Taxation Perspectives from Large Language Models: A Case Study on Additional Tax Penalties

Anonymous ACL submission

Abstract

How capable are large language models (LLMs) in the domain of taxation? Although numerous studies have explored the legal domain in general, research dedicated to taxation remain scarce. Moreover, the datasets used in these studies are either simplified, failing to reflect the real-world complexities, or unavailable as open source. To address this gap, we introduce PLAT, a new benchmark designed to assess the ability of LLMs to predict the legitimacy of additional tax penalties. PLAT comprises a total of 300 examples, (1) 100 binary-choice questions, (2) 100 multiple-choice questions, and (3) 100 essay type questions, all originally derived from 100 Korean precedents. PLAT is constructed to evaluate not only LLMs' understanding of tax law, but also their performance in legal cases that require complex reasoning beyond straightforward application of statutes. Our systematic experiments with multiple LLMs reveal that (1) their baseline capabilities are limited, especially in casees involving conflicting issues that requires a comprehensive understanding, and (2) LLMs struggles particularly with the "AC" stages of "IRAC" even for advanced reasoning models like o3, which actively employ inference-time scaling.

1 Introduction

Large Language Models (LLMs) have demonstrated promising results across various domains. Among them, the legal domain has been one of the earliest areas of application, since OpenAI's demonstration that GPT-4 passes the U.S. Uniform Bar Exam (Martinez, 2023). To solidly assess LLMs' capabilities in the legal domain beyond the bar exam, where questions may follow certain patterns, many studies have proposed benchmarks (Guha et al., 2023; Fei et al., 2024; Kim et al., 2024) and analyzed LLM performance Magesh et al. (2024); Kang et al. (2023); Trautmann et al. (2024); Chalkidis (2023).

However, in the taxation domain–despite its close relationship with the legal field, there has been little research on assessing LLM capabilities. Previous studies have primarily focused on relatively simple questions that can be answered mostly based on deductive application of statutes (Holzenberger et al., 2020; Nay et al., 2024), or have used real-world datasets without releasing them as open source, making reproduction difficult (Harvey Team, 2024; Zhong et al., 2024). With rapid progress of LLMs and advancements in LLM-based agents (or test-time scaling) (OpenAI, 2024; Guo et al., 2025), issues such as deductive reasoning (Lee and Hwang, 2025) or simple calculation errors can now be easily mitigated using external tools. This suggests that more advanced benchmarks may be necessary for comprehensive evaluation in the taxation domain.

Here, we introduce PLAT[1], a benchmark consisting of 300 questions derived from Korean precedents concerning the legitimacy of additional tax penalties. Article 48 of Korean Framework Act on National Taxes[2] allows exemptions from penalty taxes in cases of *justifiable reasons*, but the statute does not explicitly define what constitutes such reasons. Thus, we use PLAT to assess LLMs' tax law comprehension, particularly in scenarios where the issue cannot be resolved by merely referencing statutes.

PLATis designed to assess not only LLMs' domain knowledge in taxation but also their legal reasoning capabilities in complex cases—especially where resolution requires real-world considerations, such as weighing competing legal principles or judging whether it is reasonable to expect a taxpayer to recognize and comply with the law.

Our experiments with two open-source LLMs–

[1] PREDICTING THE LEGITIMACY OF PUNITIVE ADDITIONAL TAX

[2] https://elaw.klri.re.kr/kor_service/lawTwoView.do?hseq=28738

Qwen3 (Yang et al., 2025), Exaone (Research et al., 2024))–alongside five commercial LLMs (GPT-o3, o3-mini, 4o, 4.1, and Claude 3.7) show that the strongest reasoning model, GPT-o3, achieves an F_1 score of 0.79 on PLAT. A detailed analysis reveals, while LLMs perform well on relatively simple problem, their accuracy declines when a comprehensive understanding is required. For instance, all LLMs correctly recognize that ignorance or misunderstanding by taxpayers cannot serve as a justified reason. However, when the misunderstanding originates from an incorrect statement of opinion by the tax authority, accuracy drops due to a conflict between two legal principles: (1) the final responsibility lies with the taxpayer, vs (2) the principle of protection of legitimate expectations.

To address this issue, we adopt the IRAC framework and investigate how LLM performance varies under the following conditions: (1) enabling retrieval-augmented generation (RAG), (2) providing the "Application" and "Conclusion" stages, and (3) introducing more complex essay-type questions.

Our findings reveal that (1) LLMs remain relatively proficient at identifying the "Issue"; (2) consistent with prior work, they struggle to identify the correct "Rule" due to hallucinations (Dahl et al., 2024), though this can be mitigated with RAG; (3) LLMs underperform in the "Application" and "Conclusion" stages: without inference-time scaling, they often hesitate to proceed, resulting in low recall, while with inference-time scaling, they do continue but frequently reach incorrect conclusions; (4) when the "Answer" (Conclusion) and a corresponding simplified "Reason" (Application) are provided as a starting point, LLM accuracy improves significantly, highlighting the potential of backward-chaining reasoning in legal contexts; (5) regardless of inference-time techniques or task format, final "Conclusion" accuracy remains limited, even when performance on intermediate steps such as "IRA" is high.

In summary, out contributions are

- We propose a new dataset, PLAT, to evaluate LLMs' understanding of tax law, particularly in legal cases that cannot be resolved solely by referencing statutes.
- We assess nine LLMs and find that, while they exhibit some competence, their performance is limited—especially in comprehending legal cases at the "Conclusion" stage even with inference-time scaling.

Our datasets–both original Korean, and English translated version–will be released to the community under a CC BY-NC license.

2 Related Work

2.1 NLP in Taxation domain

Nay et al. (2024) studies GPT-4's capability in handling tax law inquiries with and without retrieval augmented generation (RAG). Their study uses synthetically generated multiple-choice questions based on templates, where answers can be derived from either the Treasury Regulations under the U.S. Code of Federal Regulations (CFR) or Title 26 of the U.S. Code. The datasets has not been released.

