Aligning LLMs for Thai Legal Question Answering with Efficient Semantic-Similarity Rewards

Pawitsapak Akarajaradwong[1,3], Chompakorn Chaksangchaichot[1,4], Pirat Pothavorn[1], Attapol Thamrongrattanarit-Rutherford[3], Ekapol Chuangsuwanich[4], Sarana Nutanong[1,2]
[1]VISAI AI [2]Vidyasirimedhi Institute of Science and Technology
[3]Department of Linguistics, Faculty of Arts, Chulalongkorn University
[4]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University
{pawitsapaka_visai,chompakornc_pro,piratp_visai,sarana.n}@vistec.ac.th
{attapol.t,ekapol.c}@chula.ac.th

Abstract

The Retrieval-Augmented Generation (RAG) systems' performance on Thai legal question answering is still limited, especially for questions requiring extensive, complex legal reasoning. To address these limitations, we introduce a resource-efficient approach that aligns Large Language Models (LLMs) for improved citation accuracy and response quality using Group-Relative Policy Optimization (GRPO). Our proposed method leverages BGE-M3 embeddings as a cost-efficient semantic-similarity reward, significantly reducing computational expenses up to 2.5× compared to an LLM-based reward model. Experiments on the NitiBench benchmark demonstrate substantial improvements: GRPO achieves up to 90% citation-F1 gains relative to the base model and a 31% increase in joint quality metrics over instruction tuning. Crucially, our approach provides a practical and effective solution for enhancing legal LLMs in resource-constrained environments.

1 Introduction

The ability to deliver accurate and grounded answers with relevant law citations is essential for reliable legal question answering. Most legal-domain LLM solutions (Corporation, 2025; LexisNexis, 2023; Takyar, 2024; Viriyayudhakorn, 2024) adopt Retrieval-Augmented Generation (RAG) to reduce hallucinations by attaching retrieved legal documents as supporting context. However, retrieved documents are not always fully leveraged and can contain false positives (Akarajaradwong et al., 2025; Magesh et al., 2024). A common approach to mitigate this issue is to require LLMs to emit explicit citations during generation. Yet, Akarajaradwong et al. (2025) show that even when golden contexts are provided, strong proprietary models often fail to cite all relevant law sections at generation time. These findings highlight a key limitation of current LLMs in producing factually grounded responses, thereby undermining the reliability of downstream legal applications. Ensuring accurate and well-cited responses thus remains a central open challenge in the legal domain.

While instruction tuning can partially address this gap, it provides limited control over citation behavior since its objective is to maximize next-token likelihood rather than citation accuracy. This highlights the need for more targeted alignment that not only improves factuality but also enforces verifiable citation standards. To enable fine-grained control over citation accuracy, we frame Thai legal QA as a citation-sensitive alignment challenge. Building on recent advances in Reinforcement Learning with Verifiable Rewards (RLVR) (DeepSeek-AI et al., 2025) and Group-Relative Policy Optimization (GRPO) (Shao et al., 2024), we treat citation accuracy and response quality as a verifiable reward and directly align LLMs toward higher citation quality during alignment tuning. Although computing rewards for citation is straightforward and inexpensive, evaluating response quality is far more costly, as it typically requires an LLM-based reward model during training, leading to significant computational overhead and higher alignment costs. This raises our central research question: *How can we affordably and effectively align LLMs for citation-sensitive legal QA in resource-constrained settings such as Thai law?*

To address this, we investigate the following research questions:

- **(RQ1) Reward Strategies:** What are the trade-offs between an LLM-based reward model compared with a low-cost semantic reward proxy for modeling response quality rewards?
- **(RQ2) Thai-CPT vs. Language-Generic:** Does Thai-specific continued pretraining enhance the effectiveness of alignment strategies?
- **(RQ3) RLVR vs. Instruction Tuning:** How can RLVR enhance response quality and citation accuracy compared to instruction tuning?

