EVIA 2025: Proceedings of the Tenth International Workshop on Evaluating Information Access (EVIA 2025), a Satellite Workshop of the NTCIR-18 Conference, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002106

Structured Evaluation of Legal Reasoning in LLMs: Chain-of-Thought Prompting and Human Scoring for Retrieval Robustness

Ying-Chu Yu
College of Law, National Taiwan University
Taipei, Taiwan

Sieh-Chuen Huang
College of Law, National Taiwan University
Taipei, Taiwan

Hsuan-Lei Shao*
Graduate Institute of Health and Biotechnology Law, Taipei Medical University
Taipei, Taiwan

Abstract

This study investigates the legal reasoning abilities of Large Language Models (LLMs) in Taiwan's Status Law (family and inheritance law) and evaluates the effects of Chain-of-Thought (CoT) prompting on answer quality. Six essay questions from past judicial and graduate law exams were decomposed into 68 sub-questions targeting issue spotting, statutory application, legal reasoning, and property calculation. Four LLMs (ChatGPT-4o, Gemini, Copilot, and Grok3) were evaluated using a two-stage framework: decomposed sub-question accuracy and full-length essay response performance with and without Guided Question Answering with Feedback (GQA-F) at the evaluation stage, with human scoring conducted by a law professor and a student.

Results show that GQA-F consistently improves legal reasoning quality across models, notably enhancing issue coverage, statutory citation accuracy, and reasoning structure. Gemini achieved the most significant accuracy gains (from 83.2% to 94.5%, $p < 0.05$) and was selected for detailed qualitative analysis. Beyond model-specific findings, this study contributes to retrieval evaluation research by addressing statistical consistency challenges in human scoring, proposing a diagnostic evaluation method adaptable for multilingual and multimedia legal corpora, and suggesting extensions for evaluating enterprise-level legal information systems. These findings underscore the value of structured prompting strategies in supporting more interpretable, transferable, and scalable legal AI evaluation frameworks.

CCS Concepts

• **Computing methodologies** → **Natural lSanguage generation**; • **Applied computing** → **Law**; • **Human-centered computing** → **Empirical studies in HCI**.

Keywords

Large Language Models, Legal Reasoning, Chain-of-Thought Prompting, Retrieval Evaluation, Human-Centric Assessment

1 Introduction

With the rapid advancement of Natural Language Processing (NLP) technologies, Large Language Models (LLMs) have been widely applied to tasks such as legal information retrieval, document summarization, and automated contract analysis. Increasingly, they are also being expected to handle more complex forms of legal reasoning. However, compared to general NLP tasks, legal language poses unique challenges: it is highly logical, rich in domain-specific terminology, and embedded within intricate statutory structures. Real-world legal problems often involve the interaction of multiple statutes and require the ability to subsume facts under legal norms—a capability that makes legal reasoning one of the most demanding frontiers for LLMs.

In recent years, with the rise of models such as ChatGPT and Gemini, a growing body of research has attempted to apply LLMs to legal exam questions and case-based reasoning. Yet, existing studies have predominantly focused on multiple-choice or true/false formats, which only test the model's retrieval of legal knowledge. These formats fall short in assessing the model's ability to perform structured reasoning required in open-ended legal analysis. Even when LLMs are fine-tuned on domain-specific legal corpora, or when prompt engineering is applied, or even when using commercially scaled models trained on massive datasets, a core limitation persists: the model often fails to accurately identify legal issues, or while identifying relevant facts, cannot properly subsume them under applicable legal provisions. This limitation highlights a key distinction between legal reasoning and general NLP tasks: the challenge lies not in language generation per se, but in processing a multi-step logic chain that moves sequentially from facts, to rules, to conclusions.

To guide LLMs toward legal reasoning that more closely resembles human thought, this study introduces the Chain-of-Thought (CoT) prompting strategy. By providing structured prompts, we aim to help models incrementally identify the relevant parties, clarify relationships, apply legal provisions, and synthesize conclusions when addressing essay-style legal questions—thereby improving both answer quality and legal subsumption.

To investigate the comprehension and reasoning abilities of LLMs in the context of Status Law, we design a diagnostic two-stage evaluation framework. The first stage adopts a "fine-grained decomposition" approach: six essay questions adapted from past judicial and graduate law exams are systematically split into 68 sub-questions based on legal analysis logic, covering fact recognition, statutory application, issue classification, and property calculation. The second stage returns to holistic legal writing, comparing model performance on full-length answers under two conditions: with and without Guided Question Answering with Feedback (GQA-F).

We evaluate four mainstream LLMs—ChatGPT-4o, Gemini, Copilot, and Grok3—using two assessment methods. The first is absolute scoring based on standard answers, applied to Stage 1 sub-questions. The second involves human scoring of full-length responses in Evaluation Stage, with ratings provided by both a law professor and a

*Corresponding author. Email: hlshao@tmu.edu.tw

student. This allows us to analyze scoring consistency, using both score differentials and cosine similarity. One of the questions is selected for in-depth qualitative analysis to compare the original and CoT-enhanced answers in terms of issue coverage, statutory citation, and reasoning structure.

The contributions of this study span three dimensions: methodology, empirical findings, and resource creation. Methodologically, we propose a diagnostic legal reasoning test framework for LLMs that combines decomposition and staged evaluation, enabling the observation of strengths, weaknesses, and error patterns at each step of the reasoning process. Empirically, our quantitative and qualitative analyses demonstrate that GQA-F improves performance across most models, particularly in terms of issue identification and statutory logic application. In terms of resources, we develop a set of legally grounded test items adapted from real legal exams, along with their decomposed sub-question sets, which can be used in future research, legal education, or model fine-tuning.