Holzenberger et al. (2020) develops SARA, a statutory reasoning dataset constructed from a simplified version of U.S. Internal Revenue Code. The dataset consists of two tasks: determining entailment relations and calculating tax amounts based on given statues and cases. Since all questions can be answered mostly through deductive reasoning from the given statutes, the dataset primarily comprises relatively simple questions.

Zhong et al. (2024) develops a retrieval-based LLM system designed to answer tax-related questions typically handled by tax departments. The datasets has not been released.

Compared to these studies, our dataset consist of 50 manually constructed examples, supervised by tax professionals. PLAT is particularly distinct from previous datasets in that its questions cannot be answered solely by referencing statutes. Instead, they require a comprehensive understanding of tax law and complex reasoning about real-world situations.

2.2 Agent

LLM-based AI agents are being rapidly developed. Unlike vanilla LLMs, which simply generates output text based on input text, LLM-based agents can enhance their capabilities by leveraging external tools for knowledge retrieval (e.g., search engine), improving reasoning (e.g., logic solver (Lee and Hwang, 2025)), or refining internal knowledge through memory and self-reasoning processes. These processes can be iteratively orchestrated by the LLMs themselves. Below, we highlight a few representative works.

Yao et al. (2023a) introduces the Tree-of-Thoughts inference algorithm, which allows LLMs to generate and navigate multiple reasoning paths unlike Chain-of-Thought (Wei et al., 2022), which follows only a single path.

Yao et al. (2023b) proposes REACT, which integrates reasoning and planning (such as action generation and document retrieval). The inference process is formalized into tree key steps: thought (planning), action (tool calling), and observation (interpreting tool-generated results).

Wu et al. (2024) presents AutoGen, an open-source framework for building LLM-based agent with a focus on multi-agent interaction. Similarly, Roucher et al. (2025) introduces smolagents, another open-source framework designed for simplicity and seamless Python code integration. Both frameworks are employed in this study.

3 Datasets

3.1 Motivation

An additional penalty tax can be applied to all 25 types of taxes in Korea. It is an additional economic burden imposed on taxpayers who fail to properly file or pay their taxes, in addition to the original tax liability. However, when there are objective circumstances that prevent taxpayers from fulfilling their tax obligations, it would be more reasonable not to impose the penalty tax even when there is a legal basis for imposing a penalty tax.

Indeed, the section 2 of Article 48 of Korean Framework Act on National Taxes explicitly states that a penalty tax shall not be imposed if there is a "justifiable reason." However, this phrase is an indeterminate concept, meaning that the term used in the law is abstract and lacks a clear scope, requiring interpretation in specific cases Kim and Lee (2008); Yang (2024); Park (2019). In a situation where statutes are ambiguous, interpretative standards become necessary, and this is where precedents play a crucial role. Court rulings determine, in such cases, whether a given situation constitutes a "justifiable reason" or not[3].

Thus, it requires not just referencing the statutes but to understand the individual situation comprehensively to answer the "justifiability" like human judges. In this regard, we build PLAT that are created from 100 Korean precedents–50 justifiable, 50 not justifiable cases– handling the issue regarding the legitimacy of the additional tax penalty.

We believe that this study is not merely limited to tax law, but represents a starting point for exploring dimensions of legal judgment—such as leniency, compassion, and discretionary reasoning—that are unique to human judges. These aspects are not exceptional outliers but fundamental components of real-world legal decision-making, which current LLMs are inherently unable to replicate. Accordingly, our work serves not only as a benchmark for tax-related reasoning, but also as a broader indicator for assessing the applicability of LLMs across diverse areas of law.

3.2 Dataset Construction

We first collect relevant precedents using the commercial Korean legal search engine LBOX[4], searching with the keyword "additional penalty tax". The query returned approximately 20k precedents. To further refine the dataset, we added the keyword "justifiable reasons," reducing the target cases to 3.7k. Finally, we excluded cases containing the keyword "gift tax," as such cases primarily focus on the issue related to the method of tax calculation. This results in total 2.8k candidate pools.

To extract facts and claims from precedents, we used GPT-o3. We initially prepared 10 examples, which were manually evaluated by two tax professionals (authors of this paper) based on the following criteria:

- Well-defined task: Does the input contain sufficient information to answer the question? Are the main issues of the selected cases related to an additional penalty tax?
- Information leakage: Is there any unintended disclosure of the court decision in the input?
- Hallucination: Are their any inaccuracies of fabricated information in the extracted facts and claims?
- Legal Correctness: Are the labels extracted from court ruling consistent with the actual court decisions?

Based on this criteria, we removed unrelated cases–such as those where the focus was on the original tax liability rather than the justifiability of a penalty tax–during the first. We repeated this

[3] Although Korean legal system is rooted in civil law system, higher courts' decisions, especially those of the Supreme Court, are typically followed by lower courts.

[4] lbox.kr

process until we compiled a final 100 examples, with an equal split: 50 cases where the court ruled the exemption from penalty tax was, and 50 cases where the court decided that the exemption was not justified. Each example required approximately 30–40 minutes for evaluation, resulting in total 50–67 hours of expert review time.

Based on this, we built two multiple-choice (PLAT-MC, PLAT-MC$_R$) and one essay type (PLAT-E) QA datasets.

- PLAT-MC: Each question provides two answer choices—"lawful" and "unlawful"—along with an additional "don't know" option for cases where the model is uncertain. Because our goal was to construct a dataset that closely reflects real-world legal scenarios, we added the "Cannot be determined" label. In practice, especially in the legal domain where accuracy is critical, it is important for models to be able to express uncertainty. Therefore, we included a "Cannot be determined" option to allow models to respond honestly when they cannot confidently determine the legitimacy of a case.

- PLAT-MC$_R$: We labeled choices according to court's logic and judge's decision in precedents. Each option includes not only whether the judgment is lawful or unlawful, but also the key rationale behind the judge's decision. These 400 options were all manually labeled, evaluated and modified by two tax professionals (authors of this paper).

- PLAT-E: Each essay question follows the format of the second-round essay-style exam for the Korean Certified Tax Accountant (CTA). We considered the court's reasoning and the judge's final decision as the reference answer. To extract rubrics from the precedents, we used GPT-o3. Initially, we prepared 10 examples, manually written by a tax professional (an author of this paper), based on the IRAC framework[5].