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 304–316

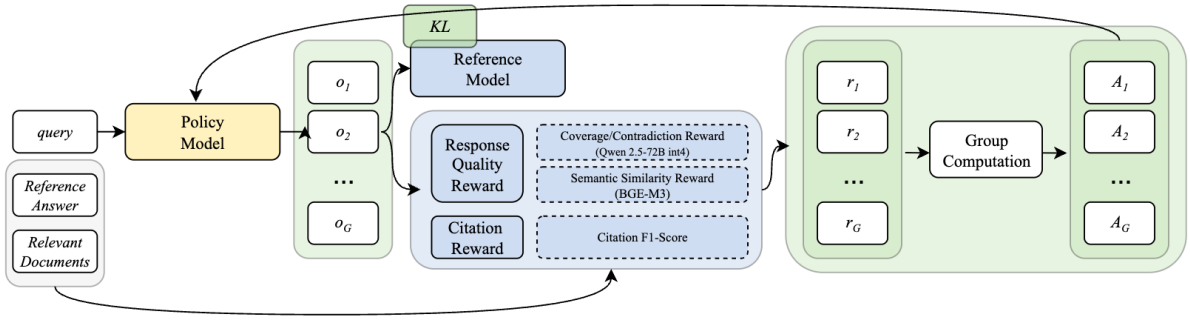

Figure 1: Demonstration of our proposed method. Here, we use GRPO objectives with specialized reward to align LLM towards better citation and response using **Response Quality Reward** and **Citation Accuracy Reward** (§3.1).

Our experiments on the NitiBench benchmark (Akarajaradwong et al., 2025) yield three key contributions. First, we propose a low-cost semantic reward proxy to compute response quality reward using general embedding models, achieving over 2.5× improvement in training efficiency while maintaining performance comparable to an expensive LLM-based reward model. Second, we show that Thai-CPT models are more receptive to alignment signals than their general-purpose counterparts. Finally, we demonstrate that GRPO consistently outperforms instruction tuning in improving citation fidelity and response quality, particularly for in-domain tasks. *Together, these findings provide a recipe for building specialized legal LLMs that achieve efficient and accurate law grounding under resource constraints.*

2 Background

Thai Legal System Thailand operates under a civil law system, characterized by a strict hierarchy of written laws. At the apex is the Constitution, followed by Organic Laws, Acts, Codes, and various subordinate legislation like Royal Decrees and Ministerial Regulations. This hierarchy mandates that lower-level laws must not contradict higher ones, creating a highly structured but complex legal corpus.

These documents are meticulously organized into a multi-tiered structure, often including divisions such as Book, Title, Chapter, Division, and ultimately, the **Section**. A Section is the fundamental unit of law, articulating a specific rule, right, or obligation. Accordingly, this work focuses on evaluating legal reasoning and citation performance at the Section level, treating it as the primary unit for retrieval and grounding.

Enhancing LLM legal citation performance A growing body of work seeks to make LLMs produce verifiable citations. CitaLaw (Zhang et al., 2025) adapts the ALCE benchmark (Gao et al., 2023) to the legal domain, introduces a syllogism-level citation metric, and supports both statutes and precedent cases. ALCE itself evaluates statement-level grounding using an NLI verifier, requiring every generated claim to be backed by retrieved evidence. Shareghi et al. (2024) compares citation accuracy across three retrieval regimes: 1) retriever-only, 2) LLM query-rewrite, and 3) hybrid method. This work focuses on Australian case-law and shows that task-specific instruction tuning yields the largest gains in improving citation accuracy. LegalBench-RAG (Pipitone and Alami, 2024) isolates the retriever's contribution by measuring precision over expert-annotated snippets while varying chunking and top-k, revealing a retrieval-quality ceiling on downstream citation F1.

Usage of embedding-based reward models Early works explored leveraging pretrained embeddings as reward signals for text generation alignment. Yasui et al. (2019) finetune BERT (Devlin et al., 2019) on Semantic Textual Similarity (STS) and employ the tuned model as a REINFORCE reward for machine translation. Kumar and Subramaniam (2019) optimize an abstractive summarizer directly for BERTScore (Zhang et al., 2020), observing higher fluency and lower redundancy than ROUGE-reward baselines. More recently, Sun et al. (2025) distil preference scores from the "gold" reward model of Dong et al. (2023, 2024) into lightweight proxies, an MLP and a LightGBM, that take paired Gemma-2B embeddings as input, achieving judge-level quality. These results indi-

cate that inexpensive embedding-based rewards can rival heavyweight LLM judges for preference-optimized generation. However, their integration into modern preference-optimization algorithms remains under-explored.

Research Gap Existing legal QA systems rely primarily on instruction tuning, with limited success on citation grounding. While GRPO and efficient reward proxies have shown promise elsewhere, their application to legal domains, particularly Thai legal QA, remains underexplored. We address this gap by investigating GRPO's impact on citation accuracy and evaluating cost-effective reward strategies.