By focusing on the deep reasoning demands of open-ended legal questions, this study addresses the current overreliance on multiple-choice or static evaluations in LLM assessment. It proposes a prompt intervention strategy centered on legal logic chains to guide models toward a more human-like legal reasoning process. Beyond improving answer quality, this experimental and analytical framework also contributes to retrieval evaluation research by enhancing statistical consistency in human scoring and offering a stepwise diagnostic method. Recognizing the multilingual and domain-specific nature of legal practice, this approach lays the groundwork for extending legal reasoning evaluation to multilingual, multimedia, and enterprise-level information retrieval systems, where reasoning transparency and retrieval accuracy are critical.

2 Related Work

With the advancement of Large Language Models (LLMs), their applications in the legal domain have grown rapidly, yielding promising results in tasks such as contract analysis, judgment summarization, legal consultation, and case prediction [1]. To promote research in legal language processing, several benchmark datasets and evaluation platforms for LLMs have emerged in recent years, including LexGLUE [2], LegalBench [10], and the COLIEE competition on statutory entailment and retrieval [7]. These benchmarks primarily cover tasks such as multiple-choice questions, case classification, statute matching, and legal question answering. However, most of them focus on English-language corpora and closed-form problems, lacking the design needed to evaluate the type of open-ended, reasoning-intensive essay questions encountered in real-world legal practice. As noted by the creators of LegalBench, current benchmarks "still fall short of comprehensively evaluating the open-ended reasoning required in law school exams and legal writing assignments" [10], which often involve deep statutory subsumption and integrated legal analysis.

Chain-of-Thought (CoT) prompting has recently emerged as a promising strategy for improving multi-step reasoning and computation. Initially proposed by Wei et al. and validated on tasks such as math word problems and commonsense reasoning [8], CoT prompting has since been shown by Kojima et al. to be effective even without in-context examples: simple natural language cues like "Let's think step by step" can activate internal reasoning chains and significantly improve performance in zero-shot settings [5]. CoT has since been widely applied in mathematical reasoning (e.g., GSM8K, MATH), logic puzzles, scientific domains, and programming tasks. Prior studies consistently find CoT especially useful for tasks that require intermediate inference steps, as it helps maintain contextual coherence and supports longer, structured chains of reasoning.

Although legal reasoning itself is inherently a multi-step logical task, systematic analysis of CoT prompting in the legal domain remains limited. Some notable attempts include the KIS team's Interpretable CoT strategy in the COLIEE 2024 entailment task, which enhances interpretability in statutory subsumption through structured prompting [4]. Another example is the LegalGPT framework, which integrates CoT modules within a multi-agent architecture to simulate the collaborative logic of real-world legal practice [6]. While such work demonstrates the utility of CoT for legal QA and statute retrieval, they fall short of addressing open-ended legal essay questions that demand issue identification, statutory chaining, and stepwise legal reasoning. To address this gap, the present study proposes a systematic evaluation framework for observing legal reasoning chains and subsumption errors in LLMs, in order to investigate how prompting strategies affect the quality of essay-style legal answers.

Mainstream approaches to evaluating legal LLMs typically rely on automated metrics (e.g., accuracy, BLEU, F1-score) or multiple-choice style datasets to compare model performance. However, such closed-form evaluations fail to reflect the logical depth and reasoning quality required for open-ended generative tasks. Recent studies have begun to incorporate human evaluation to better assess consistency and subsumption performance in long-form legal QA. Representative systems such as Length-Controlled AlpacaEval [3], MT-Bench [11], and PromptBench [9] use human preference ratings or expert judgments as quality signals, augmented by ranking-based metrics, weighted averages, or Elo-style comparisons.

Nonetheless, these benchmarks are primarily based on English tasks, and their evaluation dimensions often center on readability and preference, without offering structured analysis of legal reasoning components such as issue coverage, statute application, or subsumption logic. In response, this study designs a human evaluation framework that incorporates both quantitative and diagnostic features. By combining fine-grained decomposition with human scoring and consistency analysis, we address statistical robustness issues that are often overlooked in retrieval evaluation. Two human raters—a professor and a student—scored model outputs using a 0–10 scale, with inter-rater consistency assessed through score differentials and cosine similarity measures. Furthermore, by selecting cases with high agreement for qualitative analysis, this framework enhances process transparency and error traceability in legal reasoning evaluation. Recognizing that legal information retrieval increasingly spans multilingual, multimedia, and enterprise-specific domains, our evaluation design offers a scalable approach for diagnosing reasoning quality and retrieval reliability across diverse legal systems and information environments.

3 Experiment Design

3.1 Stage 1: Decomposed Reasoning Evaluation

3.1.1 Test Set Design. The test set used in this study consists of six essay questions, all adapted from previous Judicial Officer Examinations and the National Taiwan University Graduate Law School entrance exams in the field of Status Law. These questions cover a range of key topics, including the validity of marriage, division of marital property, legal guardianship, inheritance, bigamy, and adoption.

Unlike the holistic writing and grading approach commonly used in legal education and exams, this study adopts a decomposition strategy, breaking down each full-length question into multiple sub-questions. In total, 68 sub-questions were constructed, targeting different aspects of legal reasoning: factual judgment, statutory application, inference steps, determination of inheritance order, and property distribution calculations.

