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 is constructed to evaluate LLMs' understanding of tax law, particularly in cases where resolving the issue requires more than just applying related statutes. Our experiments with six LLMs reveal that their baseline capabilities are limited, especially when dealing with conflicting issues that demand a comprehensive understanding. However, we found that enabling retrieval, self-reasoning, and discussion among multiple agents with specific role assignments, this limitation can be mitigated.

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] that comprises of 50 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.

Our experiments with two open-source LLMs (Qwen2.5 (Qwen Team, 2024), Exaone (Research et al., 2024)), and four commercial LLMs (GPT-o1-mini, 4o, o1 (OpenAI, 2023, 2024), and Claude (Anthropic, 2023)) reveal, the strongest commercial model can achieve 75% F_1 score in PLAT. A detailed analysis reveals, while LLMs perform well on relatively simple problem, their accuracy declines when a comprehensive understanding is required.

To address this issue, we examine how LLM performance changes when enabling (1) retrieval augmentation, (2) self-reasoning, and (3) multi-agents

[1] PREDICTING THE LEGITIMACY OF PUNITIVE ADDITIONAL TAX

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

collaboration with specified roles. The resulting LLM-based agent achieves up to +11% in F_1.

In summary, out contributions are

- We propose a new dataset, PLAT, to assess LLMs' understanding of tax law specialized in cases that cannot be resolved solely based on statutes.
- We evaluate six LLMs and find that while they demonstrate some capability, their vanilla performance is limited in comprehensively understanding legal cases.
- We show that integrating agent functionality into LLMs can mitigates these limitations.

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 (IRC). 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.

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].

[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.

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 consists of 50 questions–25 justifiable, 25 not justifiable cases–made from Korean precedents handling the issue regarding the legitimacy of the additional tax penalty.

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 returend approximately 10k 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 3k candidate pools.

To extract facts and claims from precedents, we used GPT-o1 (`o1-2024-12-17`). 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 process until we compiled a final dataset of 50 examples, with an equal split: 25 cases where the court ruled the exemption from penalty tax was, and 25 cases where the court decided that the exemption was not justified. Each example required approximately 30–40 minutes for evaluation, resulting in total 25–33 hours of expert review time.

[4] `lbox.kr`

4 Experiments

We use two agentic frameworks: AutoGen (Wu et al., 2024), and `smolagent` (Roucher et al., 2025) along with following language models: `Qwen/Qwen2.5-32B-Instruct`, `LGAI-EXAONE/EXAONE-3.5-32B-Instruct`, `gpt-o1-mini-2024-09-12`, `gpt-4o-2024-08-06`, `o1-2024-12-17`, and `claude-3-5-sonnet-20241022`.

For all experiments, we set the temperature to 0.3, as intial tests with 0.0 and 1.0 resulted in degraded performance. Fro retrieval-based experiment, we use Pyserini (Lin et al., 2021) with the BM25 algorithm with default hyperparameters. Each retrieval is limited to five documents, as initial tests three of ten documents resulted in lower performance.

During evaluation, the model generates an answer among three possible choices: the penalty tax is legitimate, the penalty tax is not legitimate, uncertain. The model must also provide rational for its response. To assess performance, we compute precision, recall, and 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 LLMs' scores on PLAT

In PLAT, a model needs to decide whether an additional penalty tax is legitimate, based on given facts and claims from both the plaintiff (taxpayer) and the defendant (tax authority) (Table 3 in Appendix). The model is also permitted to refuse to answer. We evaluate six LLMs on PLAT (Table 1). The results show that while the two open-source LLMs–Qwen and Exaone–show comparable performance to lower-end commercial LLMs (row 1, 2, and 3), flagship commercial models achieve up to 0.75 F_1 scores. Interestingly, both open-source models exhibit low recall, suggesting they frequently refuse to make a decision.

5.2 LLMs' Limitation in Understanding Tax Cases Comprehensively

To gain insight into what LLMs are (not) capable of, we manually analyzed cases where either at least three LLMs answered correctly or at least three

Table 1: Accuracy of vanilla LLMs on PLAT.

Model	F1	P	R
Alibaba Qwen-2.5-32B	0.55	0.79	0.42
LG Exaone 3.5-32B	0.61	0.69	0.55
GPT-o1-mini	0.63	0.49	0.88
GPT-o1	0.67	0.61	0.74
GPT-4o	0.72	0.59	0.94
Claude-3.5-sonnet	0.75	0.71	0.79

LLMs answered incorrectly. In these cases, LLMs were able to recognize the following principles:

- Ignorance or misunderstanding of tax laws by a taxpayer does not constitute a justifiable reason.[5]
- 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).[6]
- Uncertainty due to differing opinions or conflicting views between the Board of Audit and tax authorities can constitute a justifiable reason.[7]

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.[8]
- When judges considered various taxplayer-specific circumstances, including the feasibility of fulfilling obligations, LLMs strictly adheres to principles and rules.[9]

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

5.3 Agent-Based Approach for Enhancing LLMs' Understanding of Tax Cases

To address the limitations identified above, we gradually introduce additional functionalities, including retrieval augmentation, self-reasoning with memory, and multi-agent collaboration. Retrieval augmentation may allow LLMs to search for relevant cases and legal articles, improving decision-making, self-reasoning with memory enables LLMs to track prior reasoning, making more consistent judgments, multi-agent collaboration assigns three LLMs as taxpayer, tax authority, and judge, encouraging each agent first focuses on local problem and then gradually extend the scope to the whole problem.

