with different input queries. We use the GPT3.5 to evaluate the results.

The analysis of results reveals a significant improvement across all modules when utilizing decomposed questions and rules derived from ground truth data. Particularly noteworthy is Mistral's substantial increase of 48% in results upon incorporating decomposed questions and rules. Meanwhile, Gemini maintains its position as the top performer, demonstrating improvement even when employing the same methodology.

4 Related Work

Legal Reasoning Savelka et al. (2023) analyzed how effectively GPT-4 produces definitions for legal terms found in legislation. Huang et al. (2023) addressed the challenge of improving Large Language Models (LLMs), such as LLaMA, for domain-specific tasks in the legal field. LegalBench (Guha et al., 2022) is created through an interdisciplinary procedure for legal scenario analysis using the IRAC methodology. However, their work did not utilize the same legal scenarios for the completed IRAC tasks. Large Language Models (LLMs) have demonstrated significant reasoning abilities, especially when chain-of-thought (CoT)

	query: scenario index: law			query: scenario index: interpret (text book)			query: scenario index: gpt_interpret		
	@ top5	@ top10	@ top50	@ top5	@ top10	@ top50	@ top5	@ top10	@ top50
Precision	2.60%	1.70%	1.40%	4.30%	4.90%	7.80%	3.30%	4.40%	3.20%
Recall	2.90%	3.30%	12.50%	0.90%	1.85%	15.70%	2.30%	9.00%	29.40%
F1 score	2.50%	2.00%	2.50%	1.50%	2.54%	9.50%	2.60%	5.50%	5.60%
	query: legal concept + scenario index: law			query:legal concept+ scenario index: interpret (text book)			query: legal concept + scenario index: gpt_interpret		
	@ top5	@ top10	@ top50	@ top5	@ top10	@ top50	@ top5	@ top10	@ top50
Precision	9.70%	7.50%	3.10%	11.80%	13.30%	11.80%	10.30%	9.00%	4.40%
Recall	32.20%	32.60%	37.20%	35.30%	31.20%	35.30%	33.20%	36.50%	48.50%
F1 score	13.90%	11.50%	5.60%	16.30%	17.20%	16.30%	14.60%	13.50%	7.90%

Table 3: Evaluation Results: Rule retrieval.

	Direct	RAP	DecomQ	DecomQ &RAP	DecomQ &Rule& RAP	WIN/TIE/LOSS
LLAMA	0.34	0.32	0.3	0.41	0.61	637/631/333
MISTRAL	0.23	0.32	0.49	0.36	0.71	1018/742/754
Gemini	0.71	0.67	0.65	0.73	0.76	1178/155/351

Table 4: Application Result

prompting is employed. CoT-style prompting (Wei et al., 2022; Hu et al., 2023) involves, given a complex question (Q), the LLM generating a reasoning chain (C) along with the final answer (A). Hao et al. (2023) proposed Reasoning via Planning (RAP). RAP enhances the LLM with a world model and employs principled planning, namely Monte Carlo Tree Search (MCTS), to generate high-reward reasoning traces following effective exploration, demonstrating its superiority over several contemporary CoT-based reasoning approaches. However, these approaches, including RAP, have yet to be applied in the legal domain, as Legal AI requires highly domain-specific legal knowledge rather than just common sense knowledge.

Structured knowledge graph SKILL (Moiseev et al., 2022) demonstrated that the results show improvements with pre-trained models on the Wikidata KG, beating the T5 baselines on FreebaseQA, WikiHop, and the Wikidata-answerable subset of TriviaQA and NaturalQuestions. Knowledge graphs with external knowledge can help the model improve accuracy and reduce confusion. Leveraging the power of structured knowledge graphs is able to enhance the performance of the LLMs. The current approach mainly focuses on common sense knowledge. Especially in legal reasoning, we need external knowledge to ensure that the model is capable of providing more accurate answer.

5 Conclusion

In this paper, we introduce **LegalSemi**, which consists of 54 scenarios annotated with IRAC analysis in the area of contract law and an SKG for legal knowledge extracted from a law textbook and legislation. The SKG covers legal concepts, legal rules, interpretations in lay language etc. and their relations. Legal concepts from the SKG are particularly useful for improving the quality of decomposing questions by 6%, legal analysis (Application) by 21%, and conclusions by 13%.

We observe that LLMs fall short of identifying relevant legal rules accurately by having the mean precision at top-5 below 3%. By leveraging the SKG, we achieve a remarkable improvement of the rule retrieval at 17.2% of the F1 score. Using legal concepts as queries greatly improve both precision and recall for rule retrieval.

Our analysis of various LLMs shows that self-check prompts has led to a 14% improvement in the accuracy of LLMs across four different tasks. While Mistral-7b excels in identifying legal concepts, it requires further refinement for accuracy. However, a notable limitation across LLMs is that they struggle with accurately identifying the correct rule for given scenarios. The introduction of the SKG has significantly enhanced rule retrieval. With the rules and decompose questions, the analysis result improved 48%.

