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, research dedicated to taxation remains scarce. Moreover, the datasets used in these studies are either simplified, failing to reflect the real-world complexities, or not released 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 300 examples: (1) 100 binary-choice questions, (2) 100 multiple-choice questions, and (3) 100 essay-type questions, all derived from 100 Korean court 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 cases involving conflicting issues that require a comprehensive understanding (not only of the statutes but also of the taxpayer's circumstances), and (2) LLMs struggle particularly with the "AC" stages of "IRAC"[1] 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 was an early focus since OpenAI's demonstration that GPT-4 passes the U.S. Uniform Bar Exam (Martinez, 2023). To more robustly assess LLMs' capabilities in the legal domain beyond 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). Also, with rapid progress in LLMs and advances 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[2], a benchmark consisting of 300 questions derived from Korean precedents concerning the legitimacy of additional tax penalties. Article 48 of the Korean Framework Act on National Taxes[3] allows exemptions from penalty taxes in cases of *justifiable reasons*, but the statute does not explicitly define what constitutes such reasons. In practice, this gap has been filled by judicial precedents, where the courts have elaborated on and recognized specific circumstances as qualifying "justifiable reasons". Thus, we designed PLAT to assess LLMs' tax law comprehension and inferential ability–particularly in scenarios where issues cannot be resolved by statute alone and call for real-world legal judgment, including leniency, compassion, and discretionary reasoning typically

[1] IRAC: Issue, Rule, Application, Conclusion. A structured method of legal analysis used to organize legal arguments and reasoning.

[2] PREDICTING THE LEGITIMACY OF PUNITIVE ADDITIONAL TAX

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

exercised by human judges–such as weighing competing legal principles or determining whether it is reasonable to expect a taxpayer to recognize and comply with the law.

Our experiments with two open-source LLMs–Qwen3 (Yang et al., 2025), Exaone3.5 (Research et al., 2024)–alongside eight commercial LLMs (GPT 4o, 4.1, o1, o3, o3-mini, Claude 3.7, Gemini2.5-flash-preview and 2.5-pro-preview) show that the strongest reasoning models, o3 and Gemini2.5-pro, achieve an F_1 score of 79% on PLAT. A detailed analysis reveals that, while LLMs perform well on relatively simple problems, their accuracy declines when a more inferential understanding is required.[4]

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 to explore all the "IRAC" factors.

Our findings reveal that (1) LLMs are 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 agentic RAG; (3) LLMs underperform in the "Application" and "Conclusion" stages. Without inference-time scaling, they often hesitate to proceed, resulting in low recall, whereas 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 formats, final "Conclusion" accuracy remains limited.

In summary, our 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 ten 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 the 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 the Taxation domain

Nay et al. (2024) study 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 dataset has not been released.

Holzenberger et al. (2020) develop SARA, a statutory reasoning dataset constructed from a simplified version of the U.S. Internal Revenue Code. The dataset consists of two tasks: determining entailment relations and calculating tax amounts based on given statutes 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) suggest a retrieval-based LLM system designed to answer tax-related questions typically handled by tax departments. The dataset has not been released.

Compared to these studies, our dataset consists of 300 manually constructed examples, supervised by tax professionals.[5] PLAT is particularly distinct from previous datasets in that its questions cannot be answered solely by referencing statutes. Instead, they require an 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 generate 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.

[4] For instance, all LLMs correctly recognize that ignorance or misunderstanding by taxpayers cannot serve as a justifiable 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.

[5] Authors of this paper. Two professors, one PhD student in Taxation Science and one undergraduate student holding a national tax license.

These processes can be iteratively orchestrated by the LLMs themselves. Below, we highlight a few representative works.

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

Roucher et al. (2025a) introduce `smolagents`, open-source framework designed for simplicity and seamless Python code integration. The smolagents framework is employed in this study.

3 Datasets

3.1 Motivation

In Korea, an additional penalty tax can be applied to all 25 types of taxes. It is an extra economic burden imposed on taxpayers who fail to properly file or pay their taxes, in addition to the original tax liability.[6] However, when objective circumstances prevent a taxpayer from fulfilling their obligations, it is more appropriate not to impose the penalty tax, even if a legal basis exists.

Indeed, 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 lacking a clear scope, requiring interpretation in specific cases (Kim and Lee, 2008; Yang, 2024; Park, 2019). In such situations, precedents act as a interpretative standards.[7]

Thus, resolving whether a penalty tax is justified requires not only referencing the statutes but also comprehensively understanding the individual situation like human judges.[8] In this regard, we build PLAT, a benchmark constructed from 100 Korean precedents–50 justifiable, 50 not justifiable cases–addressing the legitimacy of additional tax penalties.

3.2 Dataset Construction

We first collect relevant precedents using the commercial Korean legal search engine LBOX[9], 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 pool.

From the candidate pool, we sampled 10 precedents and extracted facts and claims from them using o3. Then a tax professional (one of our authors) manually evaluated 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 there any inaccuracies or fabricated information in the extracted facts and claims?
- Legal Correctness: Are the labels extracted from court rulings consistent with the actual court decisions?

Under the result of evaluation, we curated final 100 examples, with an equal split: 50 cases where the court ruled the exemption from penalty tax was justified, and 50 cases where the court decided that the exemption was not. Each example required approximately 30–40 minutes for evaluation, resulting in a total of 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. (Examples of PLAT are presented in Appendix B.)

[6] Korea's additional tax penalties function as punitive measures, often imposing heavy burdens even for minor taxpayer errors. Their growing scale (e.g., in 2024, VAT (Value-Added-Tax) penalties rose by 40% (∼$400m)) underscores the need to evaluate their legitimacy, a task for which PLAT aims to "plat" (balance) state authority and taxpayer burden onto the same analytical plane, thereby strengthening the protection of taxpayers' rights.

[7] 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.

[8] This also indicate that the topic studied here reflects a real-world legal decision-making, an area in which current LLMs show limited capability. 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.

[9] `https://lbox.kr/v2`

- PLAT-MC: Each question provides two answer choices–"lawful" and "unlawful". In practice, especially in the legal domain where reliability is critical, it is important for models to be able to express uncertainty. Therefore, we included an additional "don't know" option for cases where the model is uncertain. (Table 10 in Appendix B.1)

- 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. All 400 options were manually labeled, evaluated, and double-checked by two tax professionals[10]. (Table 11 in Appendix B.2)

- 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. The task-specific rubrics are designed based on the IRAC framework, curated through discussion among three tax professionals who also served as human judges for the LLM essay responses. (Appendix B.3)

Additional details about PLAT construction and the annotation GUI are provided in Appendix C.