The GUI used during the annotation is shown in the Appendix.

[5]Issue, Rule, Application, Conclusion

4 Experiments

We used two open LLMs (Qwen3-32B[6], LG EXAONE3.5-32B[7]) and five commerical LLMs (GPT4o, 4.1, o1, o3, o3-mini[8], and Claude3.7 sonnet[9]). For retrieval-based experiment, we use Pyserini (Lin et al., 2021) with the BM25 algorithm with default hyperparameters. Each retrieval is limited to three documents, which was selected during initial experiments with top-1, 3, 5 and 10. The retrieval pool comprises 100 precedents related to additional tax penalties and 4,042 articles related to Korean tax law. The articles are filtered from the Korean Statutes Corpus (Kim et al., 2024). Also, the source precedent for each question was excluded from the retrieval pool to prevent information leakage. We use `smolagent` (Roucher et al., 2025) to build LLM agents. For all experiments with non-reasoning models, we set the temperature to 0.0 to ensure the stability and reproducibility of the results. For reasoning models that do not support temperature settings, we conducted three evaluation runs.

In PLAT-MC and PLAT-MC$_R$, a model first generates (selects) an answer among possible choices followed by accompanying rationale for its choice. To assess performance, we compute accuracy or F_1. Precision is defined as $n_o/(n_o + n_x)$ while Recall is defined as $(n_o + n_x)/(n_o + n_x + n_u)$ where n_o indicates the number of correct answers, n_x is the number of incorrect answers, and n_u the number of cases where the model was uncertain and refused to make a decision.

5 Result and Analysis

5.1 Multiple-Choice Taxation Questions

5.1.1 Performance of LLMs on PLAT-MC

In PLAT-MC, a model needs to decide whether the imposition of additional penalty tax is legitimate, based on provided facts and claims from both the plaintiff (taxpayer) and the defendant (tax authority) (Table 6 in Appendix). The model is also permitted to refuse to answer if it is not confident. We evaluate nine LLMs (Table 1). The results show that except Exaone3.5, all models shows comparable F_1 scores 0.70–0.79 (col 1), with commercial reasoning model o3 achieving the highest

[6]Qwen3-32B

[7]EXAONE-3.5-32B-Instruct

[8]gpt-4o-2024-11-20,gpt-4.1-2025-04-14, o1-2024-12-17, o3-2025-04-16, o3-mini-2025-01-31

[9]claude-3-7-sonnet-20250219

Table 1: F1 scores on PLAT-MC

Model	F1	P	R	F1-easy	P-easy	R-easy	F1-hard	P-hard	R-hard
Exaone3.5-32B	0.55	0.70	0.46	0.72	0.86	0.61	0.31	0.20	0.62
Qwen3-32B	0.75	0.60	0.98	0.80	0.67	0.95	0.67	0.50	1.00
GPT4o	0.70	0.62	0.81	0.88	0.83	0.92	0.45	0.32	0.80
GPT4.1	0.68	0.67	0.69	0.75	0.69	0.82	0.44	0.31	0.74
Claude3.7-sonnet	0.74	0.63	0.91	0.68	0.53	0.94	0.75	0.65	0.89
Qwen3-32B (reasoning)	0.72(±0.05)	0.57(±0.04)	0.96(±0.06)	0.60 (±0.02)	0.44(±0.02)	0.94(±0.06)	0.69(±0.03)	0.53(±0.03)	0.97(±0.97)
o3-mini	0.69(±0.01)	0.53(0.01)	0.97(0.03)	0.90(±0.03)	0.85(±0.04)	0.95(±0.01)	0.46(±0.03)	0.31(±0.02)	0.95(±0.02)
o1	0.75 (±0.04)	0.62(±0.04)	0.96(±0.02)	0.92(±0.01)	0.86(±0.02)	0.99(±0.01)	0.62(±0.07)	0.47(±0.07)	0.94(±0.02)
o3	0.79(±0.03)	0.65(±0.04)	1.00(±0.0)	0.83(±0.02)	0.71(±0.03)	1(±0.00)	0.77(±0.05)	0.62(±0.08)	1(±0.00)

score 0.79 F_1. Interestingly, all non-reasoning models (col3, rows 1–5) tend to exhibit lower recall compared to reasoning models, suggesting they are more likely to refrain from making a decision. In contrast, they generally achieve higher precision, indicating that when they do respond, their answers are more often correct.

5.1.2 Cases LLMs Cannot Effectively Handle

To gain insight into what aspects LLMs are (not) capable of, we manually analyzed cases where either at least three LLMs answered correctly or at least three LLMs answered incorrectly. LLMs were able to recognize the following principles:

- Ignorance or misunderstanding of tax laws by a taxpayer does not constitute a justifiable reason.[10]
- Mistakes or misunderstandings by tax accountants do not exempt taxpayers from responsibility; the final responsibility always lies with the taxpayer (thus, it is not a justifiable reason).[11]

On the other hand, LLMs shows the following failure patterns.

- When a taxpayer is misled due to the tax authorities' opinion, LLMs were unable to make a clear decision due to a conflict with the principle of legitimate expectation.[12]
- When judges considered various taxplayer-specific circumstances, including the feasibility of fulfilling obligations, LLMs strictly adheres to principles and rules.[13]

[10]Daegu District Court 2015Guhap877

[11]Seoul Administrative Court 2016Guhap56936

[12]Busan High Court 2016Nu11, Seoul High Court 2020Nu43946

[13]Daegu District Court 2018Guhap20506

Based on these, We categorized them into two groups–`Easy` and `Hard`–based on observed reasoning difficulty as our analysis.

5.1.3 Case Categorization

`Easy` group consists of 36 cases where the issue can be clearly spotted and leads to a single normative conclusion based on existing legal rules or precedents as described below.

