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

Anonymous EMNLP submission

Abstract

The effectiveness of Large Language Models (LLMs) in legal reasoning is often limited due to the unique legal terminologies and the necessity for highly specialized knowledge. These limitations highlight the need for high-quality data tailored for complex legal reasoning tasks. This paper introduces LEGALSEMI, a benchmark specifically curated for legal scenario analysis. LEGALSEMI comprises 54 legal scenarios, each rigorously annotated by legal experts, based on the comprehensive IRAC (Issue, Rule, Application, Conclusion) framework. In addition, LEGALSEMI is accompanied by a structured knowledge graph (SKG). A series of experiments were conducted to assess the usefulness of LEGALSEMI for IRAC analysis. The experimental results demonstrate the effectiveness of incorporating the SKG for issue identification, rule retrieval, application and conclusion generation using four different LLMs. LEGALSEMI will be publicly available upon acceptance of this paper.

1 Introduction

Access to justice is a universal social challenge. 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 completely resolved [1]. In India, more than 10,490 legal cases in the Supreme Court of India have been pending for more than a decade (Madhana and Subhashree, 2022). These backlogs are often caused by the complexity in legal practice, as well as the scarcity of legal professionals. IRAC framework (Metzler, 2002), stands for issue, rule, application, and conclusion, is the problem solving framework widely used by legal professionals to determine the underlying legal issues, followed by extracting and transforming facts in a legal scenario for legal reasoning, which eventually leads to 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 remains a challenge for LLMs to perform IRAC analysis on legal scenarios accurately. A recent study (Kang et al., 2023) identifies two key problems in analysing legal scenarios. First, ChatGPT draws wrong conclusions on approximately 50% of the legal scenarios on average. Even if the conclusions are correct, there are mistakes in the intermediate reasoning steps. Secondly, ChatGPT is not able to cite correct legal rules when analysing majority of the legal scenarios. In real world, it is crucial for legal professionals to understand every single reasoning step that leads to the final conclusion. In addition, our empirical study finds that LLMs struggle to cope with the language gaps between legalese and everyday language. We conjecture that LLMs still cannot fully comprehend the underlying legal knowledge and perform complex legal reasoning accurately.

Recent advances show that it is possible to mitigate the hallucination problem of LLMs by leveraging structured knowledge graphs (SKGs) (Pan et al., 2024a). 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 curate LEGALSEMI, a dataset comprising legal scenarios pertaining to Malaysian Contract Law, accompanied by rich structured IRAC analysis carried out by top law students, as illustrated in Fig. 1. Compared to (Kang et al., 2023), we do not only extend their dataset by doubling the legal scenarios in Malaysian Contract Law but also introduce new

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

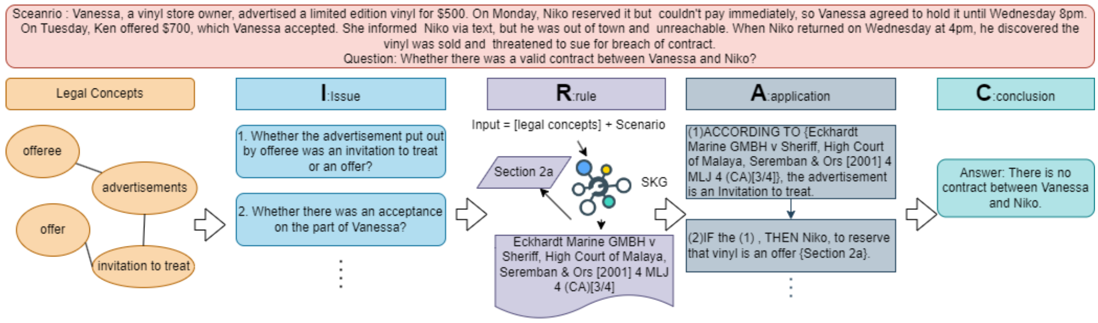


Figure 1: An example of a legal scenario pertinent to Malaysian Contract Law with annotations for IRAC analysis. The new types of annotations are legal concepts, court cases, and links to SKGs.

annotation types to all 54 scenarios, including legal concepts and court cases.

To support reasoning with structured legal knowledge, we extract semantic information from a law textbook and a legislation *automatically* to build the 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 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 issues, followed by retrieving rules and performing legal analysis to address each issue. Incorporating legal concepts from the SKG, the quality of the issue generation is improved by over 21.4% across all evaluated LLMs. The annotated issues subsequently lead to significant improvement on application generation.

- 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. We find out that legal concepts are significant 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.

- Our findings indicate that while LLMs are adept at identifying high-level legal concepts, there is still a need for improvement in recognizing the details of these concepts.

2 Dataset

IRAC provides a comprehensive problem-solving framework for legal professionals. It takes four stages to transform facts acquired from a legal scenario into legal conclusions: (i) identifying legal issues, (ii) determining the legal rules and precedents pertinent to the issues, (iii) performing analysis by applying the law to the facts and the issues, which requires strong legal reasoning skills, and (iv) drawing conclusions based on the analysis. Legal reasoning is *defeasible* such that there are often more than one reasoning traces leading to the same conclusion or different plausible reasoning traces lead to different reasonable conclusions (Billi et al., 2021). Given a legal scenario, legal professionals' concern is not just about the final conclusion, but also why the conclusion is drawn. Therefore, it is essential to build automatic IRAC analysis tools that produce outcomes for each stage and help them identify any missing reasoning steps, and suggest alternative analysis, when necessary.

While LLMs demonstrate a great potential to automate IRAC analysis in the absence of supervised training data, they suffer from three key limitations: i) wrong references to statutes and precedents, ii) weak legal reasoning capability, and iii) difficulties in filling the gaps between legalese and everyday language. In contrast, prior studies demonstrate the effectiveness of utilizing Retrieval-Augmented Generation (RAG) and neuro-symbolic approaches with knowledge bases to enhance the factuality and reasoning capability of LLMs (Gao et al., 2023). Therefore, LegalSemi builds the *first* SKG as a legal knowledge base to facilitate research on neuro-symbolic approaches for legal reasoning and provides an annotated corpus to evaluate system outcomes for each stage of IRAC.

2.1 Structured Knowledge Graphs

Neuro-symbolic systems have garnered increasing interest due to their ability in enhancing the reasoning capabilities of deep neural networks by incorporating symbolic reasoning, such as logic. Recent advances indicate that it is possible to mitigate the hallucination problem of LLMs and enhance the factual accuracy of their responses by incorporating knowledge graphs (KGs) (Pan et al., 2024b). These approaches are considered neuro-symbolic because KGs essentially implement the principles of description logic (Baader et al., 2017).

We consider Malaysian Contract Law as the target area of law due to the importance of contracts in everyday life. The corresponding SKG is automatically constructed from the textbook “Law for Business” (Trakic et al., 2022), the Contracts Act 1950 (the primary legislation governing contracts in Malaysia), and 76 court cases pertinent to contracts downloaded from Malaysia e-judgement [2]. It is easy to implement rules to extract legal knowledge from legal documents because The layout of a legal document often resembles the structure of legal knowledge, as evident by the screenshots of the textbook in Appendix E.

Legal concepts serve as the building blocks of legal doctrine, often act as bridges that connect related knowledge from diverse sources. For example, under the Contract Act 1950, *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 linked to paragraph P4-014 in the text book via the legal concept *"offer"*.

We derive the skeleton of the SKG from the textbook and enrich the skeleton with statutes from the primary legislation. The index of the book organizes the key concepts of contract law hierarchically, as illustrated in Appendix E. We extract those concepts from the index and annotate them as nodes at the corresponding levels, such as *main_concept* and *subconcept*. The children nodes are linked to their parent nodes using the relation *subconcept_of*. Additionally, we represent each chapter title as a node, indicating specific aspects of the Contract Act 1950, such as *Void Agreements*.

[2]e-Judgement: https://cms2.kehakiman.gov.my/CommonWeb/ejudgment/SearchPage.aspx?JurisdictionType=ALL

Furthermore, we extract the titles, section titles etc. from the Contracts Act 1950 and represent each as a node. Then we introduce several relations to associate the nodes derived from the book with the relevant ones in the legislation. For example, each chapter is associated with the relevant sections of the legislation. Figure 2 shows a snippet of the SKG.