4 Experiments

4.1 Models

We used two open-weight LLMs (Qwen3-32B[11], LG EXAONE3.5-32B[12]) and eight commercial LLMs (GPT 4o, 4.1, o1, o3, o3-mini[13], Claude 3.7 sonnet[14], Gemini2.5-flash-preview and 2.5-pro-preview[15]). For all experiments with non-reasoning models, we set the temperature to 0. For reasoning models that do not support temperature settings, we conducted three independent runs.

[10] All annotations for the dataset were performed by an undergraduate student holding a national tax license and by a PhD student in Taxation Science (Authors of this paper). The inter-annotator agreement rate was 92%.

[11] Qwen3-32B

[12] EXAONE-3.5-32B-Instruct

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

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

[15] gemini-2.5-flash-preview-05-20, gemini-2.5-pro-preview-06-05. We evaluated PLAT-E using the stable (non-preview) version, as the preview version is no longer available.

4.2 Retriever Setup

We used Pyserini (Lin et al., 2021) with the BM25 algorithm with default hyperparameters. We limited the retriever to the top-3 documents, based on initial experiments comparing 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. Articles are filtered from the Korean Statutes Corpus (Kim et al., 2024). The precedents in the retrieval pool were processed using o3 to extract facts and claims. The source precedent for each question was excluded from the retrieval pool to prevent information leakage. `smolagents` was employed (Roucher et al., 2025a) to build LLM agents.

4.3 Metrics

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 (PLAT-MC$_R$) or F_1(PLAT-MC). To this, we define Precision and Recall as in equation 1 and equation 2.

$$\text{Precision} := \frac{n_o}{n_o + n_x} \quad (1)$$

$$\text{Recall} := \frac{n_o + n_x}{n_o + n_x + n_u} \quad (2)$$

Where n_o denotes the number of correct answers, n_x the number of incorrect answers, and n_u the number of cases in which the model refused to answer (i.e., selected the "don't know" option). Precision measures the accuracy among the answered questions, whereas Recall measures the proportion of answered questions relative to the total.

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 10 in Appendix B.1) . We evaluate ten LLMs and permit them to refuse to answer if they are not confident. (Table 1). The results show that except for Exaone3.5 and GPT-4.1, all models show comparable F_1 scores 0.70-0.79 (col 1), with the commercial reasoning model o3 and Gemini2.5-pro achieving the highest score 0.79 F_1. Interestingly, non-reasoning models (col 3, rows

Table 1: F1 scores on PLAT-MC. Numbers in parenthesis indicate standard deviation from three independent runs.

Model	F1	P	R	F1-easy	P-easy	R-easy	F1-hard	P-hard	R-hard
Exaone3.5-32B	0.55	0.69	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
Claude3.7-sonnet	0.74	0.63	0.91	0.68	0.53	0.94	0.75	0.65	0.89
Gemini2.5-flash	0.73	0.63	0.88	0.78	0.69	0.89	0.70	0.59	0.88
GPT-4o	0.70	0.62	0.81	0.88	0.83	0.92	0.45	0.32	0.80
GPT-4.1	0.65	0.66	0.64	0.75	0.76	0.74	0.49	0.42	0.64
Qwen3-32B$^{\dagger)}$	$0.72_{(\pm 0.05)}$	$0.57_{(\pm 0.04)}$	$0.96_{(\pm 0.06)}$	$0.60_{(\pm 0.02)}$	$0.44_{(\pm 0.02)}$	$0.94_{(\pm 0.06)}$	$0.69_{(\pm 0.03)}$	$0.53_{(\pm 0.03)}$	$0.97_{(\pm 0.03)}$
o3-mini	$0.69_{(\pm 0.01)}$	$0.53_{(\pm 0.01)}$	$0.97_{(\pm 0.03)}$	$0.90_{(\pm 0.03)}$	$0.85_{(\pm 0.04)}$	$0.95_{(\pm 0.01)}$	$0.46_{(\pm 0.03)}$	$0.31_{(\pm 0.02)}$	$0.95_{(\pm 0.02)}$
o1	$0.75_{(\pm 0.04)}$	$0.62_{(\pm 0.04)}$	$0.96_{(\pm 0.02)}$	$0.92_{(\pm 0.01)}$	$0.86_{(\pm 0.02)}$	$0.99_{(\pm 0.01)}$	$0.62_{(\pm 0.07)}$	$0.47_{(\pm 0.07)}$	$0.94_{(\pm 0.02)}$
o3	$\mathbf{0.79}_{(\pm 0.03)}$	$0.65_{(\pm 0.04)}$	$1.00_{(\pm 0.00)}$	$0.83_{(\pm 0.02)}$	$0.71_{(\pm 0.03)}$	$1.00_{(\pm 0.00)}$	$0.77_{(\pm 0.05)}$	$0.62_{(\pm 0.08)}$	$1.00_{(\pm 0.00)}$
Gemini2.5-flash$^{\dagger i}$	$0.74_{(\pm 0.02)}$	$0.62_{(\pm 0.03)}$	$0.91_{(\pm 0.01)}$	$0.78_{(\pm 0.03)}$	$0.69_{(\pm 0.05)}$	$0.90_{(\pm 0.02)}$	$0.71_{(\pm 0.04)}$	$0.59_{(\pm 0.05)}$	$0.91_{(\pm 0.02)}$
Gemini2.5-pro$^{\dagger i}$	$\mathbf{0.79}_{(\pm 0.01)}$	$0.68_{(\pm 0.01)}$	$0.95_{(\pm 0.01)}$	$0.78_{(\pm 0.05)}$	$0.65_{(\pm 0.06)}$	$0.96_{(\pm 0.02)}$	$0.80_{(\pm 0.03)}$	$0.70_{(\pm 0.04)}$	$0.95_{(\pm 0.01)}$

† Reasoning mode, ^{i} thinking-budget: 4096

1-6) tend to exhibit lower recall compared to reasoning models, suggesting that they are more likely to refrain from making a decision.[16] In contrast, they generally achieve higher precision, indicating that when they do respond, their answers are more often correct. These findings align with recent works indicating that reasoning models show similar or even lower performance on complex tasks (Shojaee et al., 2025), and tend to exhibit overconfidence (Mei et al., 2025).