This approach draws inspiration from how law students often prepare for exams—using issue-spotting books, guided-answer manuals, and cram school handouts. The purpose here is not to simulate an actual exam setting, but rather to implement gradual stimulation and staged evaluation, allowing for more precise observation of how Large Language Models (LLMs) perform at each step of the legal reasoning process, and where they are prone to errors.

This design offers two major advantages:

- **Enhanced learning focus and specificity:** Decomposed sub-questions provide clearer response units for the model, improving its precision at each step and allowing researchers to assess how well it understands legal structures and the functions of legal provisions.
- **Clearer error localization:** By splitting each question, we can pinpoint whether a model's error occurs in fact recognition, statute application, logical subsumption, or numerical calculation—thus providing greater diagnostic value.

While this design differs from traditional legal assessments, its research value lies in significantly improving the resolution of model evaluation and providing a more effective means of exploring the capabilities and limitations of LLMs in legal reasoning.

The decision to focus on Status Law was based on its unique structural properties. Unlike other civil law subjects such as property or contract law, questions in Status Law (which encompasses family law and inheritance law) often require a combination of factual analysis, statutory application, and mathematical computation—for example, calculating inheritance shares or the balance of marital assets. This combination of cross-structural and multi-task demands makes Status Law a particularly well-suited domain for testing the integrated legal reasoning abilities of LLMs.

A complete list of questions and their decomposition logic is provided in Appendix A.

3.1.2 Rationale for Selecting Status Law as the Test Domain. This study adopts Status Law—which includes the family and succession sections of Taiwan's Civil Code—as the testing domain for evaluating LLMs' capabilities in legal understanding and reasoning. Compared to other areas of civil law, such as property law or contract law, Status Law presents several key characteristics that make it particularly suitable for the purposes of this study:

First, many questions in Status Law, such as those involving inheritance and the division of residual marital assets, require not only an understanding of legal provisions but also numerical calculations and proportional reasoning to arrive at specific answers. Solving these problems demands a combination of statutory interpretation and logical mathematical computation based on factual premises. This allows us to evaluate the model's integrated capabilities across semantic comprehension, legal application, and quantitative reasoning.

Second, Status Law problems are especially well-suited for systematic decomposition into sub-questions, which is crucial for the test design of this study. While traditional legal questions often involve holistic essay responses, our question design mimics the structure used in issue-spotting books and preparatory law school materials by breaking each scenario into individually assessable sub-items. These include binary questions, computational problems, and short-answer legal analyses. This decomposition allows us to apply "stepwise prompting" to the model, making it easier to observe performance and weaknesses at each individual reasoning stage.

By contrast, subjects such as property law and contract law, although theoretically robust, often focus on abstract assessments of rights and obligations. These questions typically lack the computational elements that enable precise decomposition and diagnostic testing. Therefore, Status Law provides not only the logical rigor of doctrinal legal analysis but also the operational structure of quantitative problems, making it an ideal subject for evaluating LLMs' legal reasoning capabilities.

3.1.3 Overview of the Question Set. The six essay questions selected for this study cover the following legal topics:

- Validity of marriage
- Division of residual marital property
- Limitations on parental rights in relation to children's interests
- Limited succession and creditor claims
- Legal consequences of bigamy
- Collation issues in inheritance distribution

The sub-questions are primarily framed as numerical problems and binary (yes/no) questions, with a few short-answer questions. Each sub-question presents a specific legal scenario and requires the LLM to provide a definitive judgment or calculation. To ensure stricter evaluation, this study adopts a rigorous scoring standard: if the model arrives at the correct final answer but misidentifies roles, relationships, or inheritance rankings within its reasoning, the response is marked incorrect. This prevents models from "guessing correctly" and emphasizes the need for accurate legal reasoning and comprehension.

3.1.4 Model Selection. The study evaluates four mainstream LLMs:

- ChatGPT-4o (OpenAI)
- Grok 3 (xAI)
- Gemini (Google)
- Copilot (Microsoft)

All models were tested using identical formats and prompts. Their responses to the 68 sub-questions were collected, manually graded, and subjected to statistical analysis.

3.2 Stage Two: Evaluation of GQA-F vs. Baseline)

3.2.1 Response Format and Prompting Design. This evaluation stage compares two prompting strategies: baseline answering and Guided Question Answering with Feedback (GQA-F):

(1) **Baseline answering**: The six legal questions in Status Law are provided in their original, unsegmented form, with no added prompting or step-by-step guidance. Each model is instructed to generate a single, complete answer in one shot, simulating a typical law exam environment.
(2) **Guided Question Answering with Feedback (GQA-F)**: The same legal questions are used but decomposed into sub-questions as designed in Stage One. The model is prompted to answer each sub-question sequentially. After each response, a human evaluator provides immediate confirmation or correction before proceeding to the next sub-question. Upon completion of all sub-questions with verified answers, the model is then prompted to generate a full-length answer using the confirmed intermediate reasoning as scaffolding.