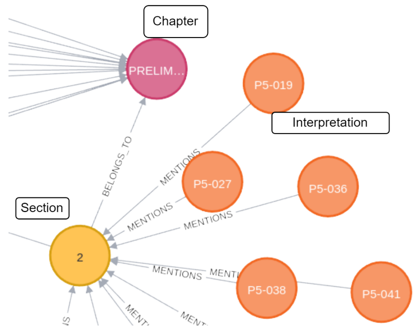


Figure 2: A snippet of the SKG.

To bridge the language gap between legalese and plain English, we extract interpretations from the book that provides layman explanations of the corresponding statutes in the legislation. Each interpretation is represented as a node in the SKG, and the *mentions* relation is used to link an interpretation to the relevant statute. Overall, the SKG comprises 3,114 nodes and 1,811 edges, stored in Neo4j for easy data exchange. Further details about the SKG can be found in Table 4 in Appendix C.

2.2 Corpus Construction

From a pool of applicants from Malaysia, we carefully selected six data annotators to assist with scenario selection and annotating IRAC analysis. This annotator team comprises four second-year law students from three distinct Malaysian universities and two junior lawyers. The annotated corpus comprises 54 legal scenarios covering five chapters in the textbook and 55 subtopics, ensuring extensive coverage of various aspects of Malaysian Contract Law. Each scenario is designed to reflect real-world legal problems. The rigorous annotation task is challenging, because it takes approximately *three* hours to annotate one scenario with legal concepts and complete IRAC analysis, though we develop an easy-to-use annotation tool (Appendix D), which significantly improve the productivity of annotators.

2.2.1 Scenario Selection

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 extracted from the subheadings of the text book.

First, we asked two annotators to create 24 scenarios which were modified from tutorial questions, books, and past exam questions. Next, for the remaining subtopics, we utilized 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, another two of the six law students evaluated the quality of the candidate scenarios using the following questions, as shown in Fig. 3.

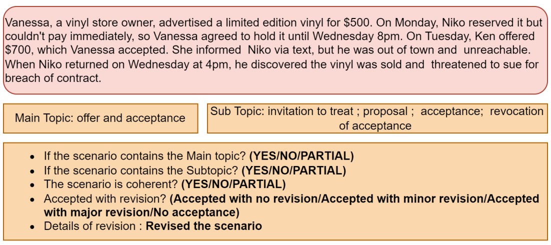


Figure 3: A scenario with quality assessment questions.

Based on annotators' feedback, our experts revised all 54 scenarios to ensure their quality. Their average length is 800 words.

2.2.2 IRAC Annotations

Legal concepts act as a bridge, connecting the facts in a scenario to the professional legal knowledge, including statutes and precedents. Therefore, we first annotated legal concepts, followed by performing IRAC analysis. The detailed annotation guideline is provided in Appendix A.

Legal Concepts. Using the legal concepts in the SKG, annotators were asked to identify and highlight relevant legal concepts within a given scenario. They were instructed to look up legal concepts first in the textbook "Law for Business", which ensures that the identified legal concepts are part of the SKG. If a concept, such as "*offeror*", is not absent from the textbook, they are allowed to add new concepts into the SKG.

Issues. A legal issue is a point of dispute that involves the interpretation, application, or violation of laws and regulations. Six annotators were asked to identify the issues in given scenarios. Across the scenarios, the main question is whether there is a valid contract between involved parties. To answer the question, the target problem is decomposed into issues, which are articulated in the style of questions. For example, an issue in our example scenario (Fig. 1) is *"Whether there was an acceptance on the part of Vanessa?"*.

Rules. A rule specifies the laws applicable to the issues. The annotators are asked to locate the appropriate cases and/or statute sections from the Contract Act 1950 pertinent to issues. For statutory law, the annotation tool offers a drop-down menu to select relevant sections from the 280 sections available in the SKG. For case law, the tool includes text fields for related court cases along with the corresponding page numbers in the cases. For instance, *Eckhardt Marine GMBH v Sheriff, High Court of Malaya, Seremban & Ors [2001] 4 MLJ 4 (CA) [3/4]*. To enable reuse and reference of those rules, the provided cases are displayed as buttons in the user interface so that annotators can refer to those cases by clicking on the buttons (see Fig. 9 in the Appendix D).

Application. In the Application section, annotators applied the rules identified in the rules section to the specific facts of the issues in a given scenario step by step. They are encouraged to use the conditional statements in form of "IF...THEN..." to articulate each reasoning step. Figure 7 in Ap-

	No. Scenario	Full IRAC	Legal Concept	Rules	Annotated Application	SKG
SIRAC	40	yes	0	58	Yes	No
LEGALSEMI	**54**	**yes**	**297**	**90**	**Yes**	**Yes**
SARA_entailment	277	no	38	9	No	No
SARA_numeric	100	no	38	9	No	No
LEGAL BENCH	59	no	0	18	No	No

Table 1: Comparison between LEGALSEMI and the most relevant datasets.

pendix B illustrates tje application section of our example. As legal reasoning is defeasible, annotators can make assumptions due to incomplete information. Different assumptions may lead to different conclusions, thus it is essential to discuss and justify these assumptions in the corresponding reasoning step. The application is the most important and challenging section of an IRAC because it develops the answer to the issue at hand.

Conclusion. The conclusion section directly answers the questions in the issue section, without introducing any new rules and analysis. Following the common practice in law, annotators were asked to write the full sentence of a conclusion, such as *" There is no contract between Vanessa and Niko."*

2.2.3 Data Quality Assurance

Given a scenario, there are many plausible IRAC analysis, because different assumptions and different interpretations of rules may lead to different conclusions. As it takes roughly three hours to perform a single IRAC analysis and it is infeasible to annotate all possible IRAC, we verified the quality of an IRAC analysis by asking another annotator to act as an evaluator. Specifically, an evaluator can either agree, disagree, or partially agree with an IRAC analysis. The ratio of overall agreements across all 54 scenarios exceeds 0.8, indicating a high level of annotation quality.

Annotator Quality. IRAC analysis is challenging. Hence, as mentioned above, we selected only annotators who have a strong legal background. The law students were required to have achieved at least a B grade in related law subjects. All annotators must pass a specialized pre-test before being recruited. Financial compensation is MYR30 per hour, above the minimum wage MYR7.21 in Malaysia, reflecting the complexity and rigour of the annotation tasks.

2.2.4 Summary of the Corpus

Our corpus comprises 54 scenarios, 243 issues as decomposed questions, 197 mentions of legal concepts (70 of them are unique), 268 sections of the Contracts Act 1950 (44 of them are unique), 76 court cases, and 607 reasoning paths. On average, each application involves 11.25 reasoning steps to draw a conclusion for the main questions. The most common legal concepts encountered include "offeror", "offeree", and "proposal", reflecting the frequent focus on contract formation. Similarly, the law sections most often cited are s2(a), s2(d), s7(a), s2(e), and s2(b).

Dataset Comparison. We compare our corpus with the publicly available corpora in Table 1. Among them, SIRAC (Kang et al., 2023) is the only one that includes annotations of full IRAC analysis for legal scenarios. LEGALSEMI improves upon their work by i) adding a SKG to support neuro-symbolic approaches, ii) introducing annotations of legal concepts to facilitate evaluation of neuro-symbolic approaches, iii) decomposing main legal questions into scenario-based issues, instead of using fixed issues across scenarios as in SIRAC, and iv) include a test set with longer scenarios, closer to the real world scenarios, that require more complex reasoning.

SARA (Holzenberger and Van Durme, 2021) focuses on legal question answering in Taxation Law. It annotates structured reasoning paths involving merely seven rules in total for the QAs but does not include any annotations of IRAC. LEGAL BENCH (Guha et al., 2022) covers diverse legal AI tasks but does not include full IRAC analysis, particularly the detailed analysis in Application. Instead, it designs simple QAs or classification tasks for a certain IRAC stage or reasoning steps, such as asking if a specific rule is relevant to a simplified scenario or not.

3 Experiments and Results

We empirically demonstrate the usefulness of LEGALSEMI for IRAC analysis and highlight the open challenges for future research.

3.1 Legal Concept Identification

Legal concepts play a key role in neuro-symbolic approaches for legal reasoning by linking facts in scenarios with legal knowledge. We investigate how accurate the state-of-the-art (SOTA) LLMs can identify legal concepts in scenarios, because LLMs demonstrate remarkable performance on zero-shot and few-shot learning (Brown et al., 2020).