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 the majority of LLMs answered correctly or incorrectly.[17] LLMs were able to recognize the following principles:

- Ignorance or misunderstanding of tax laws by a taxpayer does not constitute a justifiable reason.[18]
- 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).[19]

On the other hand, LLMs show 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.[20]
- When judges considered various taxpayer-specific circumstances, including the feasibility of fulfilling obligations, LLMs tended to strictly adhere to principles and rules.[21]

Based on these observations, we categorized them into two groups–`Easy` and `Hard`–according to observed reasoning difficulty.

5.1.3 Case Categorization

The `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.

[16] Motivated by the difference in the number of "don't know" responses between reasoning and non-reasoning models, we conducted a qualitative analysis. The non-reasoning models often provide logically sound (and sometimes even correct) explanation process well but they eventually conclude with statements such as "However, since the exact determination cannot be made," or "Yet, because the tax authority's criteria are often stringent, I answer "don't know".)

[17] We conducted exploratory analysis with GPT-4o, GPT-o1-mini, Claude-3.5-sonnet, o1, EXAONE-3.0-7.8B-Instruct, and Qwen2.5-7B-Instruct

[18] Daegu District Court 2015Guhap877

[19] Seoul Administrative Court 2016Guhap56936

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

[21] Daegu District Court 2018Guhap20506

The remaining 64 cases are classified as `Hard`.

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 (not only of the statutes but also of the taxpayer's 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 1-6 vs row 7-12). This results in a higher absolute number of "Don't know" labels overall. To further investigate this behavior, we removed the "Don't know" option from PLAT-MC creating PLAT-MC$_2$ and measured the accuracy instead of F_1.

Notably, when "Don't know" options were removed and non-reasoning models were forced to choose between two candidates, the accuracy was lower than their original precision (Table 1 col 2, row 1-6 vs Table 2 col 1, row 1-6) except for Claude3.7-sonnet and Gemini2.5-flash. This suggests that many of the previously abstained ("Don't know") 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 1 col 2, row 7-12 vs Table 2, col 1, row 7-12).

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
Claude3.7-sonnet	0.67	0.84	0.17
Gemini2.5-flash	0.68	0.79	0.11
GPT-4o	0.57	0.78	0.21
GPT-4.1	0.55	0.83	0.28
Qwen3-32B†	$0.60_{(\pm 0.05)}$	$0.73_{(\pm 0.02)}$	0.13
o3-mini	$0.53_{(\pm 0.02)}$	$0.66_{(\pm 0.02)}$	0.13
o1	$0.55_{(\pm 0.01)}$	$0.69_{(\pm 0.02)}$	0.14
o3	$0.62_{(\pm 0.01)}$	$0.77_{(\pm 0.04)}$	0.15
Gemini2.5-flash$^{\dagger i}$	$0.61_{(\pm 0.01)}$	$0.72_{(\pm 0.01)}$	0.11
Gemini2.5-pro$^{\dagger i}$	$0.66_{(\pm 0.01)}$	$0.80_{(\pm 0.00)}$	0.14

†Reasoning mode
^{i}thinking-budget: 4096

5.1.5 Analysis under the IRAC Framework

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

- **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 cited by the models, we found that many were either outdated (e.g., superseded by newer statutes) or unverifiable.

Motivated by these findings, we conducted additional experiments with RAG[22].

Interestingly, LLMs show at most a small improvement, or even degradation when using RAG (Table 3). 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; Park et al., 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

[22]Retrieval-Augmented Generation

end, we construct a new set of multiple-choice questions, where each option includes both a proposed answer ("Conclusion") and its corresponding rationale ("Application").

Table 3: RAG scores on PLAT-MC

Model	F1	P	R
Qwen3-32B	0.75	0.60	0.98
Qwen3-32B (RAG)	0.77	0.62	1.00
Claude3.7-sonnet	0.74	0.63	0.91
Claude3.7-sonnet (RAG)	0.74	0.60	0.96
GPT-4.1	0.65	0.66	0.64
GPT-4.1 (RAG)	0.64	0.57	0.76
Qwen3-32B†	$0.72_{(\pm 0.05)}$	$0.57_{(\pm 0.04)}$	$0.96_{(\pm 0.06)}$
Qwen3-32B† (RAG)	$0.73_{(\pm 0.02)}$	$0.61_{(\pm 0.05)}$	$0.91_{(\pm 0.14)}$
o3†	$0.79_{(\pm 0.03)}$	$0.65_{(\pm 0.04)}$	$1.00_{(\pm 0.00)}$
o3† (RAG)	$0.75_{(\pm 0.01)}$	$0.64_{(\pm 0.01)}$	$0.92_{(\pm 0.04)}$

Column names without bracket are taken from Table 1 for comparison with the results after adding RAG.
†Reasoning mode. ^{i} thinking-budget: 4096

5.2 Multiple-Choice Questions with Answer Rationales

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 "unlawful"–each accompanied by annotated rationales. These rationales are plausible but not necessarily correct (Table 11 in Appendix B.2).

In the resulting benchmark, PLAT-MC$_R$, LLMs achieve higher accuracy scores (+ 0.13–0.28, Table 2, col 2) despite the increased difficulty of selecting from four options (compared to the expected baseline accuracy of 0.50 for PLAT-MC$_2$ and 0.25 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 1–6) 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 analysis, we develop an essay-type benchmark.

5.3 Essay-Type Questions

For a comprehensive analysis, we construct PLAT-E, which consists of 100 questions. For automatic evaluation using LLMs, we built ten task-specific rubrics made by tax professionals. Each rubric is aligned with the elements of IRAC (Table 12 in Appendix B.3). We assign 1 point for each rubric.

5.3.1 Building a judge LLM

In essay-type examinations, scoring results can vary across graders even with the presence of rubrics (Smolentzov, 2013; Park, 2015; Doi et al., 2024). Thus, it is standard to employ multiple graders in major essay-based examinations, including legal qualification exams. Following this practice, we conducted a human evaluation by asking three tax experts[23] to grade a random sample of 30 responses generated by o3.

Next, to build a LLM judge, we benchmarked multiple LLMs by measuring their correlation with these expert averages. Among them, o3 achieved the highest correlation and was therefore selected as the judge model (Pearson r = 0.67, Spearman ρ = 0.70, Figure 1 in Appendix). The inter-human Pearson correlation scores are $r = 0.79$, $r = 0.53$, and $r = 0.58$ respectively.