Future work will focus on enhancing the content linkage within the SKG to cover a broader range of legal concepts. Additionally, we aim to implement more advanced re-ranking models to further improve rule retrieval for legal analysis. This study underscores the potential and areas for improvement in employing LLMs for IRAC analysis.

6 Limitation

In this study, our primary emphasis revolves around examining scenarios that pertain specifically to the 'Formation of Contract' as delineated within the Malaysia Contract Law. While our dataset may exhibit limitations in terms of the breadth of legal scenarios available for analysis, it remains robust in its coverage of all essential topics to contract formation. Despite potential constraints, such as data availability or accessibility, our dataset is meticulously curated to encompass a comprehensive spectrum of scenarios relevant to the legal domain, ensuring a thorough investigation into the intricacies of contract formation under Malaysian law.

Furthermore, an additional limitation inherent in our study lies in the selection of LLMs employed for our experiments. Our study opts for a more focused approach by utilizing a limited subset of these models. While this decision may result in a narrower scope of analysis compared to studies incorporating a broader array of LLMs, it ensures consistency and reliability in our experimental methodology. Despite this limitation, our choice of employing the most widely used and recognized LLM ensures that our findings are grounded in established practices within the field of natural language processing and legal analysis.

References

Neel Guha, Daniel E Ho, Julian Nyarko, and Christopher Ré. 2022. Legalbench: Prototyping a collaborative benchmark for legal reasoning. *arXiv preprint arXiv:2209.06120*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Nils Holzenberger and Benjamin Van Durme. 2021. Factoring statutory reasoning as language understanding challenges. *arXiv preprint arXiv:2105.07903*.

Tongxin Hu, Zhuang Li, Xin Jin, Lizhen Qu, and Xin Zhang. 2023. Tmid: A comprehensive real-world dataset for trademark infringement detection in e-commerce. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 176–184.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton, and Genevieve Grant. 2023. Can chatgpt perform reasoning using the irac method in analyzing legal scenarios like a lawyer? *arXiv preprint arXiv:2310.14880*.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*.

Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S Verykios. 2024. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics*, 13(3):648.

B Madhana and S Subhashree. 2022. A study on backlog of cases. *Issue 5 Int'l JL Mgmt. & Human.*, 5:942.

Jeffrey Metzler. 2002. The importance of irac and legal writing. *U. Det. Mercy L. Rev.*, 80:501.

Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. Skill: structured knowledge infusion for large language models. *arXiv preprint arXiv:2205.08184*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Haziqa Sajid. 2023. Leveraging the power of knowledge graphs: Enhancing large language models with structured knowledge. Blog.

Jaromir Savelka, Arav Agarwal, Christopher Bogart, and Majd Sakr. 2023. Large language models (gpt) struggle to answer multiple-choice questions about code. *arXiv preprint arXiv:2303.08033*.

Adnan Trakic, Nagiah Ramasamy, Cheah You Sum, Paul Linus Andrews, Sri Bala Murugan, P Vijayganesh, and Kanchana Chandran. 2022. *Law for Business, Third Edition*. Sweet & Maxwell Malaysia.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

A Appendix

A.1 Annotation Guidelines

Project Overview Develop a machine learning system for in-depth analysis of legal scenarios, specifically focusing on Contract Law utilising the IRAC (Issue, Rule, Analysis, and Conclusion) methodology.

Methodology: Apply Contract Law principles to annotate data using the IRAC framework.

Project Requirements

- Contract Law Expertise: A comprehensive understanding of Contract Law, particularly in relation to contract formation, is essential.
- Responsibility and Time Management: Commitment to assigned tasks and timely completion is crucial.
- Basic IT Knowledge: Familiarity with computer systems and basic IT concepts is preferred.
- Communication and Teamwork: Strong communication skills and ability to collaborate effectively within a team are important.
- Pass the pre-test before starting the real annotation work.

Data Annotation Outcomes

- Publication: The annotated dataset will be used for benchmarking and may be published in a journal or presented at a conference.
- Further Research: The annotated data will serve as a resource for subsequent machine learning research.

Benefits

- Research Assistant Experience: Opportunity to work as a Data Annotator on a research project.
- Flexibility: Remote work with flexible hours.
- Compensation: RM 30 per hour.

Annotation Tasks

- Evaluation of Legal Scenarios: Analyse and evaluate legal scenarios as per the IRAC framework.
- IRAC Analysis for Contract Formation: Apply IRAC methodology to analyse contract formation in provided scenarios.
- Decomposed Questions and Court Case References: Generate relevant decomposed questions for each IRAC segment and include related court cases with page numbers.