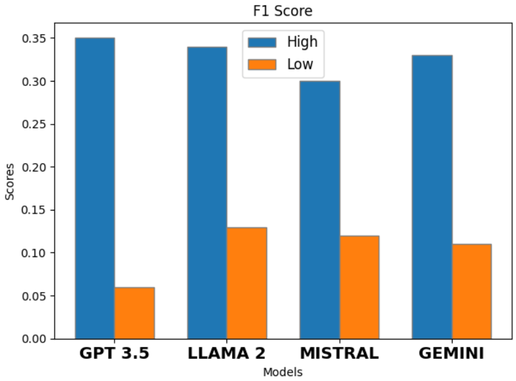


Figure 4: Comparison of F1 Score for predicting both high-level and lower-level concepts.

We adopt four LLMs for legal concepts identification: GPT-3.5 TURBO, LLAMA 2, MISTRAL, and GEMINI. The configurations of those models are detailed in Appendix G. Our prompt for those models is shown in Fig. 13. It begins with instructing the LLM to select relevant concepts from a comprehensive list of concept candidates, followed by providing a scenario and main legal questions. At the end of the prompt, it specifies the output format to a Python list for easy post-processing.

Evaluation Details. We extract a list of legal concepts from each model output, and compare them with the ground truth concepts in terms of precision, recall and F1 score. As the concepts are organized into a hierarchy in the textbook, we report the results for top-level and lower-level concepts, respectively, in order to highlight open challenges.

Results. As illustrated in Figure 4, all four LLMs perform significantly better at predicting top-level concepts compared to the lower-level ones. For the top-level concepts, GPT-3.5 TURBO achieves the highest precision (35%), while GEMINI obtains the highest recall (93%). We conjecture that compared to top-level concepts, e.g. "invitation to treat", the lower-level concepts associate with specific details of contracts, such as "audition" and "advertisement". Hence, they appear less frequently in the pre-training data of LLMs. This sheds light on the importance of constructing dedicated supervised training data for future research.

3.2 Issue Identification

Prior works (Kang et al., 2023; Guha et al., 2022) employ a set of fixed issues to decompose a main legal questions into simpler questions. Since issues can vary significantly from case to case in practice, we investigate the extent to which LLMs can generate scenario-based issues and identify the helpfulness of legal concepts at this stage.

We apply the same four LLMs as in concept identification for issue generation. The prompt instructs LLMs to break a main legal question into a list of issues by leveraging relevant ground-truth legal concepts. In the prompt, we also instruct an LLM to self-evaluate its outputs by ensuring they are *reasonable*, as inspired by (Hao et al., 2023). The prompt is detailed in Appendix H.

Evaluation Details. As issue generation is a language generation task, following (Kang et al., 2023), we apply GPT-3.5 TURBO to compare predicted issues with annotated reference issues with a list of criteria detailed in the prompt (see Appendix H). An LLM is expected to select one option from: strongly agree, neutral, or disagree, which is further mapped to a score of 1, 0, and -1, respectively.

To investigate the quality of this automatic metric, we compare the results of the automatic evaluation with those of human evaluation. In the human evaluation, we assess the quality of an IRAC analysis using a rubric that is widely used in Malaysian contract Law courses (Gerhardt, 2008; Carter, 2006). The issues of an IRAC analysis receive a *Pass* when they satisfy the corresponding criteria detailed in Appendix F.1, otherwise, they are marked as *Fail*.

We ask two annotators to independently assess the issues generated by two best performing models (GEMINI and GPT-3.5 TURBO) in three different configurations (e.g. with or without legal concepts) on 10 randomly selected scenarios. In case of disagreement between their assessments, we ask the most experienced annotator with a strong legal background to resolve it. Each generated output marked as *Pass* receives a score of 2, otherwise, a score of 1. We then rank all LLM configurations according to their average scores and compute the Spearmann rank correlation with the counterpart using the automatic metric. A strong correlation of 0.8 suggests the effectiveness of the automatic metric. Further details are in Appendix F.

Results. Figure 5 depicts the average scores of the automatic metric across all 54 scenarios. Legal concepts are beneficial for all LLMs especially for GPT-3.5 TURBO, which increased by 21.4%. The self-evaluation instruction further enhances the performance of all LLMs, with the best per-

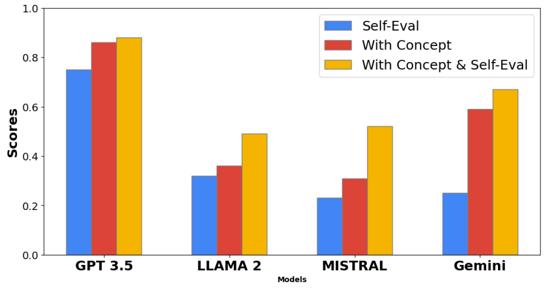


Figure 5: The results of issue identification.

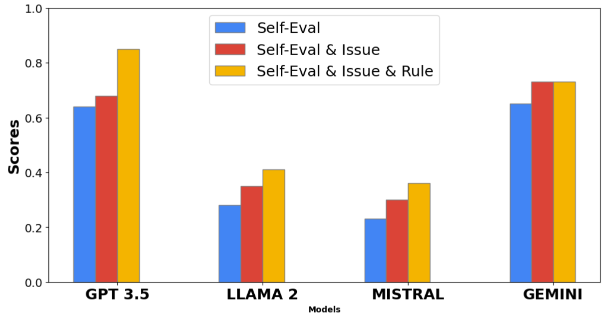


Figure 6: Results of application generation.

formance observed when both legal concepts and self-evaluation are combined.

3.3 Rule Retrieval

Given a scenario annotated with legal concepts, we investigate what information in the SKG is beneficial for retrieving relevant legal rules from the Contract Act 1950. A key challenge herein is the gaps between the lay language used in scenarios and the legalese used to express legal rules. Given a scenario as the query, we apply a TF-IDF based search engine (Pedregosa et al., 2011) to retrieve rules indexed by three types of representations, detailed below. Those experimental results are compared with the setting that only the legal rules associated with the same legal concepts pertinent to the scenario are considered for retrieval. This is achieved by performing retrieval in two stages: i) sending the legal concepts of a scenario as the structured query to the SKG to identify the set of legal rules associated with those concepts, and ii) using the TF-IDF based search engine to rank the legal rules in the results of the initial retrieval.

Indexing. We consider three representation types of a legal rule for indexing: i) original legalese, ii) interpretation extracted from the textbook, and iii) combination of the interpretations from the textbook and the additional interpretation generated by GPT-3.5 TURBO (detailed in Appendix H), because the textbook covers only 18.5% of the legal rules.

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

Results. Table 2 shows that the retrieval of legal concepts can be used effectively as the search scope for scenario-based retrieval, where it boosts both precision and recall of rule retrieval. Using the interpretations from the text book as index further improves the retrieval quality overall. The highest F1 score at top-5 reaches 16.3%. In contrast, direct application of those LLMs for rule generation falls below 3% at top-5. Those automatically generated interpretations are beneficial only in terms of recall. In addition, it can be observed that the quality of the generated interpretations are often prone to the hallucination problem of GPT-3.5 TURBO.

3.4 Application Generation

We investigate the effectiveness of utilizing issues and rules for application generation. We reuse the same four LLMs in the previous stages and apply the prompt (Figure 16 in the Appendix H) to generate legal analysis.

Evaluation Details. Similar to issue generation, we apply GPT-3.5 TURBO to compare the generated application sections with the annotated reference for each scenario. The possible outcomes of the evaluation include strongly agree (1), neutral (0), or disagree (-1). We also conduct a similar human evaluation to assess the quality of this automatic evaluation metric, and obtain a correlation of 0.86. The details are covered in Appendix F.

Results. Figure 6 shows the results when the prompts were incrementally added with issues and rules. Notably, all LLMs greatly benefit from the identified issues except for GEMINI. GPT-3.5 TURBO shows the most significant increase in performance, with an improvement of 18.9% from the self-evaluation prompt to the self-evaluation prompt with issues and rules. These results highlight the effectiveness of incorporating rules and issues in the legal reasoning steps.

3.5 Conclusion Generation

We apply the same LLMs to generate conclusions, based on given scenarios, main questions, and refer-

No initial retrieval	index: legalese			index: interpret (text book)			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%

Initial retrieval with legal concepts	index: law			index: interpret (text book)			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 2: Results for rule retrieval. GPT_interpret denotes using the interpretations generated by GPT-3.5 TURBO.

ence application sections. The details of the prompt can be found in Appendix H.