Using o3 as the judge, we conducted further analysis. We found that (1) o3 tends to assign more generous scores compared to human graders (although there is a rank-wise correlation between them), and (2) o3 cannot judge "Rule"-related rubrics where it needs to verify the cited legal documents.[24] Based on this observation, the four "Rule"-related rubrics were excluded and the maximum score per question was set to 6. (See Appendix F for more details.)

5.3.2 Result

Gemini2.5-pro achieves the top score (0.71) indicating that its generated answers, on average, satisfy approximately 4.25 out of 6 rubrics (Table 5, bottom). Notably, reasoning models do not show clearly superior performance raising questions about the credibility of reasoning process on complex real-world tasks (Shojaee et al., 2025; Mei et al., 2025).

5.4 Agent-Based Approach

5.4.1 Motivation

The experiments with PLAT-MC and PLAT-MC$_R$ reveal that LLMs struggle with the "Application"

[23] Authors of this paper. One PhD student and two professors in Taxation Science.

[24] Consistently, the responses from o3 achieves zero-score 91% times on "Rule"-related rubrics when judged by human experts. This is in line with recent study from Dahl et al. (2024)

Table 4: Agent scores on PLAT-MC (See Appendix B for the prompt).

Model	F1	P	R	F1-easy	P-easy	R-easy	F1-hard	P-hard	R-hard
GPT-4.1	0.65	0.66	0.64	0.75	0.76	0.74	0.49	0.42	0.64
o3	$0.79_{(\pm 0.03)}$	$0.65_{(\pm 0.04)}$	$1.00_{(\pm 0.00)}$	$0.83_{(\pm 0.02)}$	$0.71_{(\pm 0.03)}$	$1_{(\pm 0.00)}$	$0.77_{(\pm 0.05)}$	$0.62_{(\pm 0.08)}$	$1_{(\pm 0.00)}$
GPT-4.1 (RAG)	0.64	0.57	0.76	0.62	0.56	0.74	0.66	0.57	0.77
o3 (RAG)	$0.75_{(\pm 0.01)}$	$0.64_{(\pm 0.01)}$	$0.92_{(\pm 0.04)}$	$0.85_{(\pm 0.03)}$	$0.80_{(\pm 0.05)}$	$0.91_{(\pm 0.03)}$	$0.69_{(\pm 0.02)}$	$0.55_{(\pm 0.01)}$	$0.93_{(\pm 0.05)}$
GPT-4.1 (ReAct)	0.76	0.63	0.97	0.74	0.60	0.97	0.77	0.64	0.96
o3 (ReAct)	$0.78_{(\pm 0.02)}$	$0.64_{(\pm 0.02)}$	$0.99_{(\pm 0.01)}$	$0.76_{(\pm 0.08)}$	$0.62_{(\pm 0.10)}$	$0.97_{(\pm 0.02)}$	$0.77_{(\pm 0.04)}$	$0.63_{(\pm 0.05)}$	$0.99_{(\pm 0.01)}$

Table 5: Essay Score comparison across LLMs

Model	Rubric Score
Exaone-3.5-32B	0.43
Qwen3-32B	0.48
Claude3.7-sonnet	0.62
Gemini2.5-flash	0.68
GPT-4o	0.61
GPT-4.1	0.65
Qwen3-32B†	$0.45_{(\pm 0.00)}$
o3-mini	$0.48_{(\pm 0.01)}$
o1	$0.59_{(\pm 0.01)}$
o3	$0.66_{(\pm 0.01)}$
Gemini2.5-flash^{†,i}	$0.70_{(\pm 0.01)}$
Gemini2.5-pro^{†,i}	**0.71**$_{(\pm 0.01)}$

†Reasoning mode
^{i} thinking-budget: 4096

and "Conclusion" stages. Another experiment with PLAT-E, together with the qualitative analyses of the PLAT-MC [25], reveals vanilla LLMs cannot handle "Rule" stage often citing outdated or unverifiable statutes and precedent. Based on these observation, we introduce agentic retrieval.

5.4.2 Design

We use the ReAct prompt (Yao et al., 2023b) that consists of three main steps: thought, action, and observation. We hypothesize that this setup will benefit the "Rule" stage by iteratively refining retrieved legal documents. In this regard, we enforce a minimum of three retrieval steps.

For "action" steps, we introduce three tools: a BM25 retriever, a legal-document comparison tool, and a legal-analyzer tool. The legal document-relevance-analyzer tool classifies and summarizes relevant legal documents while the legal-analyzer tool to enables LLMs to reason following IRAC. Further details for these legal think tools are available in Appendix E.2.

5.4.3 Result & Analysis

With the agentic-based approach, GPT-4.1 achieves +0.11 F_1 (row 1 vs 5, col 1) while o3 shows -0.01 F_1. In both cases, the agentic retrieval demonstrates superior performance compared to RAG (+0.12 F_1 for GPT-4.1 and +0.03 F_1 for o3).

Further analysis reveals, under ReAct, the models initially retrieved only precedents, but gradually began incorporating statutes into the retrieval process; after the 3rd retrieval, GPT-4.1 retrieves documents comprising 56% precedents and 44% statutes while o3 retrieves 66% precedents and 34% statutes.

This suggests that under single-pass RAG with a BM25 retriever, LLMs fail to access the relevant statutes needed for legal verification, resulting in low performance. However, with ReAct, the agent iteratively retrieves relevant statutes and incorporates the retrieved knowledge into the reasoning process, leading to higher recall (col 3, row 5–6). Precision remains similar to that of vanilla LLMs, indicating that the agentic approach also cannot significantly enhance the "Application" and "Conclusion" steps.

Interestingly, o3 does not exhibit significant improvements. This may be because o3 already incorporates inference-time reasoning strategies, reducing the marginal benefit of additional prompting or agentic structuring.

6 Conclusion

We introduce PLAT, a benchmark designed to evaluate LLMs' capability in taxation. Compared to previous studies, 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 reveal that while LLMs demonstrate some capability, vanilla models struggle to comprehensively understand taxation issues, particularly at the higher reasoning stages (Application/-Conclusion of IRAC). We also show that by gradually integrating retrieval and self-reasoning, these limitations can be partially mitigated, although reaching a correct conclusion remains challenging.

[25] Section 5.1.5

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, however, focuses specifically on evaluating the justifiability of exemption from additional tax penalties, serving as a case study in which LLMs must demonstrate a comprehensive understanding of complex situations rather than simply referencing statutes. A more holistic 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.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2024. Automated essay scoring using grammatical variety and errors with multi-task learning and item response theory. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 316–329, Mexico City, Mexico. Association for Computational Linguistics.

Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. 2025. Lexam: Benchmarking legal reasoning on 340 law exams. *Preprint*, arXiv:2505.12864.

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. De-

veloping 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.

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. 2025. Reasoning about uncertainty: Do reasoning models know when they don't know? *Preprint*, arXiv:2506.18183.

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.

Jung Ha Park. 2015. Study on the improvement of essay scoring system in leet. *법조*, (9):296 – 324.

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

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. 2025a. 'smolagents': a smol library to build great agentic systems. https://github.com/huggingface/smolagents.

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

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *Preprint*, arXiv:2506.06941.

André Smolentzov. 2013. Automated essay scoring: Scoring essays in swedish.

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.

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 Detailed statistics about PLAT

Detailed statistics about PLAT are presented in Table 9. PLAT consists of 36 easy cases and 64 hard cases.

B Example of PLAT

B.1 PLAT-MC

Examples from PLAT-MC are presented in Table 10.

B.2 PLAT-MC$_R$

Examples from PLAT-MC$_R$ are presented in Table 11.

B.3 PLAT-E

Essay questions are based on PLAT-MC with prefix that follows the format of the second-round essay-style exam for the Korean Certified Tax Accountant (CTA): "Read the following case and answer the question. Explain whether the imposition of additional tax (penalty tax) in the above case is lawful. (Note: Only the lawfulness of the imposition of additional tax is to be considered; no other issues are to be taken into account.)
<Case>"
Examples from PLAT-E rubrics are presented in Table 12.

C Dataset Annotation Procedure

For PLAT-MC, one tax expert annotated 100 precedents (hours). Other tax expert subsequently reviewed all annotated cases to ensure consistency and accuracy.

For PLAT-MC$_R$, one tax expert annotated 100 precedents (22 hours) and another tax expert reviewed all the annotated precedents along with the associated multiple-choice options, revising 8 of them. This collaborative review process enabled the construction of a high-quality dataset.

For PLAT-E, three tax experts curated IRAC-grounded scoring rubrics by consensus.

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 2a and Figure 2b.

D Prompt to Answer PLAT

D.1 Prompt to Answer PLAT-MC and PLAT-MC(w RAG)

Prompts to Answer PLAT-MC and PLAT-MC(w RAG) are presented in Figure 3. Korean is translated to English using GPT-4o.

D.2 Prompt to Answer PLAT-MC$_R$

Prompt to Answer PLAT-MC$_R$ is shown in Figure 4.

D.3 Prompt to Answer PLAT-E

Prompt to Answer PLAT-E is shown in Figure 5 and prompt for LLM-as-a-Judge scoring is in Figure 6. These prompts were designed based on the answer and grade prompts introduced by Fan et al. (2025).

E Additional Agent experiment

E.1 Multi-agent collaboration

To further investigate the legal agent methodology, we employ multi-agent collaboration for the LLM to respond to PLAT.

E.1.1 Experiments

For the purpose of our experiments, we constructed the Virtual-Court Tool. Virtual-Court Tool is a role-playing tool where the plaintiff's lawyer argues against the penalty tax, the defendant's lawyer argues for it, and the judge neutrally compares both arguments to reach a lawful conclusion (Figure 10)

E.1.2 Result

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. The result is shown at Table 6.

Both models, GPT-4.1 and o3, show a slight performance drop when combined with ReAct prompting and multi-agent collaboration (bottom row). For easy cases, both GPT-4.1 and o3 achieve higher precision with ReAct + multi-agent than with ReAct alone (col 5, row 3 vs 4); recall remains unchanged for GPT-4.1 and increases for o3 (col 6, row 3 vs 4). For hard cases, in contrast, both models show notably lower precision with ReAct + multi-agent (col 8, row 3 vs 4); recall is higher for GPT-4.1 and

Table 6: Agent scores on PLAT-MC (M stands for the experiment with multiple agents with different roles.)

Model	F1	P	R	F1-easy	P-easy	R-easy	F1-hard	P-hard	R-hard
GPT-4.1	0.65	0.66	0.64	0.75	0.76	0.74	0.49	0.42	0.64
o3	0.79$_{(\pm 0.03)}$	0.65$_{(\pm 0.04)}$	1.00$_{(\pm 0.00)}$	0.83$_{(\pm 0.02)}$	0.71$_{(\pm 0.03)}$	1$_{(\pm 0.00)}$	0.77$_{(\pm 0.05)}$	0.62$_{(\pm 0.08)}$	1$_{(\pm 0.00)}$
GPT-4.1 (RAG)	0.64	0.57	0.76	0.62	0.56	0.74	0.66	0.57	0.77
o3 (RAG)	0.75$_{(\pm 0.01)}$	0.64$_{(\pm 0.01)}$	0.92$_{(\pm 0.04)}$	0.85$_{(\pm 0.03)}$	0.80$_{(\pm 0.05)}$	0.91$_{(\pm 0.03)}$	0.69$_{(\pm 0.02)}$	0.55$_{(\pm 0.01)}$	0.93$_{(\pm 0.05)}$
GPT-4.1 (ReAct)	0.76	0.63	0.97	0.74	0.60	0.97	0.77	0.64	0.96
o3 (ReAct)	0.78$_{(\pm 0.02)}$	0.64$_{(\pm 0.02)}$	0.99$_{(\pm 0.01)}$	0.76$_{(\pm 0.08)}$	0.62$_{(\pm 0.10)}$	0.97$_{(\pm 0.02)}$	0.77$_{(\pm 0.04)}$	0.63$_{(\pm 0.05)}$	0.99$_{(\pm 0.01)}$
GPT-4.1 (ReAct+M)	0.75	0.61	0.99	0.81	0.70	0.97	0.71	0.56	0.99
o3 (ReAct+M)	0.73$_{(\pm 0.01)}$	0.59$_{(\pm 0.00)}$	0.96$_{(\pm 0.03)}$	0.80($_{(\pm 0.02)}$	0.67$_{(\pm 0.01)}$	1.00$_{(\pm 0.00)}$	0.71$_{(\pm 0.01)}$	0.56$_{(\pm 0.01)}$	0.98$_{(\pm 0.01)}$

remains similar for o3 (col 9, row 3 vs 4). Overall, lower precision in hard cases drives a decrease in F1.