3 Proposed Studies

We frame the Thai legal question answering (QA) task as a citation-sensitive generative task, where the model must generate a free-form response and citations based on a user query and a set of retrieved legal documents. The response must be semantically informative, cite relevant statutes, and avoid hallucinations that reference unsupported claims.

To align LLMs with these objectives, we proposed a framework based on GRPO (Shao et al., 2024), which aligns model behavior through a carefully designed reward function that encourages response quality and citation accuracy. Our proposed method is summarized in Figure 1. This reward-based formulation enables us to study how different alignment strategies influence legal response quality under the constraints of low-resource legal domains.

3.1 Reward Strategies (RQ1)

Response Quality Reward We design a reward to ensure that the quality of the response is acceptable, given the reference answer from the ground truth. Here, we provide two setups for evaluating generated response quality.

First, using a strong LLM as a reward model where the reward model grades the coverage and contradiction score[1] between the generated response and the reference response. This is referred to as "cov/con" reward in the results table.

- **Coverage Reward** $r_{\text{response\_cov}}$ measures semantic coverage between generated response x and ground-truth responses $\hat{x}$ whether x is *no coverage* ($r_{\text{response\_cov}}(x, \hat{x}) = 0$), *partial coverage* ($r_{\text{response\_cov}}(x, \hat{x}) = 0.5$), or *full cov* ($r_{\text{response\_cov}}(x, \hat{x}) = 1$) following Laban et al. (2024); Akarajaradwong et al. (2025).
- **Contradiction Reward** $r_{\text{response\_con}}(x, \hat{x}) = 1$ if x does not contradict $\hat{x}$. $r_{\text{response\_con}}(x, \hat{x}) = 0$ otherwise.

Second, our proposed method utilizes the semantic similarity between x and $\hat{x}$. Formally, **Semantic Similarity Reward** computes the similarity score between the generated answer text and the ground-truth answer using an embedding model ($0 < r_{\text{response\_semantic}}(x) < 1$). This is referred to as "semantic reward" in the result table.

Citation Accuracy Reward We design a multi-component verifiable reward function that ensures correct legal citation. In particular, our reward formulation decomposes citation quality into three measurable dimensions:

- **Format Reward** $r_{\text{citation\_format}}(x) = 1$ if the rollout output x adheres specified XML format. In case that x doesn't adhere to XML format, the reward of that rollout is overridden to zero.
- **Grounded Citation Reward** In case that x cites law sections that are *not* provided in the retrieved documents, the reward of that rollout is overridden to zero. Additionally, to encourage correct and grounded citation, any successful citation is rewarded with $r_{\text{citation\_grounded}}(x) = 0.5$.
- **Citation F1 Reward** $r_{\text{citation\_f1}}(x) = F_1$ score of the citation in x.

3.2 Multilingual vs. Thai-CPT Model (RQ2)

We evaluate two types of pretrained models: (1) the original multilingual model instruction-tuned from Qwen2.5-7B (Qwen et al., 2025), and (2) the model that undergoes continued pretraining on Thai-centric corpora prior to instruction tuning.

1. Qwen2.5-7B-Instruct[2] (Qwen et al., 2025)
2. Typhoon2-qwen2.5-7b-Instruct[3] (Pipatanakul et al., 2024)
3. OpenThaiGPT1.5-7b-Instruct[4] (Yuenyong et al., 2025)

To assess the impact of language-specific pretraining, we compare a general-purpose model against two Thai-CPT variants, Typhoon2 and OpenThaiGPT1.5, as both models are finetuned

[1] We adopt these metrics based on Akarajaradwong et al. (2025) response evaluation metrics.

[2] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[3] https://huggingface.co/scb10x/typhoon2-qwen2.5-7b-instruct

[4] https://huggingface.co/openthaigpt/openthaigpt1.5-7b-instruct

on Qwen2.5-7B. This controlled setup allows us to isolate the effect of continued pretraining on the effectiveness of our alignment strategies for Thai legal QA.

4 Experimental Setup

4.1 Datasets

We use the WangchanX-Legal-ThaiCCL-RAG dataset[5] for training and the NitiBench benchmark for evaluation (Akarajaradwong et al., 2025).

Training Data Our training data is a multi-label dataset derived from 35 Thai financial laws via a semi-automated, expert-validated process. It was created using a semi-automated, expert-validated process where questions were generated from legal sections. Crucially, this dataset is inherently multi-label, reflecting that a single legal query can involve multiple relevant statutes.