Evaluation Details. Same as the previous IRAC stages, we apply GPT-3.5 TURBO to evaluate conclusions by comparing them with ground truth, as detailed in Appendix F. The correlation with the human evaluation results is 91%, as presented in Appendix F.

Results. Figure 21 in Appendix H illustrates the automatic evaluation results, which indicate improvements across all models. Notably, GPT-3.5 TURBO and GEMINI exhibit substantial enhancements. ChatGPT's score improved by 71.4%. These significant gains underscore the effectiveness of incorporating the application component into the prompt, aligned with findings from (Kang et al., 2023).

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 2, for domain-specific tasks in the legal field. LEGAL BENCH (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) prompting (Wei et al., 2022) is employed. Hu et al. (2023) applied LLM to generate a reasoning chain along with the final answer given the legal question. 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 artificial intelligence (AI) for legal tasks 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. For legal reasoning, we need highly specialized legal knowledge to ensure accurate answers. Unfortunately, the current approach mainly focuses on common sense knowledge.

5 Conclusion

We introduce LEGALSEMI, which consists of 54 scenarios annotated with IRAC analysis in Malaysian Contract Law, and a SKG for legal knowledge extracted from a law textbook and legislation. The SKG covers legal concepts, legal rules, interpretations in lay language and their relations. Legal concepts from the SKG are particularly useful for improving the quality of issue generation, which in turn significantly enhance legal analysis in Application across all four evaluated LLMs. Besides, 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 significant improvement in rule retrieval, with an increase of 17.2% in F1 score at top-5.

Ethical Statement

Our research practices align with the principles of the ACL Code of Ethics. Our investigation complies with these ethical standards. LEGALSEMI was created and evaluated with a keen awareness of ethical considerations, especially regarding the involvement of human annotators. We recognize that the necessity for human-annotated data to train conditional independence classifiers in our method demands significant effort. We have taken careful measures to ensure that this process is ethically sound, honoring the annotators' contributions by respecting their time and providing equitable compensation. Moreover, the central objective of LEGALSEMI is to create an IRAC methodology-based benchmark. It is designed without generating any information that could be deemed harmful or violate privacy.

Limitation

In this study, our primary emphasis revolves around examining scenarios that pertain specifically to the 'Formation of Contract' as delineated within Malaysian 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 related 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

Franz Baader, Ian Horrocks, Carsten Lutz, and Uli Sattler. 2017. *Introduction to description logic*. Cambridge University Press.

Marco Billi, Roberta Calegari, Giuseppe Contissa, Francesca Lagioia, Giuseppe Pisano, Galileo Sartor, and Giovanni Sartor. 2021. Argumentation and defeasible reasoning in the law. *J*, 4(4):897–914.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

John W Carter. 2006. Carter's guide to australian contract law. *(No Title)*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Michael J Gerhardt. 2008. How a judge thinks. *Minn. L. Rev.*, 93:2185.

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*.

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. 2024a. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Laerd Statistics. 2013. Spearman's rank-order correlation. *Laerd Statistics*, page 33.

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 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. You need to have B+ and above for the related subject.
- 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.

B Examples of the Application annotation.

Figure 7 exemplifies our annotation process for legal scenario analysis using the IRAC methodology. It demonstrates the structured approach we take to break down and evaluate each aspect of a legal problem. The figure uses logical steps to progress from identifying an initial legal issue to applying relevant rules and statutes, analyzing the facts, and drawing a conclusion. Each step is clearly annotated with references to legal cases and statutes, ensuring that the reasoning is well-supported and transparent. The annotations also include conditional statements and assumptions, highlighting how various legal principles and factual circumstances are considered to reach a final conclusion.

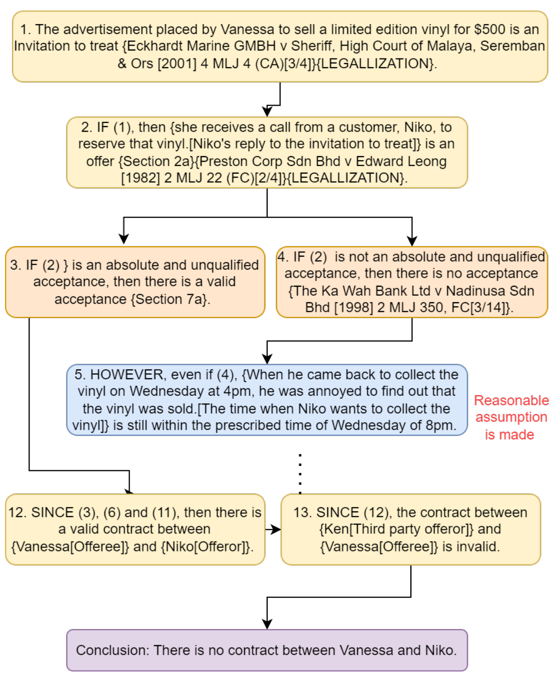


Figure 7: An example of the application section.

Table 3 lists all the condition types used in the annotation. We have a total of six different condition types. The most commonly applied condition type is the IF...THEN... structure.

C 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 details of the SKG is shown in the Table 4. Figure 2 shows partial of the graph. This graph illustrates the structure of a Social Knowledge Graph (SKG) in Neo4j, showcasing the relationships between various sections, chapters, interpretations, and main concepts through nodes and edges.

Figure 8The diagram illustrates a section of the SKG, demonstrating the hierarchical and relational structure of legal statutes. The central pink node represents a Chapter, which connects to various Section nodes (yellow) through "BELONGS_TO" relationships. Each Section, like Section with a Title (dark brown) via "HAS_TITLE" relationships and has Interpretations (orange) connected by "HAS_INTERPRETATION" links. These Interpretations, such as P4-109, detail specific provisions and connect to main Concepts (green) and sub-concepts (light brown).

D Annotation Tool

To facilitate this intricate annotation process, we developed an online data annotation platform,

Conditional Type	Use	Example	Count
If... Then	Used to state a condition and its consequence	"IF {She placed an advertisement on a social media platform selling a limited edition vinyl for the price of $500 [advertisements: invitation to treat]} is an invitation to treat, THEN the call from a customer, Niko, to reserve that vinyl is an offer."	54
According to...	Used to refer to legal cases, statutes, or authoritative sources	"ACCORDING TO {Eckhardt Marine GmbH v Sheriff (2001) 4 MLJ 49[53-54]}, IF {Lowel[offeror]} puts up {advertisement[Advertisements to treat]}, THEN it is not an offer but an invitation to treat."	39
Since... Then	Used to show a reason and its result	"SINCE {Lowel responded by saying that 500 is too cheap, and that the lowest price she is willing to sell is RM 700 .[Counter offers of initial proposal by proposee]} is not absolute and unqualified THEN there is no valid acceptance."	52
However... If... Then	Used to introduce an exception or a contrasting condition and its consequence	"{HOWEVER} IF the agreement between Penny and Tina is not supported by considerations THEN it will be void even if it is supported by intention to create legal relation."	42
Even if... Then	Used to express a consequence that applies despite a condition	"EVEN IF {Nina[Capacity contracting with]} has lied about her age, she would not be estopped from pleading incapacity."	10
Only if... Then	Used to indicate that a consequence will occur solely under a specific condition	IF (7) THEN agreement is supported by intention ONLY IF stated condition is fulfilled. {Confetti Records (A Firm) and others v Warner Music UK Ltd and another [2003] EWHC 1274 (Ch)[]}	12

Table 3: Application conditional types

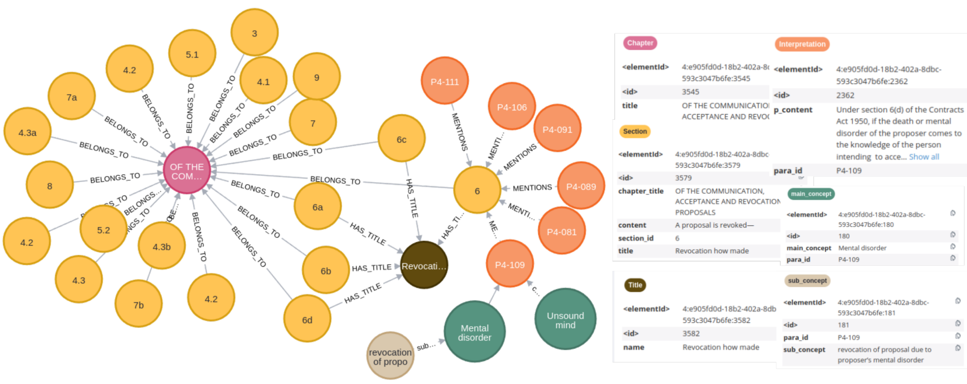


Figure 8: Details example of SKG in Neo4j.