Analysis We further analyzed how the number of correct, incorrect, and "don't know" responses changed for individual categories (Table 9).

GPT-4.1 improved in most easy categories, especially in `easy-(C)` and `easy-(D)` (Table 7). Similarly, o3 improved in most easy categories, however, unlike GPT-4.1, it degraded in `easy-(C)` (Table 8).

For hard cases, both GPT-4.1 and o3 improved in `hard-(I)`, whereas GPT-4.1 degraded in `hard-(F)` and `hard-(G)` (Table 7). Similarly, o3 also degraded in `hard-(F)` (Table 8).

Although we conducted a qualitative analysis as above, we did not observe clear patterns. Building successful Multi-agent ReAct agents remains as a future study.

Table 7: Changes in model answers across categories (GPT-4.1 ReAct to ReAct+M).

Categories	Δ Correct	Δ Incorrect	Δ "Don't know"
easy-(A)	−1	+1	0
easy-(B)	+1	−1	0
easy-(C)	**+2**	−1	−1
easy-(D)	**+2**	−2	0
easy-(E)	+1	−1	0
hard-(F)	−1	**+2**	−1
hard-(G)	−2	**+3**	−1
hard-(H)	**+3**	−2	−1
hard-(I)	**+2**	0	−2

E.2 Legal Think Tools

To further explore the legal agent methodology, we built additional tools that are employed during the agent experiments. They are commonly used in Agent-Based Approach setting.

- **Bridge between "Rule" and "Application"** *Document-Relevance-Analyzer Tool*: summarizes statutes and precedents, classifies relevant vs. irrelevant parts, and compares them with the issue (Figure 8).

Table 8: Changes in model answers across categories (o3 ReAct to ReAct+M).

Categories	Δ Correct	Δ Incorrect	Δ Don't know
easy-(A)	0	0	0
easy-(B)	+1	−1	0
easy-(C)	−2	**+2**	0
easy-(D)	+1	−1	0
easy-(E)	+1	−1	0
hard-(F)	−3	**+3**	0
hard-(G)	−1	0	1
hard-(H)	+1	0	−1
hard-(I)	**+2**	−2	0

- **Precursor to IRAC structured reasoning** *Legal-Analyzer Tool*: implements a structured IRAC analysis; identifies the key issue, states the governing rules, applies them to the facts, and concludes with the final legal determination (Figure 9).

F LLM-as-a-judge for essay grader

To refine LLM grader, we conduct rubric-level analyses comparing human graders' scores with `o3`'s scores. Our observations are as follows:

- Compared to human graders, the LLM-based grader consistently assigns more generous scores across rubrics.
- Writer LLMs, including the state-of-the-art reasoning model `o3`, frequently hallucinate in the Rule (R) component of IRAC.

F.1 Generosity Bias of LLM Graders

Figure 11 presents scatter plots comparing the LLM-based grader (o3) with each of the three human graders. Across all graders, the LLM's scores exhibit a positive correlation with human judgments, although the strength of correlation varies substantially: Pearson's r ranges from 0.44 with human1 to 0.76 with human3. The fitted regression

lines show that o3 generally gives higher scores than the human graders, as reflected in slopes less than one. This suggests that, while the LLM captures grading trends consistent with human judgment, it consistently produces more generous assessments.

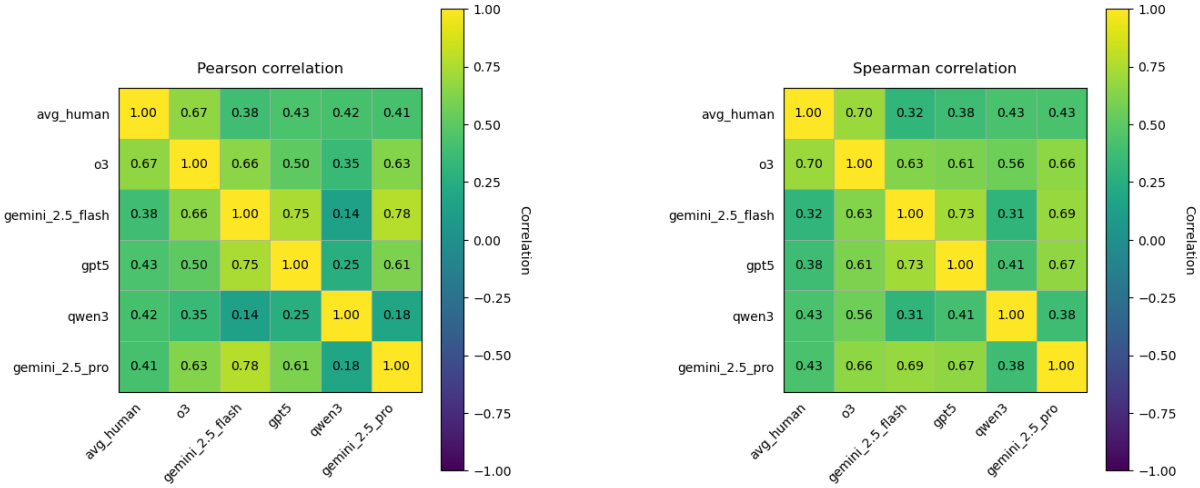
F.2 Hallucination in the Rule (R) Component

Human graders assigned almost zero points across all R-based rubrics for the o3 responses. A qualitative analysis revealed that these responses exhibited severe hallucinations, consistent with the issues discussed in Sec. 5.1.5.

Dahl et al. (2024) shows that while more recent large language models demonstrate reduced hallucination rates when evaluated on legal texts and case law, the problem remains significant. Building on this insight, our study pursued more precise scoring by excluding the four rubrics classified under R. (No 2, 4, 6, 8 rubrics. You can check them in Table 12)

G Resources and Licensing

Datasets and models used in this work, with licenses are presented in Table 13. Additionally, PLAT was preprocessed to anonymize all personally identifiable information (PII), despite the dataset's origin in court precedents.



(a) Pearson correlation

(b) Spearman correlation

Figure 1: Pearson and Spearman correlations between candidate LLMs and expert average scores.