Evaluation Benchmark The NitiBench benchmark provides two evaluation splits:

- **NitiBench-CCL (In-Domain):** Derived from WangchanX-Legal-ThaiCCL-RAG's expert-curated test set, this benchmark is primarily single-label, designed to test in-domain precision by targeting specific legal sections.
- **NitiBench-Tax (Out-of-Distribution):** This challenging benchmark evaluates generalization using tax rulings from the Revenue Department. It is inherently multi-label and features longer, more complex answers.

4.2 Evaluation Metrics

We evaluate performance using the NitiBench End-to-End (E2E) metrics (Akarajaradwong et al., 2025) with GPT-4o serving as the judge. This aligns with the benchmark's methodology, which validated GPT-4o as the most reliable automated judge due to its high agreement with human legal experts.

To provide a holistic view, we report four key metrics:

- **Citation F1:** F1-score of cited legal sections against the ground truth, measuring the accuracy of legal grounding.
- **Coverage:** A normalized score (0-1) measuring the semantic overlap between the generated and ground-truth answers, assessing answer correctness.
- **Consistency:** Factual consistency with the ground truth, calculated as `1 - Contradiction Score` to ensure reliability and align all metrics on a "higher is better" scale.
- **Joint Score:** The unweighted average of the above three metrics, providing a single, comprehensive measure of overall performance.

4.3 Training Setups

Prompt Construction Our training prompt composed of three components: 1) Instruction , 2) Positive and negative law sections[6], and 3) the question. We limit the number of positive and negative law sections to 10 where negatives are mined from the retriever. Our max prompt length set to 8,192 tokens and the positives and negatives documents order in the prompt are shuffled for every batch sampled. If the constructed prompt is longer than 8192 tokens, we iteratively remove the longest mined negatives and filled it with the negative of the next highest-ranked until the prompt length fits the context limit. The target output format for both IT and GRPO is structured XML-like text including `<reasoning>`, `<answer>`, and `<citation>` tags.

For the retriever, we use the `Human-Finetuned BGE-M3`[7], established by Akarajaradwong et al. (2025) as the top-performing model for Thai legal retrieval. Additional details regarding input and output formatting are provided in Appendix B.

Training Objectives We fine-tune all models using Low-Rank Adaptation (LoRA) (Hu et al., 2021) ($r = 256$, 16-bit precision), applying adapters to all attention layers[8]. We trained all GRPO models for one epoch on a single NVIDIA A100 80GB GPU using the Unsloth (Daniel Han and team, 2023), with a learning rate of 5e-6 and a rollout size of 10. Full hyperparameters are detailed in Appendix A.1.

Baseline (RQ3) For the baseline, we used

- **Base Instruction Tuned Model:** The base instruction-tuned model provided by the original authors.
- **Instruction Tuned Model with LoRA:** The instruction-tuned model with LoRA adapter targeting the same layers with the same rank configuration. We finetuned for 3 epochs on the training set.

[5] https://huggingface.co/datasets/airesearch/WangchanX-Legal-ThaiCCL-RAG

[6] Here, we use the term 'section' by the mean of law section as a retrieved document from the database.

[7] https://huggingface.co/VISAI-AI/nitibench-ccl-human-finetuned-bge-m3

[8] We apply LoRA on `q_proj`, `k_proj`, `v_proj`, `gate`, `up_proj`, `down_proj`

4.4 Inference and Result Aggregation

We report the mean and standard deviation over three inference runs for each model, using vLLM (Kwon et al., 2023) with different random seeds to ensure robust evaluation (see Appendix A.2 for details).

5 Results

Table 1 presents our main results, averaged over three runs. We note that the Citation F1 scores are constrained by the retriever: the upstream BGE-M3 retriever achieves a maximum F1 of 0.9220 on NitiBench-CCL and 0.4809 on NitiBench-Tax. This represents the theoretical upper bound, as models cannot cite unretrieved documents.

5.1 GRPO Reward Strategies (RQ1)

Our results reveal a clear trade-off between the two reward strategies, with performance being highly context-dependent.

For in-domain tasks, the cost-efficient semantic similarity reward proves highly effective. Models that were aligned using semantic reward often outperform the other two variants (instruction tuning and cov/con reward) in the joint score. This confirms its value as an efficient reward proxy when a strong ground-truth answer provides a clear semantic target, offering a cost-effective alternative to an expensive LLM-based reward model.