This clear distinction emphasizes that while baseline answering reflects zero-shot full answer generation, GQA-F introduces interactive, stepwise feedback to scaffold legal reasoning. This differs from the commonly referenced Chain-of-Thought prompting, which is evaluated in the full-length essay setting without intermediate intervention. The model receives the entire legal question and generates

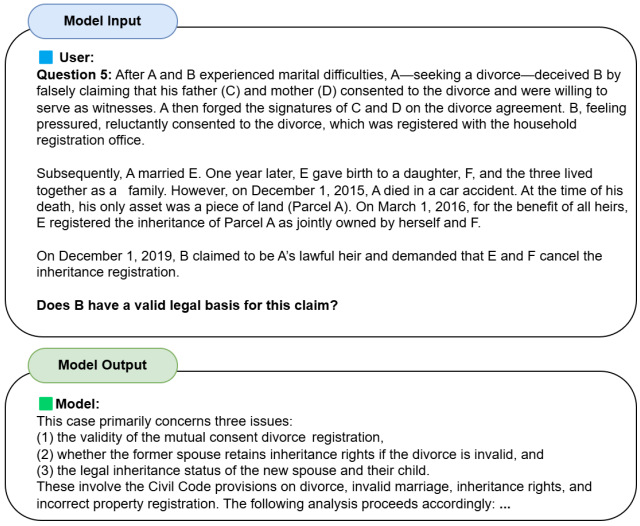

Figure 1: Example of Baseline Answering

a complete answer without stepwise guidance or feedback.

To visually illustrate the experimental design, Figures 1 and 2 demonstrate the interaction patterns between the user and the model under the two response conditions. The example used in both figures is derived from Question 5 of the test set, which involves issues of divorce validity, remarriage, and inheritance registration under Taiwan's Civil Code.

Figure 1 presents the baseline setting, where the model receives the entire legal question in its original, unstructured format and generates a one-shot, complete response without any intermediate guidance or feedback. This simulates the typical exam-taking environment faced by law students.

In contrast, Figure 2 depicts the Chain-of-Thought (CoT) prompting setting. The model is prompted to answer each sub-question in a step-by-step manner. After each step, the user provides immediate feedback by confirming or correcting the model's answer. This scaffolding prevents early-stage reasoning errors from propagating and encourages structured legal analysis. The final step prompts the model to synthesize a full answer to the original question, based on the previously verified steps.

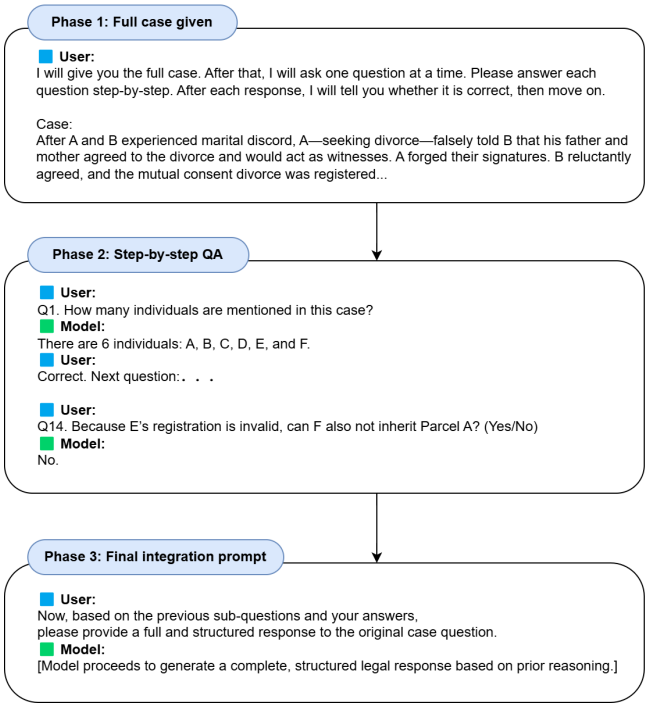

Figure 2: Example of Guided Question Answering with Feedback (GQA-F)

The model answers decomposed sub-questions one at a time, receiving immediate evaluator feedback before synthesizing a full-length response.

This contrast illustrates how GQA-F transforms the model's reasoning process from a monolithic, opaque response into an interactive, modular sequence of logic steps, allowing for clearer observation and evaluation.

3.2.2 Scoring Mechanism. Both response versions were evaluated by scorers with formal legal training, who rated each answer holistically on a 0–10 scale, with higher scores reflecting better overall quality. Scoring was based on three key criteria:

(1) Issue coverage: Whether the model identified and addressed the key legal issues and factual disputes in the question.
(2) Accuracy of statutory application: Whether the cited or applied legal provisions were correct and logically relevant.
(3) Clarity of legal reasoning: Whether the reasoning was coherent, structured, and logically sound.

Each response was independently scored by four raters:

(1) Rater A: A university professor with expertise in Status Law, specializing in family and inheritance law, with extensive experience in legal instruction and examination.
(2) Rater B: A second-year undergraduate law student, representing the perspective of a typical learner and assessing the accessibility of the model's reasoning.

All scorers evaluated the answers independently and blindly—they were not informed which model had generated the response or whether it had received GQA-F. The final score for each response was calculated as the average of the two raters' scores. This design ensures a balance between expert judgment and learner perception, aiming to enhance scoring reliability and capture potential differences in answer acceptability across rater types.

Although the number of raters was limited, the study mitigated subjectivity by introducing additional consistency metrics, including score discrepancy and cosine similarity between raters. The responses with the highest inter-rater agreement were further selected for qualitative analysis to support the robustness of the findings.

4 Experiment Result

4.1 Performance under Binary Evaluation

4.2 Performance under Binary Evaluation

This evaluation stage compares Large Language Model (LLM) performance under two prompting strategies using binary scoring: **Baseline Answering** and **Guided Question Answering with Feedback (GQA-F)**. In the baseline setting, models were asked to generate full responses to unsegmented legal essay questions. In the GQA-F strategy, models answered decomposed sub-questions one-by-one, receiving immediate human feedback before synthesizing the final answer.