Table 9: Penalty Tax Justifiability Topics by Difficulty

Difficulty	Topic	Count
easy	(A) Clerical errors or omissions that do not substantially affect the underlying tax amount are not subject to penalty taxes.	1
easy	(B) When the tax authority issued an incorrect tax disposition that misled the taxpayer, a penalty on the delayed base tax is considered unlawful.	3
easy	(C) Mere misunderstanding or ignorance of the law does not constitute a justifiable reason.	17
easy	(D) Claiming ignorance of facts that the taxpayer could have reasonably known is not accepted as a justifiable reason.	11
easy	(E) 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.	4
hard	(F) Cases where the taxpayer faced unavoidable circumstances that hindered payment.	22
hard	(G) Cases requiring proper assessment of whether differences among tax authorities indicate genuine divergence in interpretation or whether the tax law itself was ambiguous.	18
hard	(H) Cases requiring assessment of whether the taxpayer, despite delayed payment, promptly fulfilled obligations–or neglected duties and failed to exercise due care.	12
hard	(I) Cases dealing with whether an official interpretation (e.g., from a tax officer or written inquiry response) qualifies as a public opinion.	12

Table 10: 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 120 billion out of KRW 145,381,546,613 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 "when 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 11: Example of PLAT-MC$_R$ (The corresponding questions are given in Table 10.)

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.

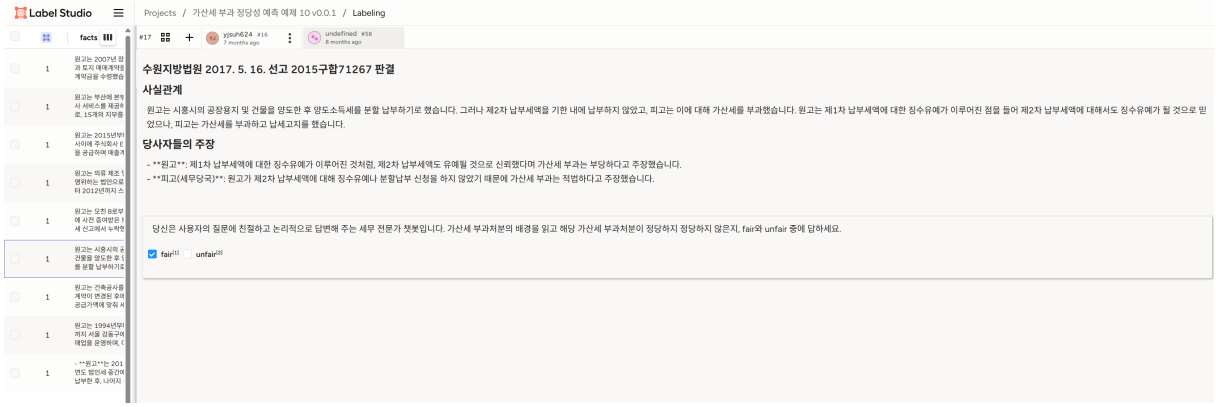
A	Opened a charging station and opened five deposit accounts (hereinafter 'the accounts in question') at Nonghyup Bank, but did not report the opening of business accounts to the competent tax office by June 30, 2015. Therefore, the imposition of the penalty tax is illegal.
B	The imposition of a penalty tax is lawful in special circumstances where the taxpayer's unawareness of their obligation can be considered not unreasonable.
C	The imposition of the additional tax is lawful. The 'Not Applicable' on the notice merely signifies that it is not subject to the additional tax for failure to open a business account, and does not imply that the account has already been reported. Therefore, the plaintiff's misunderstanding cannot constitute a valid reason.
D	The plaintiff actually used each of the deposit accounts in question as a business account and, when filing the comprehensive income tax return, recorded all year-end balances of the accounts in the sincere filing confirmation form. Although the defendant could have ascertained the account details through this, imposing a penalty tax merely on the formal ground of non-reporting runs counter to the substantive taxation principle of tax law. Therefore, the imposition of the penalty tax is illegal.

Table 12: PLAT-E IRAC classification of PLAT-E rubrics

No	Rubric Items	IRAC element
1	Nature and Definition of Additional Tax (1 point): Did the answer clearly describe the legal nature of additional tax (a sanction for violation of tax law obligations)?	I
2	Citation of Relevant Laws (1 point): Did the answer accurately mention the relevant statutes such as the Framework Act on National Taxes, and correctly cite article numbers and provisions?	R
3	Requirements and Scope of Imposing Additional Tax (1 point): Did the answer accurately explain the requirements for imposing additional tax and its scope of application in relation to the problem scenario?	I
4	Legal Basis for Exemption due to "Justifiable Reason" (1 point): Did the answer correctly present the legal basis for exemption from additional tax (Article 48 of the Framework Act on National Taxes)?	R
5	Applicability of "Ignorance of Law," etc. (1 point): Did the answer clearly state that ignorance or misunderstanding of the law, or the lack of awareness by an expert, does not constitute a "justifiable reason"?	I
6	Accuracy of Relevant Case Law (1 point): Was the cited case law an actual, existing case, and was its holding accurately identified?	R
7	Appropriateness of Case Holding Description (1 point): Was the holding of the cited case clearly articulated and organically connected to the issue, contributing to the logical development of the argument?	A
8	Accuracy of Case Number and Source (1 point): Was the case number (e.g., Supreme Court Decision 20XXDuXXXX) accurately indicated?	R
9	Identification of Core Issues and Introduction (1 point): Did the introduction clearly identify the core issue(s) in the problem, and was the overall logical structure systematic?	A
10	Analysis of Facts in Relation to Legal Basis (1 point): Did the answer analyze the facts presented in the problem (e.g., plaintiff's claims) in light of the relevant legal provisions, and was the conclusion logically derived?	C

Name	License	URL
EXAONE3.5-32B (Research et al., 2024)	MIT License	https://huggingface.co/LGAI-EXAONE/EXAONE-3.5-32B-Instruct
Qwen3-32B (Yang et al., 2025)	CC BY 4.0	https://huggingface.co/Qwen/Qwen3-32B
smolagents (Roucher et al., 2025b)	MIT License	https://github.com/huggingface/smolagents
PLAT(ours)	CC BY-NC	Dataset will be released on GitHub upon publication.

Table 13: Datasets and models used in this work, with licenses.



(a) Label Studio interface for PLAT-MC.

(b) Label Studio interface for PLAT-MC$_R$.

Figure 2: Label Studio interfaces for our datasets. Each subfigure shows annotation setup for (a) PLAT-MC and (b) PLAT-MC$_R$.

System Prompt

You are a tax expert chatbot that provides friendly and logical answers to users' questions.