- Clerical errors or omissions that do not substantially affect the underlying tax amount are not subject to penalty taxes.
- When the tax authority issued an incorrect tax disposition that misled the taxpayer, a penalty on the delayed base tax is considered unlawful.
- Mere misunderstanding or ignorance of the law does not constitute a justifiable reason.
- Claiming ignorance of facts that the taxpayer could have reasonably known is not accepted as a justifiable reason
- Even if a tax attorney, legal representative, or employee was involved in the filing process, the final legal responsibility lies with the taxpayer; thus, no justifiable reason is accepted.

Remaining cases are classified as `Hard` (64 cases).

The categorization reveals that LLMs generally perform well on `Easy` cases (Table 1, col 4–6) where rigid application of rules is sufficient. However, they struggle with `Hard` cases that require flexible legal reasoning, case-specific consideration, or weighing of competing principles depending heavily on the specific factual context as shown below (col 7–9) as described in detail below.

- Cases where the taxpayer faced unavoidable circumstances that hindered payment — these often depend on the judge's perspective and discretion regarding the taxpayer's circumstance.

- Cases requiring proper assessment of whether differences among tax authorities indicate genuine divergence in interpretation and whether the tax law itself was ambiguous.

- Cases requiring assessment of whether the taxpayer, despite delayed payment, promptly fulfilled their obligations upon becoming aware and was otherwise compliant — or, conversely, whether they neglected their duties and failed to exercise due care in tax compliance.

- Cases dealing with whether an official interpretation (e.g., from a tax officer or written inquiry response) qualifies as a public opinion.

This analysis suggests that all LLMs struggle with cases that lack clear reasoning patterns and require a more comprehensive evaluation of all relevant circumstances to reach a decision.

5.1.4 Causes of Low Recall

Non-reasoning models, that do not explicitly employ inference-time scaling, generally exhibit lower recall compared to reasoning models (Table 1 row 1–5 vs row 6–9). This results in a higher absolute number of "Cannot be determined" labels overall. To further investigate this behavior, we removed "Cannot be determined" option from PLAT-MC creating PLAT-MC$_2$ and measured the accuracy instead of F_1.

Notably, when "Cannot be determined" options were removed and non-reasoning models were forced to choose between two candidates, the resulting accuracy was lower than the original precision (Table 1 col 2, row 1–5 vs Table 2 col 1, row 1–5) except Claude3.7-sonnet. This suggests that many of the previously abstained ("Cannot be determined") cases were not simply uncertain but would likely have been incorrectly answered.

Interestingly, while reasoning models are more likely to respond under uncertainty, their decisions are not always reliable when forced to choose, as reflected in their accuracy scores (Table 2, col 1, row 6–9).

5.1.5 Analysis under the IRAC Framework

To further investigate the low performance of LLMs on our task, we investigated non-reasoning and reasoning models through the lens of the IRAC framework.

Table 2: Accuracy comparison of vanilla LLMs on PLAT-MC$_2$ (2 options w/o reasons) and PLAT-MC$_R$ (4 options w/ reasons). Their difference (Δ) is shown at final column.

Model	PLAT-MC$_2$	PLAT-MC$_R$	Δ
Exaone3.5-32B	0.60	0.79	0.19
Qwen3-32B	0.60	0.73	0.13
GPT-4o	0.57	0.78	0.21
GPT-4.1	0.55	0.83	0.28
Claude-3.7-sonnet	0.67	0.84	0.17
Qwen3-32B (reasoning)	0.60 (±0.05)	0.73(±0.02)	0.13
o3-mini	0.53 (±0.02)	0.66(±0.02)	0.13
o1	0.55 (±0.01)	0.69(±0.02)	0.14
o3	0.62 (±0.01)	0.77(±0.04)	0.15

- **I (Issue)**: Both models are generally able to identify the legal issue accurately. In many cases, they correctly articulated the core dispute and built their reasoning on it.

- **R (Rule)**: Upon examining the legal sources and case law cited by the models, we found that many were either outdated (e.g., superseded by newer statutes) or unverifiable in terms of their legal validity or existence.

Motivated by these findings in the Rule component–particularly regarding the reliability and traceability of legal sources–we conducted additional experiments with RAG.

Interestingly, LLMs show similar or decreased performance when using RAG (Table 5) especially in GPT4.1. There may be two potential explanations: (1) even when provided with the appropriate legal rules, LLMs may still struggle with the "Application" and "Conclusion" stages; (2) retrieving truly relevant legal documents remains challenging, as highlighted in recent studies (Zheng et al., 2025; Hou et al., 2025; Minhu Park and Hwang, 2025).

Given that our retrieval pool is relatively small (consisting of 100 precedents and 4,042 statutory articles), we focus first on hypothesis (1). To this end, we construct a new set of multiple-choice questions, where each option includes both a proposed answer ("Conclusion") and its corresponding rationale ("Application").

5.2 Multiple-Choice Questions with Answer Rationals

We extend the binary "lawful" and "unlawful" options from PLAT-MC$_2$ to a set of four answer choices—two labeled as "lawful" and two as

Table 3: RAG scores on PLAT-MC

Model	F1	P	R
Qwen3-32B	0.77 (+0.02)	0.62 (+0.02)	1.00 (+0.02)
GPT4.1	0.60 (-0.08)	0.55 (-0.12)	0.65 (-0.04)
Claude3.7-sonnet	0.74 (0)	0.60 (-0.03)	0.96 (+0.05)
Qwen3-32B (reasoning)	0.73 (+0.01)	0.61 (+0.04)	0.91 (-0.05)
o3	0.75 (-0.04)	0.64 (-0.01)	0.92 (-0,08)

"unlawful"—each accompanied by annotated rationales (see Table 7 in the Appendix). These rationales are plausible but not necessarily correct.

In the resulting benchmark, PLAT-MC$_R$, LLMs achieve higher accuracy scores (ranging from 13% to 28%, as shown in Table 2, col 2) despite the increased difficulty of selecting from four options (compared to the expected baseline accuracy of 50% for PLAT-MC$_2$and 25% for PLAT-MC$_R$under random guessing).