As shown in Fig. 3, all models performed better under GQA-F than under baseline prompting. For instance, GPT-4's accuracy increased from 0.826 (baseline) to 0.870 (GQA-F), a gain of +4.4 percentage points. Similar trends were observed for Claude 2 and GPT-3.5. These improvements suggest that interactive decomposition and feedback help LLMs better capture legal reasoning steps.

To evaluate the performance difference between conditions, we used a non-parametric sign test over the 69 binary sub-question pairs. The sign test is appropriate given the binary nature of the responses, which violates the interval scale assumption required by a paired-sample t-test. GPT-4 and Claude 2 exhibited statistically significant improvements under the GQA-F condition ($p < 0.05$), supporting the advantage of stepwise guidance.

Based on the results of the paired-sample t-tests, all four models demonstrated higher average accuracy when GQA-F was applied. Notably, Gemini and Grok3 showed statistically significant improvements ($p < 0.05$), suggesting that structured reasoning prompts had

Model	*Raw Accuracy*	*CoT Accuracy*	*t-value*	*p-value*
ChatGPT	0.842	0.866	-0.92	0.398
Gemini	0.833	0.9445	-3.71	0.013
Copilot	0.822	0.864	-2.14	0.089
Grok3	0.843	0.895	-2.98	0.031

Figure 3: Performance Comparison of LLMs in Decomposed Reasoning Tasks With and Without Chain-of-Thought Prompting

a substantial positive effect on their ability to address fine-grained legal issues, apply statutory provisions, and make logical inferences.

Although ChatGPT did not show a statistically significant difference, it already had the highest baseline accuracy among all models (84.2%). While its improvement margin under GQA-F was modest at 2.43%, it nonetheless exhibited consistent and stable gains. In contrast, Copilot showed an accuracy increase of approximately 4.2%, but the improvement fell short of significance ($p = 0.089$), indicating a marginal effect. This may suggest lower sensitivity to prompting or a baseline answering strategy that partially overlaps with CoT reasoning.

Overall, even though the unit of observation in this stage was the model's performance on decomposed sub-questions, the results clearly demonstrate that GQA-F significantly enhanced accuracy for certain models. The effect was particularly pronounced in question types that involved multi-step reasoning and structured analysis, such as inheritance calculation, classification of legal status relationships, and precise statutory mapping. These findings provide a quantitative foundation for the holistic answer evaluations conducted in Evaluation Stage (GQA-F vs. Baseline).

4.3 Holistic Evaluation Results

In addition to binary scoring, we conducted holistic human evaluation on the full-length answers produced under each strategy. Two raters—a legal expert and a senior law student—independently rated each answer on a 1–10 scale using three criteria: (1) issue coverage, (2) statutory application, and (3) reasoning clarity. Each criterion was scored individually, and the holistic score for each answer was computed as the average of these three dimensions. Annotators were instructed to consider all three aspects equally when assigning the overall score.

As shown in Fig .4, LLMs generally received higher holistic ratings under the **GQA-F** condition compared to the **Baseline Answering** condition. For example, GPT-4 improved from a mean score of 7.7 to 8.2, while Claude 2 improved from 7.5 to 7.9. The scoring differences reflect gains in legal interpretability and structured reasoning resulting from stepwise guidance and verified intermediate outputs.

We assessed inter-rater agreement using Spearman's rank correlation, which is more appropriate than Pearson for ordinal rating scales such as the 1–10 scores used in our evaluation. The resulting coefficients (ranging from 0.72 to 0.85) suggest moderately strong agreement between expert and student raters.

It is important to note that these scores are aggregated over only six legal case questions. Hence, minor variations in mean values

should be interpreted cautiously. Still, the consistent advantage of GQA-F across systems indicates its potential for improving coherence and legal plausibility in essay-style legal responses.

The goal of the Evaluation experiment was to simulate realistic legal exam conditions and assess whether the overall quality of LLM-generated responses to unsegmented, full-length legal questions could be improved by introducing Chain-of-Thought (CoT) prompting. The same six Status Law questions used in Stage One were employed, but this time they were presented in their entirety—without decomposition—requiring the model to generate a complete answer in one go.

Two response conditions were tested:

(1) Baseline answering: No prompting was provided; the model responded to the full question as-is, simulating natural generation without guidance.

Model	*Raw Average Score*	*CoT Average Score*	*Average Improvement*	*Improvement by Student Rater*	*Improvement by Professor Rater*	*Scoring Consistency (Pearson's r)*
ChatGPT	6.50	9.17	**+2.67**	+3.00	+2.33	0.716
Gemini	6.12	8.04	+1.92	+2.83	+1.00	0.853
Copilot	5.83	7.42	+1.58	+2.00	+1.17	0.752
Grok3	6.25	8.08	+1.83	+2.17	+1.50	0.835

Figure 4: Human Evaluation of Full-Length Answers: Comparison Between Baseline and CoT Conditions (Average Scores and Inter-Rater Consistency)

(2) CoT-prompted answering: Systematic prompts were added to guide the model through identifying legal relationships, applying statutes, performing legal calculations, and synthesizing conclusions.