Vanilla User Prompt

Please read the background and materials related to the imposed penalty tax presented above.
Determine whether the penalty tax is "Legitimate", "Not legitimate", or, if a clear conclusion cannot be reached, answer "don't know". Then, provide an explanation for your choice.
Problem description: {precedent}

LLM with RAG User Prompt

Please read the background and materials related to the imposed penalty tax presented above.
Determine whether the penalty tax is "Legitimate", "Not legitimate", or, if a clear conclusion cannot be reached, answer "Don't know". Then, provide an explanation for your choice.
Problem description: {precedent}
Reference material: {raged doc}

###Answer:

Figure 3: Prompts to Answer PLAT-MC: Vanilla vs. with RAG

System Prompt
You are a tax expert chatbot that provides friendly and logical answers to users' questions.

User Prompt

The following is a multiple-choice question on the imposition of additional tax. Read the case and each option carefully, then choose only one option (A, B, C, or D) that is most correct regarding the imposition of the additional tax. You must respond in the format `Answer: A` and provide an explanation.
Question: {question}
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
Answer:

Figure 4: Prompt to Answer PLAT-MC$_R$.

System Prompt
You are a tax expert chatbot that provides friendly and logical answers to users' questions.

User Prompt
You are an expert in tax law and must address legal issues in a systematic, exam-style manner. Unless otherwise specified, assume that South Korean tax law applies. However, where contextually justified, you should also address legal issues beyond South Korean tax law. Use precise legal terminology, and always use honorifics in your responses. Do not mention disclaimers or the need for external legal advice. Do not ask the user to look up the law themselves. Provide focused legal analysis and individualized advice. Your answers should be decisive and authoritative; do not state that it is merely general information. You must include Korea-specific legal terminology. If you identify relevant legal considerations, present a concise and clear legal analysis. You must always cite specific statutory provisions, specifying Article (조), Paragraph (항), Subparagraph (호), and Item (목). Example: "Article 6, Paragraph 1 of the Framework Act on National Taxes." Using only general references (e.g., "Framework Act on National Taxes") is not permitted. If there is no relevant information, you must explicitly state that none exists. If reliable sources are available, share substantive guidance or insights. The response must be in the same language as the question. If the question requests a "simple answer," you must provide a concise answer. If you are asked to analyze a specific case from the certified tax accountant examination, but the case text or details are not provided, you must explicitly point out that the essential case material is missing.

Read the following case and answer the question. Explain whether the imposition of additional tax (penalty tax) in the above case is lawful.
(*Note: Only the lawfulness of the imposition of additional tax is to be considered; no other issues are to be taken into account.*)

<Case> {precedent}

Figure 5: Prompt to Answer PLAT-E.

User Prompt

You are a grader objectively and officially evaluating answers for a Korean Tax Law exam. You are an expert in tax law, and you must handle legal issues systematically and in an exam style.

Goal: Your task is to determine how closely the candidate's answer aligns with the model answer provided. Refer to the model answer to evaluate accuracy, completeness, and legal reasoning.

Context: You will be given both a candidate's answer and a model answer.

Output Format:
After reviewing the answer:
Explanation: Briefly explain how the candidate's answer aligns with or deviates from the model answer.
Scoring: The score must reflect how well the candidate's answer fulfills the requirements of the model answer, using the following strict format:
Example: "Comprehensive Score: [[5]]"
Comprehensive Score: Assign a final score between 0 and 10 (whole numbers only).
- 10 = fully satisfactory (100%)
- Lower scores proportionally reflect deficiencies (e.g., 5 = 50% fulfilled).
- If the candidate includes material not in the model answer, you must generally deduct points unless you are certain it is legally accurate and relevant.
- Assume the model answer contains all information necessary for a complete response.
- If the model answer cites books or academic articles, candidates do not need to reproduce them.
- However, statutory provisions must be cited accurately with precise identification of Article, Paragraph, Subparagraph, Item, etc.

Reference: {Grading Rubrics}

Figure 6: Prompt for LLM-as-a-Judge Scoring

System Prompt

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.

User Prompt

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: `{retrieved doc}`

Answer:

Figure 7: Prompt for Agentic RAG (The default prompt from ToolCallAgent of `smolagents` library is used.)

User Prompt

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.

Figure 8: Prompt for Document-Relevance-Analyzer tool

User 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>

Figure 9: Prompt for Legal-Analyzer tool

User Prompt (Plaintiff's lawyer role-playing tool)

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.

User Prompt (Defendant's lawyer role in role-playing tool)

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.

User Prompt (Judge role in role-playing tool)

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.

Figure 10: Prompt for Virtual-court tool

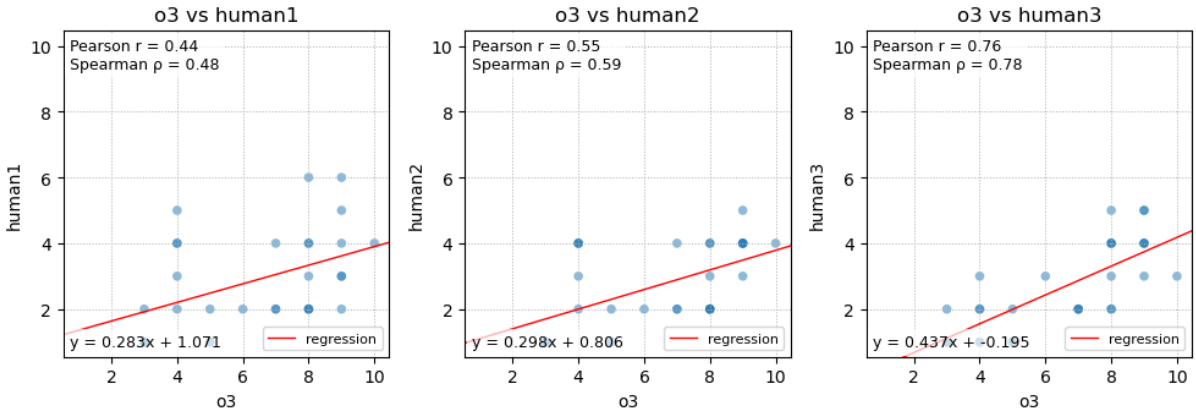


Figure 11: Scatter plots comparing the scores assigned by the LLM-based grader (o3) with those of three human graders. Each panel shows the regression line (red) with the corresponding Pearson's r, Spearman's ρ, regression equation and . While o3 scores are positively correlated with all human graders, the strength of correlation varies across graders (r = 0.44, 0.55, and 0.76, respectively).