On the other hand, for complex generalization tasks, rewards from an LLM-based reward model yield consistent performance. On NitiBench-Tax, cov/con reward shows positive performance across all metrics compared to semantic reward, e.g., typhoon2 (semantic reward) performs worse on consistency score compared to baseline. This suggests that for more difficult legal reasoning, the higher-fidelity signal from a capable reward model offers a tangible advantage. Importantly, *both GRPO strategies vastly outperform their instruction-tuned counterparts*.

5.2 Impact of Base Model Priors (RQ2)

GRPO's effectiveness is highly dependent on the base model's priors, especially for the out-of-domain tasks. While the language-generic Qwen2.5 model struggles, GRPO delivers significant gains on the Thai-aligned models (Typhoon2, OpenThaiGPT1.5). This supports the hypothesis that RL enhances sampling efficiency, finding correct reasoning paths that the model can already access, rather than creating new reasoning capacity (Yue et al., 2025). Thai-CPT models appear to possess stronger priors for these complex tasks, and GRPO capitalizes on this by biasing outputs towards correct, pre-existing pathways.

In contrast, instruction tuning consistently degrades performance, likely by disrupting these pathways or overfitting. However, the modest gains in Coverage and Consistency on the NitiBench-Tax set suggest that improved sampling alone is insufficient to fully address complex answer generation, highlighting the limits of RL in expanding a model's core reasoning boundary (Yue et al., 2025).

5.3 Effectiveness of GRPO (RQ3)

GRPO better improves the performance of the LLM compared to instruction tuning across both benchmarks. On the in-domain task, GRPO yields substantially higher gains (e.g., +27-31% gain for Typhoon2 GRPO vs. +15% for instruction tuned on Joint Score). Critically, on the challenging out-of-domain task, GRPO provides a stable performance uplift, whereas instruction tuning consistently degrades model performance.

GRPO models citation performance is comparable with proprietary LLMs on the in-domain benchmark. Under NitiBench-CCL (in-domain) setups, two models (OpenThaiGPT1.5 GRPO with cov/con reward and Qwen2.5-7B GRPO with semantic reward) outperform GPT-4o in citation F1, showing a promising result in aligning LLM towards better citation with an RL-based approach. On the out-of-domain generalization task, all listed larger models significantly outperform our tuned 7B models. NitiBench-Tax requires complex, challenging legal reasoning, so larger models might have advantages while incurring significant inference cost.

6 Reward Composition Analysis

To understand reward contributions, we performed ablations on OpenThaiGPT1.5-7B[9] (see Table 1), comparing our main GRPO variants against configurations using: (1) combined semantic and coverage/consistency rewards ('semantic + cov/con reward'), and (2) only citation-related rewards ('w/o answer reward').

[9]We selectively chose OpenThaiGPT1.5-7B due to its superior performance on NitiBench-CCL joint score.