This result suggests that LLMs struggle in the absence of guidance for the "Application" and "Conclusion" stages. It also implies that reasoning from the conclusion—i.e., backward chaining—may be beneficial in legal domains (Poole and Mackworth, 2023; Zhou et al., 2023; Kazemi et al., 2023; Lee and Hwang, 2025).

Notably, non-reasoning models (rows 1–5) show a larger improvement in accuracy compared to reasoning models (rows 6–9), suggesting that hints embedded in plausible rationales provide greater leverage for less capable models. To further explore this observation, we develop an essay-type benchmark.

5.3 Essay-Type Questions

For a comprehensive analysis, we construct PLAT-E, which consists of 100 questions accompanied by corresponding rubrics extracted from the "reasoning" sections of legal precedents. Each question is annotated with either 6 rubrics (94 examples) or 7 rubrics (6 examples). Among these, one rubric in each example specifically evaluates the correctness of the final answer (i.e., whether it is "lawful" or "unlawful").

We assign a score of 1 for each satisfied rubric and normalize the total score based on the number of rubrics. For example, if an answer satisfies 5 out of 6 rubrics, the resulting score is $0.83 = 5/6$. We employ GPT-o3 as a "LLM-as-a-Judge" for automatic evaluation (the prompt is provided in the Appendix). The most competent reasoning model, o3, achieves an average score of 0.82—indicating that its generated answers, on average, satisfy approximately 5 out of 6 rubrics (see Table 4, Column 1, Row 8). However, its accuracy on the conclusion rubric alone is only 0.69. This suggests that among the IRAC stages, the "Conclusion" stage remains the most challenging—even for advanced reasoning models.

Table 4: Accuracy and score comparison across LLMs

Model	Rubric Score	Conclusion Acc (%)
Qwen3-32B	0.57	0.33
Exaone-3.5-32B	0.58	0.29
GPT-4o	0.71	0.40
GPT-4.1	0.68	0.61
Claude3.7-sonnet	0.69	0.43
Qwen3-32B (reasoning mode)	0.69	0.49
o1	0.77	0.52
o3	0.82	0.69
o3-mini	0.52	0.31

5.4 Agent-Based Approach

To address the limitations identified above, we introduce agentic retrieval with multiple personas. Agentic retrieval is proposed because simple retrieval alone does not improve performance (Table 5). We apply the REACT prompt (Yao et al., 2023b) while enforcing a minimum of three retrieval steps. This enforcement is intended to increase the likelihood of identifying relevant legal documents. GPT-4.1 achieves a performance gain of +0.09 F_1 (comparing row 1 to row 5).

In our multi-agent collaboration experiments, three LLMs are assigned specific roles: an attorney for the taxpayer, an attorney for the tax authority, and a judge. We hypothesize that this setup can benefit the "Application" stage by enforcing diverse legal perspectives, thereby bypassing the need to determine which viewpoint is lawful during inference. GPT-4.1, when combined with ReAct prompting and multi-agent collaboration, shows a slight performance improvement (row 1 vs. row 5 vs. row 7).

Interestingly, GPT-4o does not exhibit significant changes. This may be due to the fact that GPT-4o already incorporates inference-time reasoning strategies, reducing the marginal benefit of added prompting or agentic structuring.

6 Conclusion

Here, we introduce PLAT, a benchmark designed to evaluate LLMs' capability in taxation. Compared

Table 5: Agent scores on PLAT-MC. M stands for the experiment with multiple agents with different roles. See Appendix for the prompt

Model	F1	P	R
GPT4.1	0.68	0.67	0.69
o3	0.79(±0.03)	0.65(±0.04)	1.00(±0.0)
GPT4.1 (RAG)	0.60	0.55	0.65
o3 (RAG)	0.75	0.64	0.92
GPT4.1 (REACT)	0.77	0.64	0.99
o3 (REACT)	0.79	0.65	1.00
GPT4.1 (REACT+M)	0.79	0.65	1.00
o3 (REACT+M)	0.74	0.59	0.98

to previous study, our dataset includes cases where answers cannot be determined solely by referencing statutes, requiring a deeper understanding of legal and contextual factors of individual legal issues. Our experiments reveals that while LLMs demonstrate some capability, vanialla models struggle to comprehensively understand taxation issues. We also show that by gradually integrating retrieval, self-reasoning, and multi-agent collaboration with specific roles, these limitations can be partially mitigated, although reaching a correct conclusion remains challenging.

7 Limitation

Tax accountants require a broad range of knowledge and advanced reasoning skills. For instance, the Korean Certified Tax Accountant (CTA) exam, a professional qualification for tax practitioners, covers multiple subjects: multiple-choice exams in Public Finance, Introduction to Tax Law, and Introduction to Accounting; written exams in Tax Law I (covering Corporate Tax Law, Income Tax Law, etc.) and Tax Law II (covering Value-Added Tax Law, Inheritance and Gift Tax Law, etc.). On the other hand, our study focuses specifically on evaluating the justifiability of exemption from additional tax penalties, serving as a case study where LLMs must demonstrate a comprehensive understanding of complex situations, rather than simply referencing related tax statutes. A more wholistic evaluation of LLMs in the tax domain remains as a future work.

References

Ilias Chalkidis. 2023. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark. *SSRN*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Preprint*, arXiv:2401.01301.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Harvey Team. 2024. Harvey co-builds custom model for tax with pwc. Accessed: 2025-02-12.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257*.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2025. CLERC: A dataset for U. S. legal case retrieval and retrieval-augmented analysis generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7898–7913, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2023. LAMBADA:

Backward chaining for automated reasoning in natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6568, Toronto, Canada. Association for Computational Linguistics.

Sung Kyun Kim and Bian Lee. 2008. Effects of "broad and/or vague concept(unbestimmter rechtsbegriff)" in light of "principle of essence(wesentlichkeitstheorie)" —concerning tax law—. *조세법연구*, 14(1):99–135.

Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Wonseok Hwang. 2024. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA. Association for Computational Linguistics.

Jinu Lee and Wonseok Hwang. 2025. Symba: Symbolic backward chaining for structured natural language reasoning. *Preprint*, arXiv:2402.12806.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools.