Node name	Details	No of nodes	Example
Chapter	Each chapter covers a specific aspect of the Contract Act 1950.	24	Void Agreements
Title	Each title focuses on specific legal points within the Contract Act 1950.	210	Misrepresentation,Acceptance must be absolute
Section	Sections of the Contract Act 1950, providing detailed legal provisions and serving as the main reference for the statute.	304	Section 5.2, An acceptance may be revoked at any time before the communication ...
Interpretation	Interpretations of the contract law, automatically extracted from the book content. Each is labeled with its content and paragraph IDs.	1623	If a statement does not satisfy the elements of proposal as discussed&above, the statement would more likely be an invitation to treat or ....
Extend content	Footnotes or extensions of the interpretations, sharing the same paragraph ID as the related interpretation.	307	Partridge v Crittenden [1968] 1 WLR Facts: The defendant advertised the ...
Main concept	Key legal concepts summarized from the interpretations, auto extracted from the index of the book content.	189	proposal revocation
Sub Concept	Detailed extensions of the main concepts.	351	communication
Sub sub concept	More detailed information on the sub-concepts.	106	thrid party
	Total	3114	
Edge name	Details	No of eges	Example
Belongs_to	Connects the Section and Chapter.	304	chapter_title : OF CONTRACTS, VOIDABLE CONTRACTS AND VOID AGREEMENTS content: the court regards it as immoral, or opposed to public policy. section_id : 24e title;What considerations and objects are lawful, and what not
has_title	Connects the Section and Title.	304	Same as above
mentions	Connects the Interpretation and Section.	193	The definition of “agent” and “principal” is provided in section 135 of;the Contracts Act 1950... title “Agent” and “principal”
related_to	Connects the Interpretation and Extend Content.	307	Hughes v Metropolitan Rly Co (1877) 2 App ......
concept_of	Connects the Main Concept and Interpretation.	184	Acceptance concept_of Under section 3, an acceptance can.....
subconcept_of	Connects the Main Concept and Sub Concept.	364	communication sub concept of proposal revocation
subSubconcept_of	Connects the Sub Sub Concept and Sub Concept.	155	third paty sub sub concept of communication
	Total	1811	

Table 4: The table illustrates the structure of the Social Knowledge Graph (SKG) implemented in Neo4j. It includes detailed information about the nodes representing entities and edges depicting relationships between these entities. The nodes can represent various entities such as people, organizations, and concepts, while the edges capture the interactions and connections among them. Attributes associated with nodes and edges are also detailed, providing a comprehensive view of the SKG.

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. Figure 9 shows an example of the annotation.

E Textbook details

Figure 10 showcase various sections and subsections that illustrate the organization of legal knowledge within the textbook. The index of the book's structured format, including headings, subheadings, and bullet points, mirrors the hierarchical nature of legal documents, making it conducive for rule-based knowledge extraction. At the end of the index is a link to the paragraph on that legal concept.

F Human Evaluation

Three human evaluators participate in the evaluation session. We select 10 scenarios and two models (GEMINI and GPT-3.5 TURBO) with all the experiment settings for them to evaluate. We select GPT-3.5 TURBO and GEMINI since it has the best performance from the auto evaluation result. They attend a briefing meeting to discuss and clarify their understanding of the marking rubric. After the briefing, they independently evaluate the scenarios. The third evaluator, who is more experienced, serves as the final decision-maker in cases where the first two annotators disagree, ensuring the reliability and accuracy of the final results. This method follows the steps for identifying issues, which involve decomposing questions for this experiment. The remainder of the evaluation focuses on the application of these guidelines

F.1 Human Evaluation Guidelines

These guidelines are based on the marking rubric and evaluation criteria for contract law (Carter, 2006). They determine how and why the conclusion is made. The answer needs to demonstrate the facts, which include the description of the circumstances, the main questions, and the issues (decomposed questions) raised by the question. It should also cover the rules and principles related to the issues, as well as the conclusions drawn. The proper approach is similar to that of the court's judgments. The grade in figure 11 shows is a helpful step-by-step process to evaluate the legal reasoning process.

Human Evaluation Results Human evaluation is conducted for issue identification, application, and conclusion. These aspects require human judgment based on expertise and credentials to ensure accuracy and reliability in the evaluation process.

We compared the human evaluation results with the auto-evaluation results by examining the rankings of the models' outcomes.

Figure 12 displays the human evaluation results for all experiments. We calculate the Spearman's rank correlation coefficient (Statistics, 2013) for each experiment to assess the alignment between human and automated evaluations.

We rank models based on the 'Pass' rate and compare it to the 'Agree' rate from the automated evaluation matrix. For human evaluation, we use a five-level grading scale: 1 (Fail), 2 (Pass), 3 (Credit), 4 (Distinction), and 5 (High Distinction), and compare these rankings with the 'Agree' rate from the automated evaluation. In conclusion, we compare the entailment rate with the 'Agree' rate from the automated evaluation matrix.

From the result, the average Spearman's rank correlation coefficient, indicated by the red dashed line, is approximately 0.89. This average further emphasizes the overall strong positive correlation between human and automated evaluations across different criteria.

G Models Details

We apply four Large Language Models (LLMs): GPT-3.5 TURBO, LLAMA 2, MISTRAL, and GEMINI.

In the GPT-3.5 TURBO, our settings include a temperature of 0.7, a common and general setting for GPT models, balancing creativity and coherence in responses. We also set the maximum token count to 1000, allowing for extensive and detailed answers.

We choose LLAMA 2 70B, MISTRAL 7B, and GEMINI as comparative models to analyze their performance against GPT-3.5 TURBO Turbo in handling complex legal scenarios. This comparison aims to assess the efficacy of each model in terms of accuracy, coherence, and relevance to the given legal context. LLAMA 2 is selected for its extensive parameter count, which may enhance its ability to understand intricate details. MISTRAL 7B is known for its efficiency and speed, making it an interesting