model	Citation F1 ↑	SD	gains (%)	Coverage ↑	SD	gains (%)	Consistency ↑	SD	gains (%)	Joint score	gains (%)
Nitibench-CCL (In-Domain)											
qwen2.5-7b-instruct	0.4103	0.0015		0.5908	0.0041		0.8402	0.0030		0.6138	
+LoRA instruction tuning	0.5691	0.0040	38.70	0.5832	0.0075	-1.29	0.8341	0.0024	-0.72	0.6622	7.88
+LoRA GRPO (cov/con reward)	0.6796	0.0020	65.63	0.6322	0.0010	7.00	0.8598	0.0009	2.34	0.7239	17.94
+LoRA GRPO (semantic reward)	0.7146	0.0009	74.14	0.7197	0.0023	21.81	0.8232	0.0024	-2.02	0.7525	22.60
typhoon2-qwen2.5-7b-instruct	0.3597	0.0042		0.5587	0.0061		0.8553	0.0076		0.5912	
+LoRA instruction tuning	0.5744	0.0028	59.71	0.6214	0.0030	11.23	0.8572	0.0030	0.22	0.6843	15.75
+LoRA GRPO (cov/con reward)	0.6514	0.0013	81.10	0.7092	0.0039	26.95	**0.9032**	0.0019	5.60	0.7546	27.63
+LoRA GRPO (semantic reward)	0.6828	0.0028	89.84	**0.7735**	0.0012	38.45	0.8757	0.0028	2.38	**0.7773**	31.48
openthaigpt1.5-qwen2.5-7b-instruct	0.4299	0.0048		0.5556	0.0010		0.8234	0.0048		0.6030	
+LoRA instruction tuning	0.5613	0.0069	30.56	0.5930	0.0024	6.73	0.8371	0.0031	1.66	0.6638	10.08
+LoRA GRPO (cov/con reward)	**0.7197**	0.0020	67.40	0.6680	0.0034	20.23	0.8705	0.0034	5.72	0.7527	24.84
+LoRA GRPO (semantic reward)	0.7017	0.0016	63.23	0.7214	0.0041	29.84	0.8554	0.0021	3.89	0.7595	25.96
+LoRA GRPO (semantic + cov/con rewards)	0.6912	0.0024	60.77	0.6109	0.0049	9.95	0.8529	0.0032	3.58	0.7183	19.13
+LoRA GRPO (w/o answer reward)	0.6704	0.0022	55.95	0.5484	0.0042	-1.29	0.8037	0.0086	-2.39	0.6742	11.82
gpt-4o-2024-08-06	0.7140			0.8520			0.9450			0.8370	
gemini-1.5-pro-002	0.6510			0.8650			0.9520			0.8227	
claude-3-5-sonnet-20240620	0.5950			0.8970			0.9600			0.8173	
Nitibench-Tax (Out-of-Domain)											
qwen2.5-7b-instruct	0.2110	0.0272		0.3333	0.0082		0.5733	0.0340		0.3726	
+LoRA instruction tuning	0.0975	0.0192	-53.82	0.2867	0.0249	-13.99	0.5067	0.0094	-11.63	0.2969	-20.30
+LoRA GRPO (cov/con reward)	0.1678	0.0196	-20.47	0.2933	0.0047	-12.00	0.5633	0.0094	-1.74	0.3415	-8.34
+LoRA GRPO (semantic reward)	0.1555	0.0135	-26.31	0.3167	0.0249	-4.99	0.5667	0.0249	-1.16	0.3463	-7.05
typhoon2-qwen2.5-7b-instruct	0.1272	0.0150		0.3333	0.0411		0.5467	0.0249		0.3357	
+LoRA instruction tuning	0.1072	0.0315	-15.71	0.2633	0.0205	-21.00	0.5667	0.0189	3.66	0.3124	-6.95
+LoRA GRPO (cov/con reward)	0.2035	0.0197	60.03	**0.3800**	0.0294	14.00	0.5833	0.0189	6.71	**0.3889**	15.85
+LoRA GRPO (semantic reward)	0.2113	0.0134	66.18	0.3633	0.0411	9.00	0.4933	0.0525	-9.76	0.3560	6.04
openthaigpt1.5-qwen2.5-7b-instruct	0.1850	0.0247		0.3367	0.0519		0.5400	0.0849		0.3539	
+LoRA instruction tuning	0.1039	0.0387	-43.84	0.3267	0.0450	-2.97	0.5800	0.0283	7.41	0.3368	-4.81
+LoRA GRPO (cov/con reward)	0.2085	0.0328	12.73	0.3667	0.0205	12.24	0.5600	0.0748	3.70	0.3784	6.93
+LoRA GRPO (semantic reward)	**0.2482**	0.0054	34.16	0.2500	0.0424	-25.74	**0.6000**	0.0490	11.11	0.3661	3.44
+LoRA GRPO (semantic + cov/con rewards)	0.1830	0.0048	-1.04	0.3067	0.3682	-8.91	0.5267	0.0499	-2.47	0.3388	-4.26
+LoRA GRPO (w/o answer reward)	0.1662	0.0090	-10.16	0.3133	0.0125	-6.93	0.5333	0.0189	-1.23	0.3376	-4.60
gpt-4o-2024-08-06	0.4380			0.5000			0.5400			0.4927	
gemini-1.5-pro-002	0.3320			0.4400			0.5200			0.4307	
claude-3-5-sonnet-20240620	0.4570			0.5100			0.5600			0.5090	

Table 1: Performance comparison (avg ± SD, 3 runs) on Nitibench-CCL and Nitibench-Tax: Baseline vs. instruction tuning, GRPO (cov/con reward), GRPO (semantic reward). Relative performance gains over baseline are indicated. Comparison provided against 3 proprietary LLM results from Akarajaradwong et al. (2025) on the same settings.