All responses were evaluated by two human raters with legal backgrounds—a university professor specializing in Status Law and a second-year undergraduate law student. Each answer was scored holistically (0–10 scale) based on three criteria: issue coverage, accuracy of statutory application, and clarity of legal reasoning and explanation.

4.3.1 Overall Model Scoring Results. The results show that all models demonstrated improved performance when GQA-F was applied. Among them, ChatGPT exhibited the largest improvement (+2.67 points) and the most consistent performance across questions. Gemini and Grok3 also showed marked improvements, each with gains exceeding 1.8 points. Although Copilot lagged behind the other models in terms of raw scores, it too displayed consistent improvement under GQA-F.

In terms of inter-rater agreement, Pearson correlation coefficients ranged from 0.71 to 0.85 across the four models, indicating a moderate to high level of scoring consistency between the two raters. Notably, Gemini and Grok3 achieved the highest consistency, suggesting particularly stable performance as evaluated by both expert and student raters.

4.3.2 Per-Question Score Differences and Selection of Representative Questions. To further examine whether the effects of GQA-F are reflected at the individual question level, this study conducted

	Question	*Raw Avg. Score*	*CoT Avg. Score*	*Score Gain*	*p-value*
1		4.50	7.75	+3.25	0.068
2		6.12	7.12	+1.00	0.430
3		6.25	6.38	+0.12	0.919
4		3.88	5.50	+1.62	0.080
5		**3.88**	**7.00**	**+3.12**	**0.002**
6		**4.63**	**7.50**	**+2.88**	**0.011**

Figure 5: Average Score Differences for Each Question: Baseline vs. CoT Answering (with Paired t-Test Results)

paired-sample t-tests comparing the average scores of baseline and CoT answers across the six questions. The test results are shown in Table 3.

As shown in Table 3, Questions 5 and 6 exhibited the greatest improvements in average scores after GQA-F, both of which reached statistical significance ($p < 0.05$). This suggests that these types of questions benefit most from logical guidance and structured prompting. Accordingly, Questions 5 and 6 were selected for subsequent qualitative analysis to demonstrate how GQA-F improves the reasoning process in practice.

4.4 Qualitative Analysis of Model Responses on a Representative Question

4.4.1 Selection and Rationale for the Representative Case. To select a representative case for qualitative analysis, we further examined inter-rater consistency for Questions 5 and 6. The results showed that Question 5 exhibited a stronger positive correlation between the two raters' scores across both baseline and CoT responses ($r = 0.720$), compared to Question 6, which showed a moderate correlation ($r = 0.619$). Since the goal of our qualitative analysis is to examine cases in which raters tended to agree in their evaluations, Question 5 was selected for in-depth analysis.

For the model selection, we chose to focus on Gemini, analyzing both its original and CoT-prompted responses. This choice was based on the following reasons: (1) both the professor and the student gave low scores to the original version, while (2) the CoT-enhanced version received substantially higher scores from both raters (6 and 7, respectively). This case thus offers sufficient contrast and interpretive space without reaching a ceiling effect. Furthermore, Gemini was among the models that showed the greatest improvement in performance across the study.

> [Representative Case]:
> Question 5 – Validity of Divorce, Remarriage, and Inheritance Registration Dispute
> "After marrying, A (husband) and B (wife) experienced marital discord. In order to obtain a divorce, A deceived B by falsely claiming that his father C and mother D agreed to the divorce and were willing to act as witnesses. A then forged the signatures of C and D on the divorce agreement. B, feeling pressured, reluctantly consented to the divorce, and the mutual consent divorce was registered with the household registration office.
> Subsequently, A married E, and one year later, E gave birth to a daughter, F. A, E, and F lived happily

together as a family. However, on December 1, 2015, A tragically died in a car accident. At the time of his death, A's only estate was a piece of land (Parcel A). On March 1, 2016, for the benefit of all heirs, E completed the inheritance registration of Parcel A, registering the property under joint ownership of herself and F.
On December 1, 2019, B asserted that she was A's lawful heir and demanded the cancellation of the inheritance registration made by E and F."
Legal Question: "Does B have valid legal grounds to claim inheritance and request the cancellation of the inheritance registration in favor of E and F?"

4.4.2 Key Analytical Dimensions and Findings. Based on the ideal reference answer, the qualitative evaluation was conducted across three dimensions: **(1) Issue Coverage, (2) Accuracy of Statutory Citation, and (3) Clarity of Legal Reasoning.** The findings for each dimension are described below:

(1) Issue Coverage

A logically correct response should address the following **four legal issues**, in order:

(1) **Validity of A and B's divorce:** The divorce lacked the actual signatures of witnesses, violating Civil Code Article 1050. Thus, the divorce is invalid, and B remains A's legal spouse and retains inheritance rights.
(2) **Whether E qualifies as a lawful spouse:** Since the marriage between A and B remains valid, A's subsequent marriage to E constitutes bigamy. Moreover, E was not in good faith, making the marriage invalid under Articles 985 and 988. Therefore, E is not a lawful heir.
(3) **Whether F is a legal heir:** F is a non-marital child, but was legally recognized through presumed acknowledgment. Under Articles 1065-1 and 1069, F is entitled to inheritance.
(4) **Whether the inheritance registration may be revoked:** Since E is not a legitimate heir, her registration should be revocable under Article 767. However, F's registration remains valid due to her legitimate inheritance status and may not be revoked.

(2) Accuracy of Statutory Citation

Figure 6 provides a visual comparison of key statutes that should be cited in an ideal answer.