Eric Martinez. 2023. Re-evaluating gpt-4's bar exam performance.

Eunkyung Choi Minhu Park, Hongseok Oh and Wonseok Hwang. 2025. Lrage: Legal retrieval augmented generation evaluation tool. *Preprint*, arXiv:2504.01840.

John J. Nay, David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. 2024. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philos. Trans. R. Soc. A*, 382(2270).

OpenAI. 2024. O1 system card. Accessed: 2025-02-12.

Hun Park. 2019. Interpretation analysis of judicial precedents on the borrowing concept in tax law. *법조*, 68(3):511–552.

David L. Poole and Alan K. Mackworth. 2023. *Artificial Intelligence: Foundations of Computational Agents*, 3 edition. Cambridge University Press.

LG AI Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Woohyung Lim, Sangha Park, Sooyoun Park, Yongmin Park, Sihoon Yang, Heuiyeen Yeen, and Hyeongu Yun. 2024. Exaone 3.5: Series of large language models for real-world use cases. *Preprint*, arXiv:2412.04862.

Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. 'smolagents': a smol library to build great agentic systems. `https://github.com/huggingface/smolagents`.

Dietrich Trautmann, Natalia Ostapuk, Quentin Grail, Adrian Pol, Guglielmo Bonifazi, Shang Gao, and Martin Gajek. 2024. Measuring the groundedness of legal question-answering systems. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 176–186, Miami, FL, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

In Jun Yang. 2024. Reasonable cause as an exemption requirement of tax penalty. *조세와 법*, 17(1):165–201.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, CSLAW '25, page 169–193, New York, NY, USA. Association for Computing Machinery.

Yan Zhong, Dennis Wong, and Kun Lan. 2024. Tax intelligent decision-making language model. *IEEE Access*, 12:146202–146212.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Example

A.1 PLAT

Table 6: Examples from PLAT-MC.

Facts	Claim from Plaintiff (Taxpayer)	Claim from Defendant (Tax Authority)	Label
1. The plaintiff is a company established for the purpose of shipbuilding and sales. 2. On March 25, 2009, the plaintiff applied to the head of the Jungbu Tax Office for an extension of the payment deadline for KRW 1,200 billion out of KRW 1,453,815,466.13 in corporate tax for the 2008 tax year, and it was approved. 3. The plaintiff paid the remaining corporate tax of KRW 253,815,466.13, for which no extension application was filed, on March 31, 2009, and paid the inhabitant tax on corporate tax to the defendant on April 30, 2009. 4. On June 25, 2009, the plaintiff applied for an additional extension of the corporate tax for which the payment deadline had been extended, and the payment deadline was approved until September 30, 2009. 5. The plaintiff paid KRW 6,313,838,780 in inhabitant tax on corporate tax for the extended corporate tax payment deadline to the defendant on October 30, 2009. 6. The defendant imposed an additional tax of KRW 1,609,397,490, claiming that the plaintiff had not separately applied for an extension of the inhabitant tax payment deadline, even though it had received an extension of the corporate tax payment deadline. 7. After paying the additional tax, the additional tax was revised to KRW 1,105,805,430 according to the reduction decision. 8. The plaintiff applied for an exemption from the additional tax on February 5, 2010, but the defendant rejected it.",	Plaintiff's Arguments and Grounds 1. Illegality of Imposing Penalty Tax Due to Justifiable Cause - The plaintiff received approval for an extension of the corporate tax payment deadline, and therefore mistakenly believed that the resident tax payment deadline was also extended accordingly. - The defendant's staff member answered that the resident tax payment deadline would be extended, leading the plaintiff to believe this. - Therefore, the plaintiff had a justifiable reason for failing to pay the resident tax within the deadline. - Relevant Laws: Framework Act on National Taxes Article 6, Local Tax Act Article 27-2 Paragraph 2. Penalty Tax Exemption is a Mandatory Act and Meets the Exemption Requirements - The plaintiff was facing a significant crisis in its business, which constitutes a reason for penalty tax exemption. - The defendant has an obligation to accept the exemption application. - Relevant Laws: Local Tax Act Article 27-2 Paragraph 2, Enforcement Decree Article 13-2, Article 11 Paragraph 1 Item 4	Defendant's Arguments and Grounds 1. Illegality of Penalty Tax Exemption Application - The application for penalty tax exemption must be made by the end of the statutory payment deadline. - The plaintiff applied on February 5, 2010, past the statutory payment deadline, making it an illegal application. - Relevant Laws: Local Tax Act and related Enforcement Decree 2. Non-Existence of Justifiable Cause - The extension of the corporate tax payment deadline is irrelevant to the extension of the resident tax payment deadline. - The plaintiff merely misunderstood the laws and administrative interpretations, and there was no response from the defendant's staff member. - Therefore, the plaintiff has no justifiable reason. 3. Does Not Fall Under the Penalty Tax Exemption Requirements - The plaintiff does not fall under the penalty tax exemption requirement of ẅhen the business is in a significant crisis. - Therefore, the rejection of the exemption application is lawful. - Relevant Laws: Local Tax Act Enforcement Decree Article 11 Paragraph 1 Item 4	Not legitimate.
1. The plaintiff operated a charging station from 2011 and opened and reported a business account. 2. In 2013, the plaintiff's revenue exceeded 300 million won, making them obligated to use double-entry bookkeeping from January 1, 2014 (Article 160, Paragraph 3 of the former Income Tax Act; Article 208, Paragraph 5, Subparagraph 2 of the former Enforcement Decree of the Income Tax Act). 3. The plaintiff newly opened this charging station on April 1, 2014, and opened five deposit accounts (hereinafter referred to as 'the deposit accounts in this case') at NongHyup Bank, but did not report the opening of a business account to the competent tax office by June 30, 2015 (Article 160-5, Paragraph 3 of the former Income Tax Act). 4. In early May 2015, the plaintiff confirmed that 'Not applicable' was written in the 'Non-establishment of business account' item among the 'Penalty items' of the '2014 Comprehensive Income Tax Return Guidance' received from the defendant. 5. As a person subject to faithful reporting, the plaintiff reported the ending balance of some of the deposit accounts in this case when filing comprehensive income tax returns for 2014 and 2015.	Plaintiff's Arguments 1. Existence of Justifiable Grounds: The plaintiff received a '2014 Tax Year Comprehensive Income Tax Notice' in May 2015, which stated 'Business Account Not Established' as 'Not Applicable.' Therefore, the plaintiff believed that these deposit accounts had already been reported. Thus, there are justifiable grounds for failing to fulfill the reporting obligation, and the plaintiff should be granted a reduction of penalties under Article 48 of the Framework Act on National Taxes. 2. Violation of the Principle of Taxation Based on Substance: The plaintiff actually used these deposit accounts as business accounts and fully listed the ending balance of the accounts in the faithful declaration confirmation form when filing comprehensive income tax. The defendant was able to understand the account details through this, but the defendant imposed penalties for the formal reason that it was not simply reported, which violates the principle of taxation based on substance in tax law.	Defendant's Arguments 1. Legality of Penalty Imposition: The defendant argues that the penalty imposition is lawful because the plaintiff, as a person obligated to use double-entry bookkeeping, did not fulfill the obligation to report business accounts under Article 160-5 of the former Income Tax Act.	Legitimate.