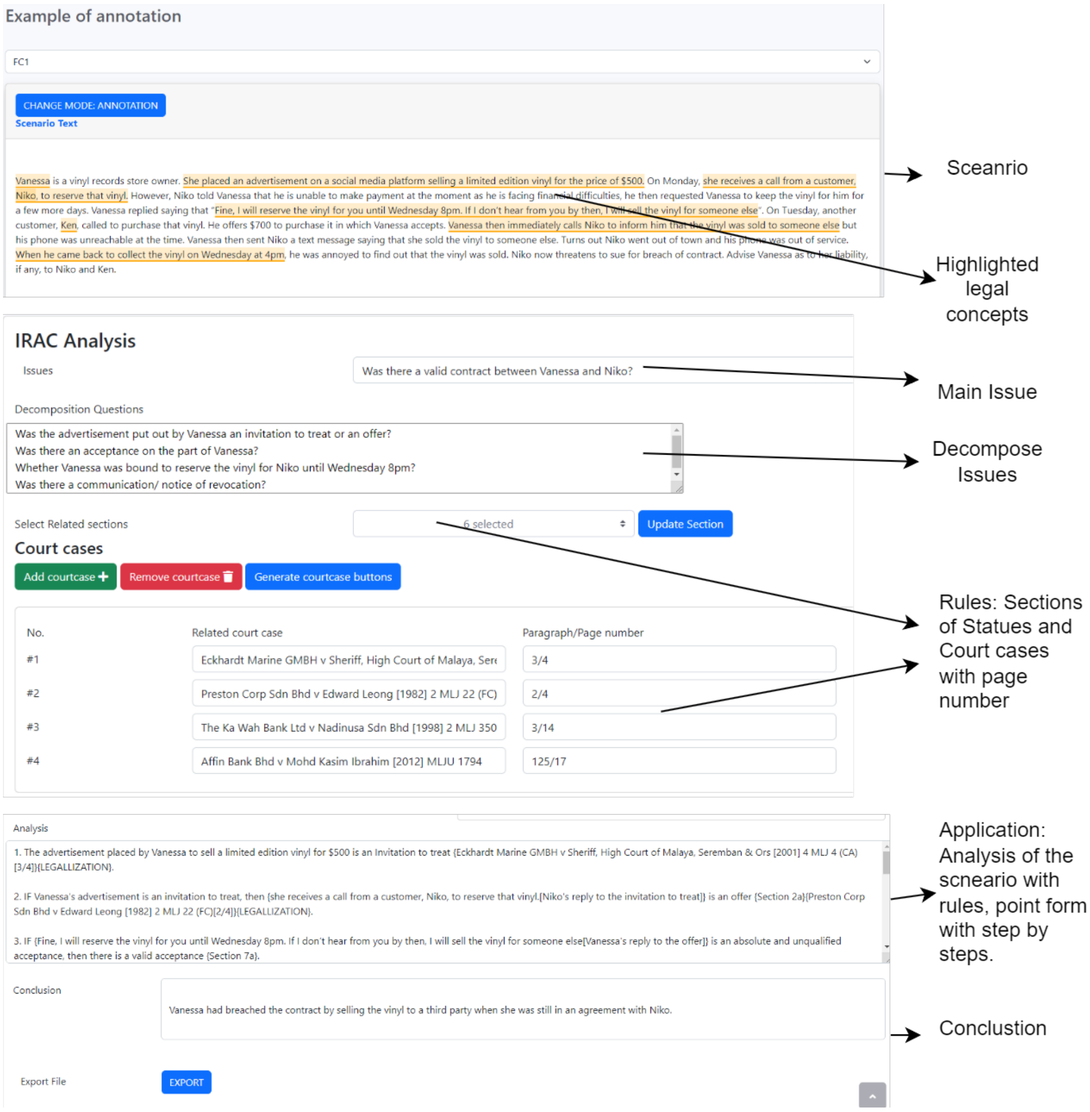
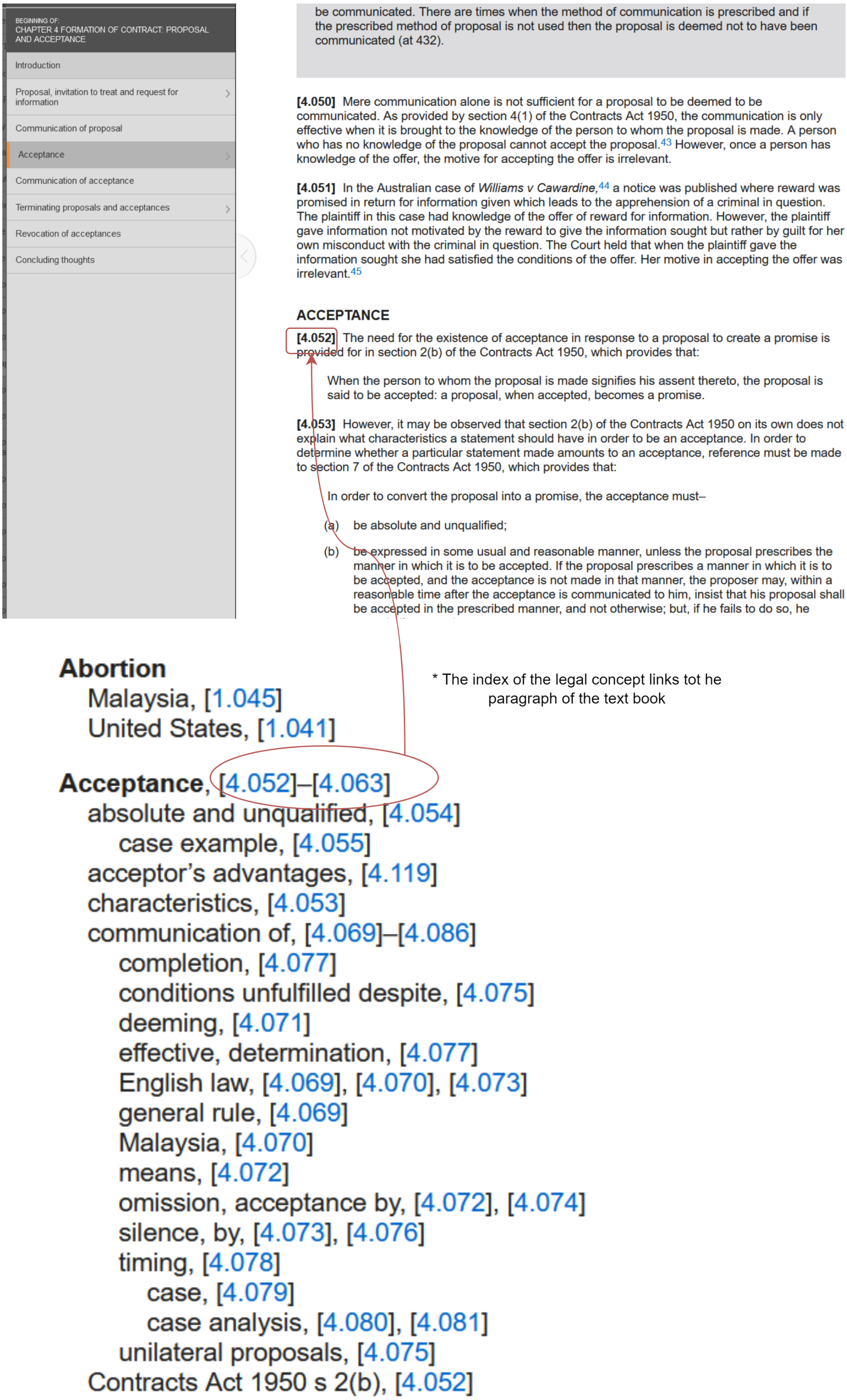


Figure 9: The web-based annotation tool developed to enable the legal scenario analysis.



BEGINNING OF:
CHAPTER 4 FORMATION OF CONTRACT: PROPOSAL AND ACCEPTANCE

- Introduction
- Proposal, invitation to treat and request for information
- Communication of proposal
- Acceptance
- Communication of acceptance
- Terminating proposals and acceptances
- Revocation of acceptances
- Concluding thoughts

be communicated. There are times when the method of communication is prescribed and if the prescribed method of proposal is not used then the proposal is deemed not to have been communicated (at 432).

[4.050] Mere communication alone is not sufficient for a proposal to be deemed to be communicated. As provided by section 4(1) of the Contracts Act 1950, the communication is only effective when it is brought to the knowledge of the person to whom the proposal is made. A person who has no knowledge of the proposal cannot accept the proposal.[43] However, once a person has knowledge of the offer, the motive for accepting the offer is irrelevant.

[4.051] In the Australian case of *Williams v Cawardine*,[44] a notice was published where reward was promised in return for information given which leads to the apprehension of a criminal in question. The plaintiff in this case had knowledge of the offer of reward for information. However, the plaintiff gave information not motivated by the reward to give the information sought but rather by guilt for her own misconduct with the criminal in question. The Court held that when the plaintiff gave the information sought she had satisfied the conditions of the offer. Her motive in accepting the offer was irrelevant.[45]

ACCEPTANCE

[4.052] The need for the existence of acceptance in response to a proposal to create a promise is provided for in section 2(b) of the Contracts Act 1950, which provides that:

> When the person to whom the proposal is made signifies his assent thereto, the proposal is said to be accepted: a proposal, when accepted, becomes a promise.

[4.053] However, it may be observed that section 2(b) of the Contracts Act 1950 on its own does not explain what characteristics a statement should have in order to be an acceptance. In order to determine whether a particular statement made amounts to an acceptance, reference must be made to section 7 of the Contracts Act 1950, which provides that:

> In order to convert the proposal into a promise, the acceptance must–
>
> (a) be absolute and unqualified;
>
> (b) be expressed in some usual and reasonable manner, unless the proposal prescribes the manner in which it is to be accepted. If the proposal prescribes a manner in which it is to be accepted, and the acceptance is not made in that manner, the proposer may, within a reasonable time after the acceptance is communicated to him, insist that his proposal shall be accepted in the prescribed manner, and not otherwise; but, if he fails to do so, he

* The index of the legal concept links tot he paragraph of the text book

Figure 10: The screenshot of the structure of the textbook.