Table 2: Accuracy of LLM-baed agents on PLAT. aRAG refers to "agentic-RAG", while `roles` denotes a multi-agent setup with distinct role assignments.

Model	F1	F1 (easy)	F1 (hard)
GPT-4o	0.72 $\pm$0.023	0.90 $\pm$0.046	0.56 $\pm$0.006
GPT-4o + RAG	0.78 $\pm$0.002	0.79 $\pm$0.023	0.77 $\pm$0.015
GPT-4o + aRAG	0.83 $\pm$0.027	0.76 $\pm$0.016	0.88 $\pm$0.043
GPT-4o + `roles`	0.83 $\pm$0.019	0.86 $\pm$0.015	0.82 $\pm$0.024
GPT-4o + aRAG + roles	0.72 $\pm$0.025	0.59 $\pm$0.012	0.79 $\pm$0.025

Indeed, we found adding RAG results in +6% F_1 (Table 2, row 2, col 2), adding reasoning capability with retrieval tool +11% F_1 (agentic RAG with REACT framework(Yao et al., 2023b), row 3, col 2), multi-agents with specific roles results in +11% F_1 (row 4, col 2). However, when we combine all functionality no improvement observed (row 5, col 2, +0% F_1).

Based on the analysis in previous section, we manually categorize 50 examples into 21 "easy" cases and 29 "hard" cases. The results shows while use of external tools somehow reduces F_1 on "easy" cases, they improve performance on "hard" cases. Further analysis is ongoing.

6 Conclusion

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

[5] Daegu District Court 2015Guhap877
[6] Seoul Administrative Court 2016Guhap56936
[7] Seoul Administrative Court 2010Guhap32402
[8] Busan High Court 2016Nu11, Seoul High Court 2020Nu43946
[9] Daegu District Court 2018Guhap20506

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

Anthropic. 2023. Introducing claude. https://www.anthropic.com/index/introducing-claude.

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

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*.

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.

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.

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. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

LG AI Research, :, Soyoung An, Kyunghoon Bae, Eunbi Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Yeonjung Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Euisoon Kim, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Moontae Lee, Seungjun Lee, Woohyung Lim, Sangha Park, Sooyoun Park, Yongmin Park, Boseong Seo, Sihoon Yang, Heuiyeen Yeen, Kyungjae Yoo, and Hyeongu Yun. 2024. Exaone 3.0 7.8b instruction tuned language model. *Preprint*, arXiv:2408.03541.

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*.

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)*.

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

A Example

A.1 PLAT

Table 3: An example from PLAT.