The correct response should reference the following statutes:

- **Article 1050:** Formal requirements for divorce
- **Article 1138:** Legal order of succession
- **Articles 985 & 988:** Bigamy and invalidity of marriage
- **Articles 1065-1 & 1069:** Inheritance rights of non-marital children
- **Article 767:** Conditions for revoking land registration

The original version erroneously cited Article 92, which pertains to the revocation of declarations of intent due to fraud. This reflects a misunderstanding of the legal nature of the problem—it misclassified the issue as a defect in intent rather than a formal defect that invalidates the divorce.

Furthermore, it failed to mention several key statutes, including those governing bigamy and non-marital inheritance.

Provision	Original Answer	CoT Answer	Remarks
Civil Code §1050 (Mutual Consent Divorce)	✓	✓	Both correct
Civil Code §92(Intent Defect – Fraud/Duress)	✓	✓	Correct, but CoT explanation lacks depth
Civil Code §1052(Judicial Divorce)	✓	✗	Incorrect reference in original version
Civil Code §767(Cancellation of Land Registration)	✗	✗	Both missing: major omission
Civil Code §985 (Invalidity of Bigamous Marriage)	✗	✗	Not mentioned in either
Civil Code §1138(Order of Successsion)	✗	✓	Only CoT version cited
Civil Code §1148(Inheritance by Spouse)	✗	✗	Not cited in either
Civil Code §1065-1(Persumption of Birth)	✗	✗	Concept mentioned, but article not cited

Figure 6: Statutory References Cited in Original and CoT Responses

The CoT version correctly cited Articles 1050 and 1138, capturing part of the statutory logic. However, it omitted Article 767 and did not clearly reference the provisions governing non-marital children, resulting in a fragmented presentation of the statutory framework.

(3) Subsumption and Reasoning Logic

In the original response, the model developed its reasoning under the assumption that the divorce between A and B was valid. While the internal logic of the answer was coherent, this false premise led the reasoning astray, ultimately producing an incorrect conclusion. This represents a classic example of misapplication of legal rules due to a flawed premise.

By contrast, the CoT response demonstrated significantly improved structural organization and logical coherence. The model was able to unfold its reasoning step-by-step and partially identify the relevant legal conditions. However, the precision of factual subsumption and legal application remained insufficient. For example, it overgeneralized the inheritance rights of non-marital children by stating that "as long as a parent-child relationship exists, inheritance is allowed," without explicitly addressing the legal procedure of presumed acknowledgment and the necessary conditions for establishing kinship.

(4) Summary

The qualitative analysis of this case indicates that GQA-F can effectively improve the logical structure and issue coverage in model-generated responses. However, there remains considerable room for improvement in legal precision and the accurate application of statutory provisions. In contrast, while the baseline responses occasionally arrived at correct conclusions, they were often built upon flawed premises and lacked overall coherence in legal reasoning and subsumption. This underscores the fact that, although large language models may exhibit surface-level fluency, they still require structured guidance—such as GQA-F—to reconstruct the proper sequence and scope of legal reasoning.

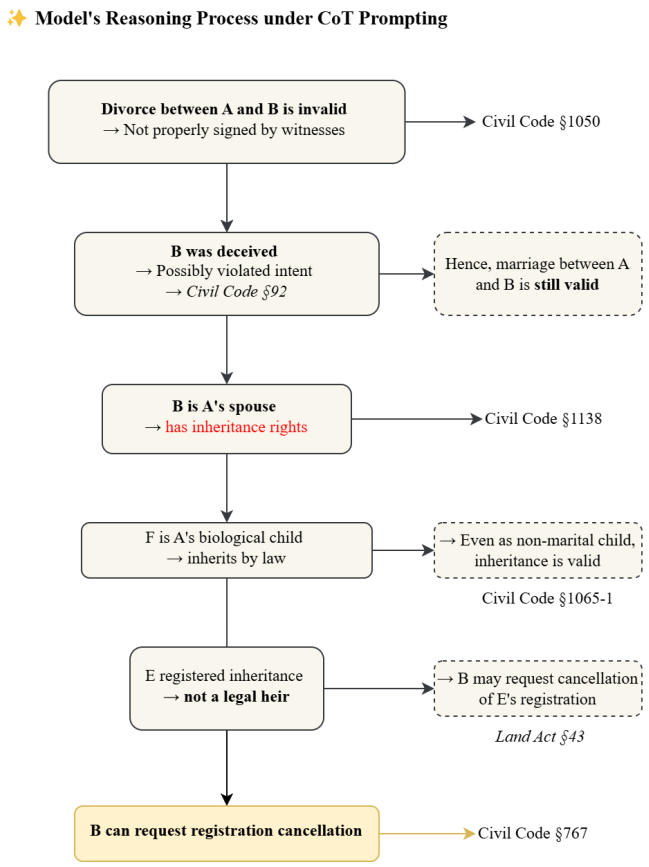

Figure 7: Model's Reasoning Process under GQA-F

5 Discussion

5.1 Contribution

This study illustrates both the promise and complexity of evaluating legal reasoning in LLMs. Our results show that structured prompting (GQA-F) and decomposed feedback pipelines can enhance legal performance, particularly for models such as GPT-4 and Claude 2. These findings suggest that legal evaluation frameworks must go beyond raw accuracy and account for the procedural structure of legal tasks. Interactive strategies like GQA-F may offer a better approximation of real-world legal reasoning workflows than single-shot prompts. Future research should also explore multilingual legal corpora and integrate model latency and inference cost into evaluation metrics for practical deployment.