Table 7: PLAT-MC$_R$version with precedents in Table 6

A	It is lawful to impose a penalty because an additional extension was requested for the extended deadline for corporation tax payment.
B	Since approval was obtained from the Jungbu Regional Tax Office for an extension of the payment deadline for a portion of the corporate tax for the 2008 tax year, the imposition of penalties is lawful.
C	Since the business is facing a significant crisis and falls under the grounds for exemption from additional tax, the imposition of additional tax is not lawful.
D	It is not legitimate to impose a penalty tax because the company was established for the purpose of shipbuilding and sales.
A	Even if an amended import tax invoice is issued in cases where a certificate of origin is prepared differently from the facts, penalties will be imposed.
B	It is lawful to impose additional tax when the plaintiff secures exclusive domestic and Asian distribution rights for imported premium overseas clothing.
C	The imposition of penalties is lawful for the issuance of amended import tax invoices, regardless of the issuance date.
D	The legality of a penalty depends on whether there is a justifiable reason for non-compliance with the tax obligation. Since the plaintiff had reasonable grounds to believe in the origin based on valid certificates of origin and labels, the imposition of a penalty is unlawful because a justifiable reason is recognized.

Table 8: PLAT-E rubrics with precedents in Table 6

Rubric Items
Case 1: Penalty Tax Exemption (Precedent 1) 1. Whether the nature and definition of the penalty tax are described (1 point) 2. Abundant description of regulations related to justifiable reasons for penalty tax exemption (1 point) 3. Whether the relationship between the rejection of penalty tax exemption and the imposition of penalty tax, and whether the defect is succeeded, are described: The relationship between the imposition of penalty tax and the rejection is described, but the description of whether it is obviously invalid is well done (1 point) 4. Describe whether the application for penalty tax exemption was made after the application deadline, and describe the legality of the timing of the application for penalty tax exemption (1 point) 5. Specify whether the reason for penalty tax exemption, 'when the business is in a serious crisis,' exists and whether the judgment of its existence is organically connected to the facts and legal basis: The plaintiff misunderstood that the deadline for reporting and paying corporate income tax surtax would also be extended, and failed to pay the resident tax within the statutory payment deadline, so there are some mitigating circumstances for the failure to fulfill the resident tax payment obligation, and the legislative intent is to exempt penalty tax if there is a reason to extend the reporting and payment deadline. Mentions of these points are well made (1 point) 6. Clearly state that there is a justifiable reason for penalty tax exemption (1 point) 7. The correct answer is "The additional tax is unlawful", and whether the correct answer was provided (1 point)
Case 2: FTA Origin Misclassification (Precedent 2) 1. Whether the legal basis and nature of the penalty tax are clearly described (1 point) 2. Whether the existence of a justifiable reason for exemption from the penalty tax is logically described based on the legal basis and specific facts related to the justifiable reason (1 point) 3. Whether it is specifically described whether the plaintiff has fulfilled the duty of reasonable care, such as submitting a valid certificate of origin issued by a certified exporter and labeling, and whether a justifiable reason for non-compliance with the tax obligation is acknowledged as a result (1 point) 4. Whether the existence of the obligation to issue a revised import tax invoice is logically explained based on the relevant laws and regulations (Article 35 of the Value-Added Tax Act, Article 72 of the Enforcement Decree, etc.) and whether the importer is responsible or has made a simple error in determining the illegality of the refusal to issue the revised import tax invoice (1 point) 5. Whether the credibility of the defendant's grounds for verification of origin (internet postings, etc.) and the absence of the plaintiff's fault are specifically described, and whether the illegality of the penalty tax and revised import tax invoice-related dispositions are clearly judged in conclusion (1 point) 6. The correct answer is "The additional tax is unlawful", and whether the correct answer was provided (1 point)

Table 9: LLM-as-a-Judge Scoring Prompt

Prompt Template
You are a tax law expert chatbot who responds kindly and logically to users' questions. You are also a fair and objective grader who strictly follows the evaluation rubric. Evaluate the following answer according to the given rubric. At the beginning of your response, state the total score in the following format: **"Total Score: X point(s)"**. **Answer to be evaluated:** {model_ans} **Evaluation Rubric:** {rubric} **Total Score:**

A.2 Dataset Annotation Procedure

For PLAT-MC, one author searched for and annotated 100 precedents over a span of 15 hours. Another author subsequently reviewed all annotated cases to ensure consistency and accuracy.

For PLAT-MC$_R$, one author annotated 100 precedents over 22 hours. Another author reviewed all the annotated precedents along with the associated multiple-choice options, revising 29 of them. This collaborative review process enabled the construction of a high-quality dataset.

For PLAT-E, we randomly selected 10 precedents—comprising 5 lawful and 5 unlawful cases. These were manually annotated and used for 10-shot learning. The `gpt-4.1-2025-04-14` model was utilized in this phase.