Facts	Claim from Plaintiff (Taxpayer)	Claim from Defendant (Tax Authority)	Label
1. Plaintiffs' Family Relations: The plaintiffs, Yu CC and Yu DD, are siblings, and Kwon EE is their mother. 2. Ownership Status of the Building: - The building located in GGG-dong, FFF-gu, Seoul (hereinafter 'the Building in Question') is divided into multiple units. - Yu DD owns a portion of the building and the land. - Kwon EE previously owned a portion of the building and the land but donated it to the plaintiffs and others on 0000-00-00. - The plaintiffs and others completed the ownership transfer registration on 0000-00-00. 3. Lease Agreements: - Yu DD and Kwon EE entered into lease agreements with tenants. - Yu DD was granted full authority over leasing matters by Kwon EE, allowing him to enter into lease agreements and collect rent. 4. Rent Collection and Legal Disputes: - After receiving the donation, the plaintiffs requested new lease agreements from the tenants, but they refused. - The plaintiffs filed a lawsuit against the tenants for the return of unjust enrichment but lost the case. - Yu DD filed a lawsuit against the plaintiffs, demanding the removal of the building and the return of the land. - The plaintiffs, in response, filed a counterclaim to confirm their share of rental income. - In the appellate court, a settlement was reached on 0000-00-00, dividing rental income as follows: - Yu DD: 60- Plaintiffs and others: 40 5. Amended Income Tax Return and Penalty Tax Imposition: - The plaintiffs filed an amended income tax return for rental income from 0000 to 0000 and paid the corresponding tax. - However, on 0000-00-00, the defendant (tax authority) imposed a penalty tax, claiming that the plaintiffs had failed to pay the additional penalty tax. - The plaintiff's sibling, B (the decedent), passed away on March 24, 2018. - The plaintiff inherited the land specified in Appendix 1 (hereinafter 'the Land in Question') from the decedent. - On March 31, 2019, the plaintiff assessed the officially announced land price at 591,474,900 KRW and reported and paid inheritance tax based on this valuation. - The defendant (tax authority) later confirmed that the decedent had purchased the Land in Question within two years before the inheritance start date. - After a review by the valuation review committee, the inheritance tax value of the Land in Question was reassessed at 1,899,900,000 KRW. - On April 16, 2020, the defendant reassessed and notified the plaintiff of an inheritance tax adjustment for March 24, 2018, amounting to 797,054,920 KRW (including a late payment penalty of 68,825,680 KRW). - Of this amount, 62,313,740 KRW in late payment penalties related to the Land in Question is the subject of this dispute. - The plaintiff filed an appeal on July 21, 2020, but it was dismissed on December 8, 2020.	1. Plaintiff's Claim - Claim: The plaintiffs argue that they had a justifiable reason for failing to meet their tax reporting and payment obligations on time. - Basis: - Until the rental income rights regarding the Building in Question were legally confirmed through litigation, they could not determine their exact share or amount of rental income. - Given the unresolved legal status of the lease agreements with tenants and the rental income distribution ratio with Yu DD, fulfilling their tax reporting and payment obligations was either impossible or extremely difficult. - Therefore, the imposition of the penalty tax is unjust.	2. Defendant's Claim - Claim: The defendant asserts that the penalty tax imposition is lawful, as the plaintiffs had no justifiable reason for failing to fulfill their tax obligations. - Basis: - By accepting the donation of the Building in Question from Kwon EE, the plaintiffs inherited all rights and obligations as landlords. - They could have calculated their share of rental income and met their tax reporting and payment obligations on time. - Yu DD had already submitted a tax authority report specifying the rental income distribution ratio as 60:40 between himself and the plaintiffs. - Thus, the plaintiffs had no valid justification for failing to comply with their tax obligations.	Not legitimate.

A.2 Prompt for Vanilla LLM and RAG

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

System Prompt	Input
You are a tax expert chatbot that provides friendly and logical answers to users' questions.	Based on the background provided regarding the imposition of the penalty tax, please determine whether the imposition of the penalty tax is "lawful", "unlawful", or "unknown" if a conclusion cannot be reached. Provide an explanation for your answer.: ... precedent

A.3 Prompt for multi-agents

Table 5: Prompt with plaintiff role.

System Prompt	Input
You are a tax expert chatbot that provides friendly and logical answers to users' questions. You are a tax attorney who provides friendly and logical answers to users' questions. You always argue that the imposition of penalty taxes is not lawful. You reach conclusions step by step with clear reasoning and rational justification.	Based on the background provided regarding the imposition of the penalty tax, please determine whether the imposition of the penalty tax is "lawful", "unlawful", or "unknown" if a conclusion cannot be reached. Provide an explanation for your answer.: ... precedent

A.4 Agentic RAG

Table 6: Prompt with defendant role.

System Prompt	Input
You are a tax expert chatbot that provides friendly and logical answers to users' questions. You are a tax attorney who provides friendly and logical answers to users' questions. You always argue that the imposition of penalty taxes is lawful. You reach conclusions step by step with clear reasoning and rational justification.	Based on the background provided regarding the imposition of the penalty tax, please determine whether the imposition of the penalty tax is "lawful", "unlawful", or "unknown" if a conclusion cannot be reached. Provide an explanation for your answer.: ... precedent

Table 7: Prompt with judge role.

System Prompt	Input
You are a tax expert chatbot that provides friendly and logical answers to users' questions. You are a tax judge who provides friendly and logical answers to users' questions. You critically analyze the imposition of penalty taxes and make sharp and precise judgments. Among the given two arguments, you always select the most accurate and correct one, explaining your reasoning in detail.	Based on the background provided regarding the imposition of the penalty tax, please determine whether the imposition of the penalty tax is "lawful", "unlawful", or "unknown" if a conclusion cannot be reached. Provide an explanation for your answer.: ... precedent

Table 8: Agent RAG. The default prompt from ToolCallAgent of `smolagent` libary is used.

Input
You are a tax judge who provides friendly and logical answers to users' questions. You critically analyze the imposition of penalty taxes and make sharp and precise judgments. You effectively utilize the given materials to make accurate and well-reasoned decisions as a tax judge. Based on the background provided regarding the imposition of the penalty tax, please determine whether the imposition of the penalty tax is "lawful", "unlawful", or "unknown" if a conclusion cannot be reached. Provide an explanation for your answer.: ... precedent