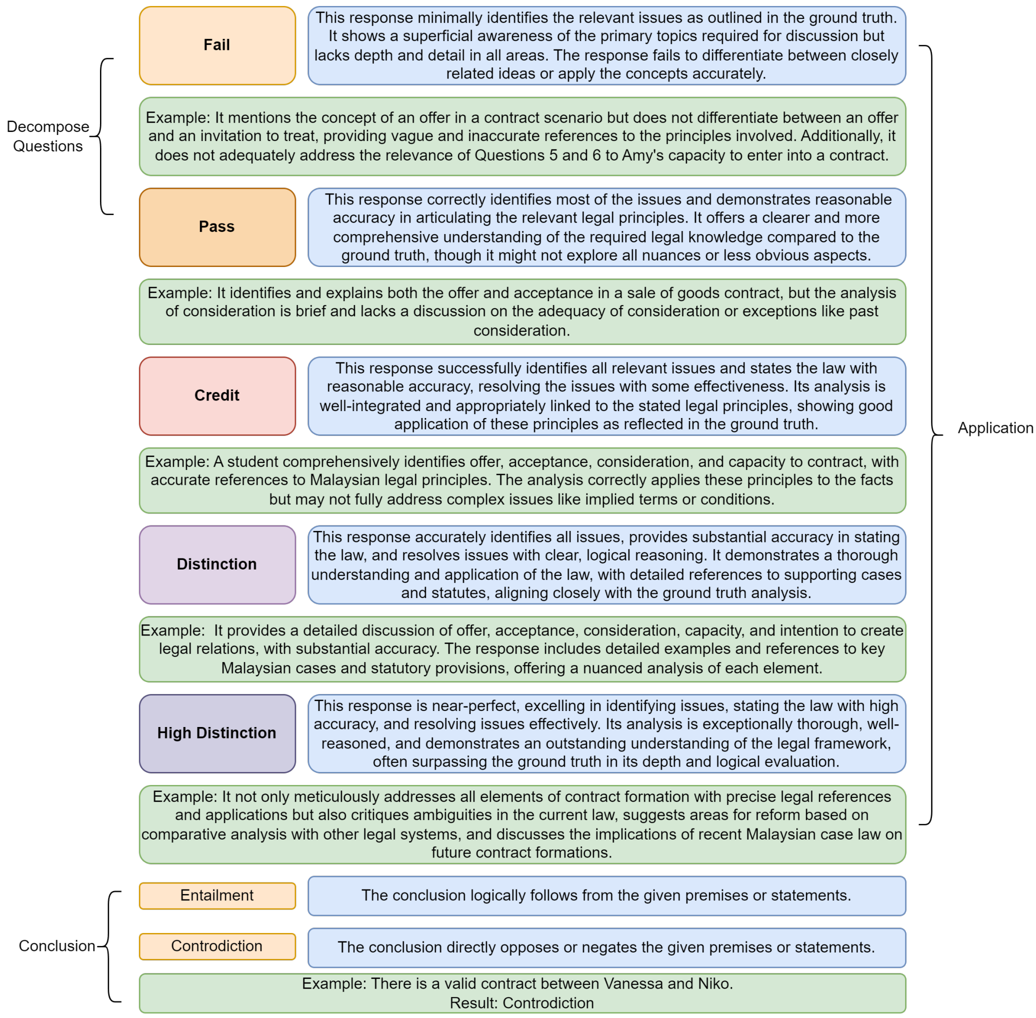


Figure 11: Human Evaluation Guidelines.

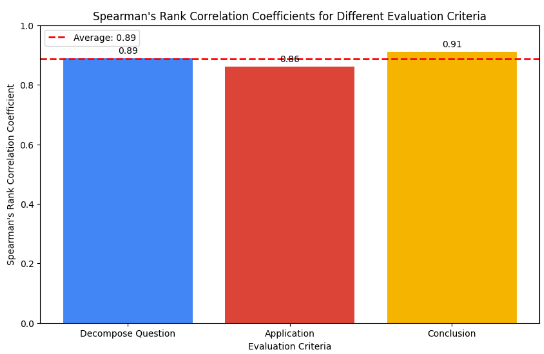


Figure 12: Human Evaluation Results.

contrast to the larger models. GEMINI is included for its promising performance in previous legal text analyses, providing a benchmark for evaluation. By comparing these models, we aim to determine which is best suited for tasks requiring precise legal understanding and reasoning.

H Experiment Details

Legal Concepts prediction Figure 13 displays the structure of the legal concepts prediction. The figure shows the different components of the prompt. For different experimental settings, we sometimes remove the legal concepts list from the potential legal concepts to compare the results.

Legal Concept Evaluation We use automatic evaluation metrics for this task. The outcome of the legal concept list is compared with the ground truth. The comparison is separated into two different levels: top-level and lower levels. The top level refers to more general concepts, such as "invitation

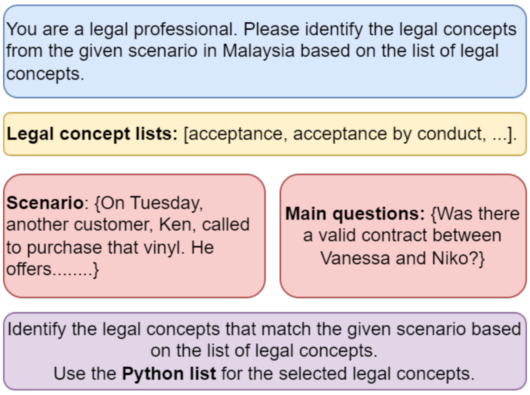


Figure 13: The prompt for legal concept identification.

to treat." The lower level includes more detailed aspects of the concept. For example, under "invitation to treat," there are specifics like "audition," "advertisement," etc.

To evaluate the accuracy of our predictions, we use precision, recall, and F1 score. Precision measures the proportion of correctly identified legal concepts out of all identified concepts. Recall measures the proportion of correctly identified legal concepts out of all relevant concepts in the ground truth, indicating the model's completeness. The F1 score provides a mean of precision and recall, offering a single metric that balances both aspects of accuracy.

Issue Identification Figure 14 displays the structure of the issue identification. The figure shows different components of the prompt. For different experimental settings, we sometimes remove the legal concepts list from the ground truth to compare the results.

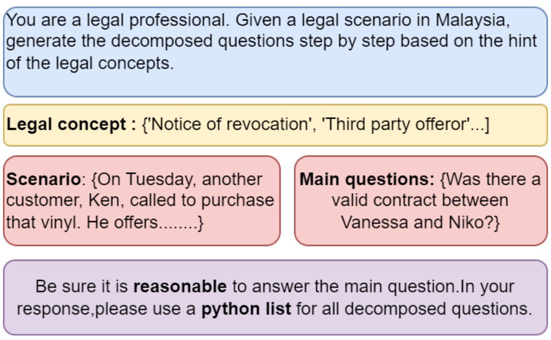


Figure 14: Details of the experiment of the decompose questions

Issue Identification Evaluation Figure 15 displays the structure of the issue identification evaluation prompt. The evaluation is based on the evaluation guidelines.

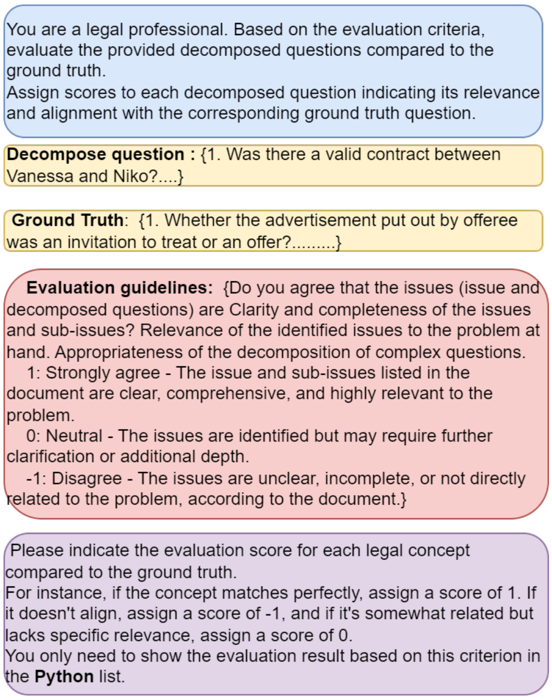


Figure 15: Prompts used for the automatic evaluation of the decomposed questions.