5.2 Study Limitations

Although this study adopted a two-stage evaluation design with decomposed questions to closely examine the performance of large language models (LLMs) on legal essay tasks, and while its framework and findings are of exploratory value, several limitations remain, as outlined below:

First, in terms of human evaluation, the scoring of full-length answers in the second stage was carried out by only two raters, one being a professor with expertise in family and inheritance law and the other an undergraduate law student. This limited number of evaluators may introduce subjective bias into the scoring process. To mitigate this, the study also analyzed scoring variance and cosine similarity to assess inter-rater consistency and selected high-agreement cases for qualitative analysis to strengthen reliability and interpretability. Future research may incorporate more evaluators with diverse legal backgrounds and adopt double-blind rating or preference-ranking methods (e.g., Elo scoring) to further enhance evaluation robustness and mitigate subjectivity. Although the current data set covers the status law under the Civil Code of Taiwan, expanding to additional legal domains and jurisdictions, including multilingual and multimedia legal corpora, is essential to assess the generalizability of findings in different legal systems and information environments. In addition, adapting the evaluation framework to enterprise retrieval systems, where heterogeneous document formats, incomplete information, and domain-specific variance are common, represents an important direction for future research. Exploring variations in prompt design, prompt length, and prompting strategies (e.g., few-shot prompting) will also help to identify model sensitivities and further optimize structured prompting for complex legal reasoning tasks.

5.3 Statistical Limitations

We acknowledge statistical limitations in the interpretation of mean scores over a small number of test cases (n = 6). Furthermore, we applied non-parametric tests (sign test, Spearman correlation) to accommodate binary and ordinal data structures, avoiding assumptions of interval scale distributions. Future work may adopt bootstrapped confidence intervals or Bayesian methods to further validate model comparisons.

5.4 Dataset and Prompt Availability

To support replicability and community benchmarking, we are publicly releasing the anonymized legal test set (six full legal essay questions) along with the structured prompts used in both baseline and GQA-F settings. The data set will be made available on an open source platform. A dedicated repository is currently under construction and will be linked in the final version.

We emphasize that all test cases were authored by the Taiwanese Bar Exam authority and were adapted for academic use under fair use and educational clauses. All examples shown in English in this paper were translated from the original Chinese.

6 Conclusion

This study evaluated the legal understanding and reasoning capabilities of large language models (LLMs) in Taiwan's status law domain, examining how structured prompting strategies, specifically Chain-of-Thought (CoT) prompting, affect answer quality. Through a two-stage experimental design, we assessed model performance on decomposed sub-questions and full-length essay responses (Evaluation Stage) under baseline and CoT conditions, with evaluation by both legal experts and students.

Results showed that GQA-F consistently improved reasoning quality across models, particularly in issue identification, statutory application, and logical structuring. Gemini and Grok3 achieved

statistically significant accuracy gains, while qualitative analyzes highlighted CoT's role in strengthening reasoning chains and statutory interpretation, although challenges remain in the precision of legal subsumptions.

Beyond empirical findings, this study contributes to broader retrieval evaluation research by addressing statistical robustness through fine-grained human scoring and diagnostic decomposition methods. It also lays the groundwork for future extensions into multilingual, multimedia, and enterprise retrieval systems, where structured legal reasoning evaluation is crucial for improving retrieval reliability, system transparency, and user trust. The stepwise-structured legal test framework developed here offers a scalable resource for legal AI training, evaluation, and education across diverse legal domains and jurisdictions.

Our evaluation contributes to the EVIA agenda in three ways. First, we propose structured evaluation metrics beyond raw accuracy, including human judgment and rating agreement, addressing open concerns in legal retrieval evaluation. Second, while our dataset is in Chinese, our framework generalizes to multilingual legal settings and encourages further investigation of language-specific model behaviors. Third, the procedural design of our evaluation, which simulates legal problem solving in guided substeps, has implications for enterprise-grade legal information systems, where explainability, review cycles, and incremental QA are crucial.

Together, our findings suggest that the integration of retrieval robustness, procedural reasoning, and human evaluation can form a more comprehensive and practice-aligned assessment strategy for legal LLMs.

References

[1] [n. d.]. COLIEE Official Site. https://sites.ualberta.ca/~rabelo/COLIEE.html. Accessed April 28, 2025.

[2] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559* (2020).

[3] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. LexGLUE: A benchmark dataset for legal language understanding in English. *arXiv preprint arXiv:2110.00976* (2021).

[4] Yuan Chen, Ronglai Shen, Xiwen Feng, and Katherine Panageas. 2024. Unlocking the power of multi-institutional data: Integrating and harmonizing genomic data across institutions. *Biometrics* 80, 4 (2024), ujae146.

[5] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475* (2024).

[6] Masaki Fujita, Takaaki Onaga, and Yoshinobu Kano. 2024. LLM Tuning and Interpretable CoT: KIS Team in COLIEE 2024. In *Proceedings of the JSAI International Symposium on Artificial Intelligence*. Springer, 140–155.

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

[8] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems* 35 (2022), 22199–22213.

[9] Juanming Shi, Qinglang Guo, Yong Liao, and Shenglin Liang. 2024. LegalGPT: Legal Chain of Thought for the Legal Large Language Model Multi-agent Framework. In *Proceedings of the International Conference on Intelligent Computing*. Springer, 25–37.

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

Received 2 May 2025; revised 3 June 2025