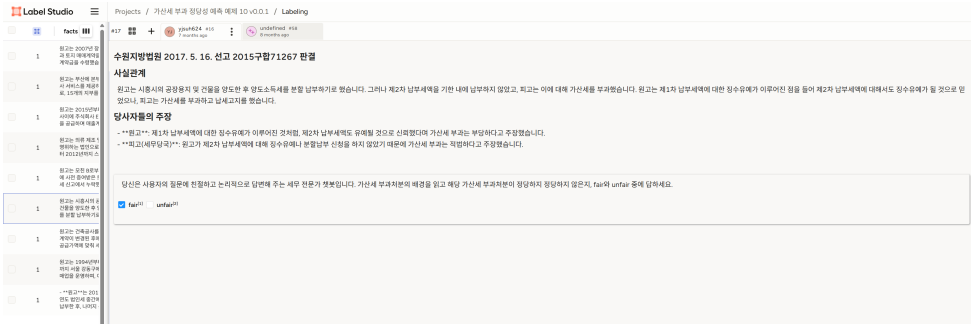

Figure 1: The screen interface of Label Studio used for manual annotation for PLAT-MC.

All datasets were annotated in accordance with the Court's legal reasoning and the Judge's final decisions. The annotation GUI are shown in Figure 1 and Figure 2.

Figure 2: The screen interface of Label Studio used for manual annotation for PLAT-MC$_R$.

A.3 Prompt for Vanilla LLM and RAG

Table 10: Example. Original Korean is translated to English using GPT-4o

System Prompt	Vanilla LLM Input	LLM with RAG Input
You are a tax expert chatbot that provides friendly and logical answers to users' questions.	Please read the background and materials related to the imposed penalty tax presented above. Based on this information, determine whether the penalty tax is "Legitimate", "Not legitimate", or, if a clear conclusion cannot be reached, answer "Unknown". Then, provide an explanation for your choice. Problem description: {precedent}	Please read the background and materials related to the imposed penalty tax presented above. Based on this information, determine whether the penalty tax is "Legitimate", "Not legitimate", or, if a clear conclusion cannot be reached, answer "Unknown". Then, provide an explanation for your choice. Problem description: {precedent} Reference material: {raged_doc} ###Answer:

A.4 Agentic RAG

Table 11: Agent RAG. The default prompt from ToolCallAgent of `smolagent` libary is used.

System Prompt	Input
Here are the rules you must always follow to complete the task successfully: You must provide at least one tool invocation. If you do not, the task will fail. Use the correct arguments for each tool. Do not pass variable names as arguments—always use actual values. Only call tools when necessary. If you already have enough information, do not call the search agent. Try to solve the task yourself first. The more tools you call, the more hints you will gather, which will guide you toward the correct final answer. You must call at least two different tools besides the final answer tool. If you can determine the final answer, return it using the final answer tool. The retrievertool must be called at least three times. The retrievertool works best in synergy with other tools. So, whenever you call retrievertool, follow up with calls to other relevant tools. Do not repeat tool calls with the exact same parameters as a previous invocation.	Please read the background and materials related to the imposed penalty tax presented above. Based on this information, determine whether the penalty tax is "Legitimate", "Not legitimate", or, if a clear conclusion cannot be reached, answer "Unknown". Then, provide an explanation for your choice. Problem description: {precedent} Reference material: {raged$_{doc}$}$###Answer:$

A.4.1 Various tool we built in Agentic RAG

Table 12: Prompt with Legal-Analyzer tool

Prompt
You are a skilled legal expert tasked with evaluating legal reasoning responses. Use the given context to answer the question accurately and naturally. You must strictly adhere to the following formatting rules: After completing your analysis and reasoning, the final line of your response must be in the format: "The answer is final conclusion" Do not include any additional explanation or reasoning after the phrase "The answer is". The phrase "The answer is" must appear exactly once, and only as the last line of your response. Analyze the given legal case scenario by following the structured steps below: <issue> Identify the key legal issue in the case. </issue> <rule> Clearly state the statutes or legal principles that govern the identified issue. </rule> <application> Analyze how the above rules or principles apply to the specific facts of the case. Discuss the legal validity of the case based on that application. </application> <conclusion> Synthesize your analysis and provide the likely legal conclusion based on the application of law to the issue. </conclusion> Problem description: precedent

A.4.2 Prompt for virtual-court tool in Agentic RAG

Table 13: Prompt with Document-Comparison Tool

Prompt Template
The above document contains statutes and precedents retrieved in relation to penalty taxes. You must carefully review the document. Summarize the relevant statutes and precedents according to the following format: Format: The relevant statutes are as follows. "Statute1 ..." The parts of the statute that are relevant to the issue are as follows. "Relevance1-1 ... Relevance1-2 ..." The parts of the statute that are not relevant to the issue are as follows. "Irrelevance1-1 ... Irrelevance1-2 ..." The relevant precedents are as follows. "Precedent1 ..." The parts of the precedent that are relevant to the issue are as follows. "Relevance2-1 ... Relevance2-2 ..." The parts of the precedent that are not relevant to the issue are as follows. "Irrelevance2-1 ... Irrelevance2-2 ..." Generate only the results you identified from the document. Do not include any additional explanations.

Table 14: Prompt with plaintiff's lawyer role-playingtool

Prompt
You are a lawyer representing the plaintiff (the taxpayer) in a tax case. For the following issue, argue unconditionally from the taxpayer's perspective that the imposition of the penalty tax is not legitimate.

Table 15: Prompt with defendant's lawyer role in role-playingtool

Prompt
You are a lawyer representing the defendant (the tax authority) in a tax case. For the following issue, argue unconditionally from the tax authority's perspective that the imposition of the penalty tax is legitimate.

Table 16: Prompt with judge role in role-playingtool

Prompt
You are a neutral tax judge. Below are the arguments from both parties: [Plaintiff's Argument] claim_a [Defendant's Argument] claim_b Please compare the arguments from both sides, evaluate their validity, and reach a final conclusion based on legality and logical reasoning. If you identify any flaws, point them out and correct them to present the proper conclusion.