6.1 Impact of Combining Answer Rewards (RQ1)

Simply combining the semantic and cov/con rewards degrades the performance. This combined configuration underperformed compared to using either reward individually across most metrics. The performance drop was particularly notable for Coverage and Consistency on NitiBench-CCL, and the model showed poor generalization on NitiBench-Tax. We hypothesize that naive summation creates balancing issues or negative interference, indicating that more sophisticated reward scaling or normalization is required.

6.2 Impact of Removing Answer Rewards (RQ1)

Using only citation-only reward enhances LLM citation ability for in-domain, but does not improve the quality of the generated response. While in-domain Citation F1 improved over baseline (+56%), Coverage and Consistency degraded below baseline levels. This variant also performed worst among GRPO configurations on CCL citation and failed to generalize on NitiBench-Tax (-10% gain). This strongly indicates that **both citation and generation aspects are coupled**; optimizing citations alone harms overall quality and generalization, demonstrating the need for answer quality rewards even to maximize citation performance within GRPO.

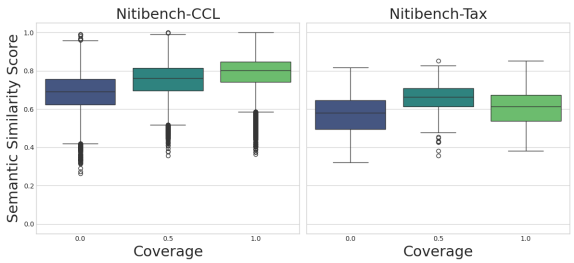

Figure 2: Semantic Similarity distributions by Coverage score level on (a) NitiBench-CCL and (b) NitiBench-Tax. Median similarity tends to increase with coverage on CCL, a trend not observed on Tax.

6.3 Correlation of Semantic Similarity with Coverage and Consistency (RQ1)

To further investigate the reason why the model generalizes poorly on NitiBench-Tax, we con-

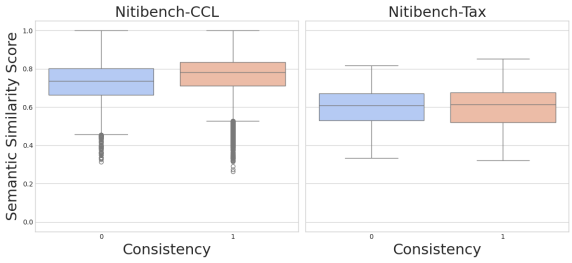

Figure 3: Semantic Similarity distributions by Consistency score on (a) NitiBench-CCL and (b) NitiBench-Tax. Consistent answers on CCL show higher similarity; this distinction is less clear on Tax.

ducted an analysis on the relationship between semantic rewards and cov/con rewards. Figure 2 and 3 show a box plot comparing the coverage and consistency score rated by Qwen2.5-72B reward model with the semantic similarity score from BGE-M3, respectively. Our analysis shows a strong positive correlation between coverage/consistency score and semantic similarity on NitiBench-CCL, explaining the effectiveness of the semantic reward for in-domain setups. Conversely, this correlation disappears on the complex NitiBench-Tax set, where simple semantic overlap is insufficient to capture nuanced factual correctness.

This contrast highlights a critical limitation: while semantic similarity is a viable low-cost proxy for tasks where answers are semantically close to the ground truth, it is an unreliable indicator for complex generalization tasks that demand deeper reasoning and synthesis (see Appendix D).

7 Error Analysis

To better understand the trade-offs between different response quality reward strategies, we performed a qualitative analysis on the outputs of OpenThaiGPT1.5-7B GRPO with semantic and cov/con reward. We highlight three cases as shown in Table 2.

7.1 Success of Integrating Substantive and Procedural Law in Both Reward Choices

In Case 1, a correct answer must combine the substantive rule (THB 20M per spouse; 5% on the excess) with the procedural requirement (payment at the Land Office at the time of registration under Sec. 52). Both reward strategies increased the likelihood that models included this procedurally binding detail, indicating that optimizing for a complete ground-truth match promotes procedural grounding.

7.2 Semantic reward can omit necessary intermediate steps

In tasks that require multi-step reasoning over retrieved context (Case 2), we observe that the GRPO-trained model with a semantic reward can short-circuit the analysis. In this case, the model trained with an LLM-judge reward first *characterizes the instrument*, treating the 'funding agreement' as a 'contract for work' under Civil Code Sec. 587, which makes the instrument taxable, and only then evaluates the exemption under Revenue Code Sec. 121. The semantic-reward variant omitted this foundational step and predicted 'exempt.'