Application Generation Figure 16 displays the structure of the application generation prompt. The figure shows different components of the prompt. For different experimental settings, we sometimes remove the issues or self-evaluation prompt to compare the result.

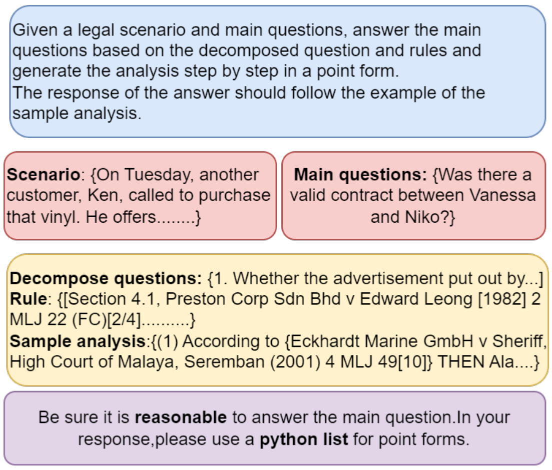


Figure 16: Application Prompt.

Application Generation Evaluation Figure 17 displays the structure of the application evaluation prompt. The evaluation is based on the evaluation guidelines.

Rule Retrieval: Details GPT-3.5 TURBO interpretation We generate the interpretation of

You are a legal professional, Evaluate the provided analysis in comparison to the ground truth based on the evaluation criteria. Assign scores for the analysis and alignment with the corresponding ground truth analysis.

Analysis : {1. Was there a valid contract between Vanessa and Niko?....}

Ground Truth: {1. Whether the advertisement put out by offeree was an invitation to treat or an offer?.........}

Evaluation guidelines: Do you agree that the analysis are Thorough examination of the facts and application of the law to the facts.Consideration of counterarguments or alternative interpretations. Transparency regarding any assumptions made during the analysis.
1: Strongly agree - The document's analysis is logically structured and coherent, comprehensive, and makes reasonable assumptions. It addresses counterarguments and alternative interpretations.
0: Neutral - The analysis is logically structured but may lack comprehensiveness or may have some minor logical flaws according to the document.
-1: Disagree - The analysis lacks clear logical structure, comprehensiveness, or contains major logical flaws as indicated in the document.

Please indicate the evaluation score for the analysis compared to the ground truth. For instance, if the analysis matches perfectly, assign a score of 1. If it doesn't align, assign a score of -1; if it's somewhat related but lacks specific relevance, assign a score of 0. You only need to show the evaluation result based on this criterion in the **Python** list.

Figure 17: Application Evaluation Prompt.

the Malaysia Contract Act 1950 using GPT-3.5 TURBO to interpret the law. The interpretation is then compared with the interpretation from textbooks.Figure 18 display the structure of the prompt. Table 5 shows the example of the output of the interpretation.

Conclusion Generation Figure 19 displays the structure of the conclusion generation prompt. The figure shows different components of the prompt. For different experimental settings, we sometimes remove the application or self-evaluation prompt to compare the result.

Conclusion Generation Evaluation Figure 20 displays the structure of the conclusion evaluation prompt. The evaluation is based on the evaluation guidelines.

Conclusion Result Figure 21 display the result of the conclusion. The comparison is between adding the application and not adding the application in the prompt.

Please provide an explanation of a given law with illustration and a real-life example, structured as a Python list containing elements for the law, illustration, and example. The output a Python list that includes:
Illustration:A simplified explanation of the given law in common English and explanation to further clarify the law.
Example: A real-life scenario demonstrating the application of the law.

Example: Given Law:
"A “bailment” is the delivery of goods by one person to another for some purpose, upon a contract that they shall, when the purpose is accomplished, be returned or otherwise disposed of according to the directions of the person delivering them. The person delivering the goods is called the “bailor”. The person to whom they are delivered is called the “bailee”."

Example output : [{
Illustration: "Every person has a right to enter into any type of contract they decide\n to be appropriate for them. However, there are some circumstances where\n the law limits the ability of a party to enter into contracts. The \nreason for such limitation usually is based on policy. The requirement \nof capacity may be found in section 10(1) of the Contracts Act 1950, \nwhich provides:"
Example: "A 16-year-old high school student attempts to take out a loan for a large sum of money to start a business. The bank, upon discovering the student's age, declines to process the loan application, citing the student's lack of legal capacity to enter into such a contract according to the Contracts Act 1950. This act protects minors from being bound by contracts that they may not fully understand or that may not be in their best interest."}]

Figure 18: Rule interpretation Prompt.

{Given a legal scenario occurring in Malaysia and the main questions and analysis, please give a conclusion for the main question.

Scenario: {On Tuesday, another customer, Ken, called to purchase that vinyl. He offers........}

Main questions: {Was there a valid contract between Vanessa and Niko?}

Analysis: {1. IF {She placed an advertisement on a social media platform selling a limited edition vinyl for the price of $500[advertisements: invitation to treat]}is......ston Corp Sdn Bhd v Edward Leong [1982] 2 MLJ 22 (FC)[2/4]}{LEGALLIZATION}.
2. According to {The Ka Wah Bank Ltd v Nadinusa Sdn Bhd [1998] 2 MLJ 350, FC[3/14]}{Section 7a}, an a......]

Be sure it is **reasonable** to answer the main question. In your response, please use a **python** list for the conclusion,the conclusion should be in **one sentence.**

Figure 19: Conclusion Prompt.

section_id	content	interpreation	real life example
s_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;	A bailment is simply the ;transfer of possession and control of personal property (goods) from one ;person (the bailor) to another person (the bailee) for a specific purpose, ;with the understanding that the goods will b e returned to the bailor or otherwise disposed of according to their instructions once the purpose is fulfilled.	John wants to sell his car and approaches his friend Mark with an offer. John tells Mark that he is willing to sell his car for $10,000. In this scenario, John is making a proposal by expressing his willingness to sell his car at a certain price. Mark can either accept or reject the proposal.
s_2c	the person making the proposal ;is called the "promisor" and the person accepting the proposal is called the "promisee";	In a legally enforceable promise, there are two parties involved. The person who makes the offer is known as the 'promisor' and the person who accepts the offer is referred to \as the 'promisee.'	"In a real-estate transaction, the seller proposes to sell their property to the buyer at a specific price. The buyer accepts this offer, resulting in a legally binding contract between them. In this scenario, the seller is the 'promisor' because they made the offer, and the buyer is the 'promisee' because they acceptedthe offer.

Table 5: Example of the interpretation.

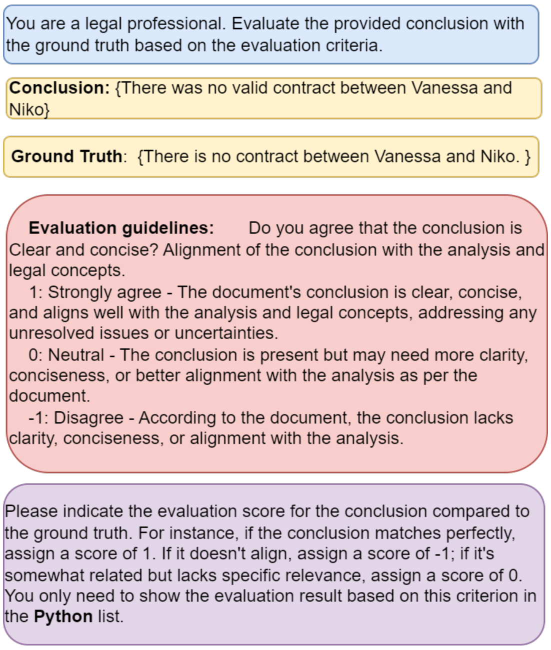


Figure 20: Conclusion Evaluation Prompt.

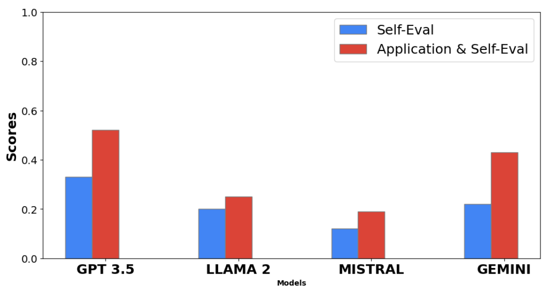


Figure 21: Result of Conclusion.