We hypothesize that a semantic-similarity signal emphasizes proximity to the final answer text and may underweight *coverage* of required intermediate premises. In contrast, the LLM-based reward allows explicit coverage/contradiction (cov/con) checks via prompting, which better preserves necessary steps. This pattern is consistent with the results in Table 1, where the semantic-reward model underperforms on NitiBench-Tax, a setting with more multi-step legal analyses.

7.3 Failure of Abstract Reasoning in Both Reward Strategies

A common challenge for both models appeared in tasks requiring abstract legal synthesis, where a term in the question must be re-characterized to connect with the provided law.

In Case 3, the query involves a 'Memorandum of Understanding' (MOU), and the provided law discusses tax on a 'provision of service'. The expert reasoning requires a multi-step conceptual leap: 'MOU' $\rightarrow$ is a type of 'contract for work' $\rightarrow$ which qualifies as a 'provision of service' $\rightarrow$ making the VAT statute applicable.

Neither model could bridge this conceptual gap. Semantic reward incorrectly defined an MOU as non-binding and thus non-taxable, a common misconception, but wrong in this legal context. LLM-based reward model focused on the surface-level term mismatch between "income tax withholding" (in the question) and "Value Added Tax" (in the law), correctly noting they are different but failing to solve the underlying legal problem of whether the transaction was taxable at all.

This shared failure highlights a boundary for current alignment techniques. While the models can follow explicit rules, they struggle to perform the abstract, multi-step synthesis that is central to legal

Case 1: Success in Integrating Procedural and Substantive Law	
Question	A married couple jointly gifts marital property to their child. They argue they should each get a separate THB 20 million tax exemption. Is this correct?
Relevant Law	**Revenue Code, Sec. 52:** Specifies that for property transfers, tax must be paid to the official at the Land Office *at the time of registration*.
Expert Reasoning Path	1. Address the substantive question about the exemption amount. 2. Integrate the procedural requirement of how and where to pay the tax.
Gold Answer	**Correct**. Each spouse gets a THB 20 million exemption. For any excess value, they must pay 5% tax. **Crucially, per Sec. 52, this tax must be paid to the Land Office official at the moment of registration.**
Semantic Reward Model	**Yes**. Each spouse qualifies for an individual exemption. **Also, the tax must be paid at Land Office official (according to Revenue Code, Section 52).**
LLM-Based Reward Model	**Yes, your understanding is correct.** Each spouse is entitled to their own exemption. **And the tax must be paid to the official at the Land Office during the registration process** as stated in Section 52 of Revenue Code.
Case 2: Challenges in Foundational Reasoning of the Semantic Proxy	
Question	A state agency signs a funding agreement for a research project. Is the agency exempt from stamp duty on this agreement?
Relevant Law	**Civil Code, Sec. 587:** Defines a 'contract for work'. **Revenue Code, Sec. 121:** Grants stamp duty exemption to government agencies, *unless* they conduct commercial activities.
Expert Reasoning Path	1. (Foundational Step) Characterize the 'funding agreement' as a 'contract for work' under Sec. 587. This makes it a taxable instrument. 2. (Main Question) Evaluate if the agency's commercial activities disqualify it from the exemption under Sec. 121.
Gold Answer	The agreement is a 'contract for work' (Sec. 587), making it taxable. The agency's activities are commercial, disqualifying it from the Sec. 121 exemption. Therefore, it **is not exempt**.
Semantic Reward Model	**Yes, it is exempt. [Failure: Skips the foundational step of characterizing the agreement.]** Because it is a state agency, it is exempt under Section 121.
LLM-Based Reward Model	**The institute is not exempt. [Success: Follows the correct reasoning path.]** The agreement is a 'contract for work' (Sec. 587) and thus taxable. Its commercial activities disqualify it from the exemption under Sec. 121.
Case 3: Shared Failure in Abstract Reasoning	
Question	A state agency uses an MOU for project funding. Is the funding subject to income tax withholding?
Relevant Law	**Revenue Code, Sec. 77/2:** States that a 'provision of service' is subject to Value Added Tax (VAT).
Expert Reasoning Path	1. (Abstract Leap) Recognize that this MOU functions as a 'contract for work'. 2. (Abstract Leap) A 'contract for work' is a 'provision of service'. 3. (Conclusion) Therefore, the transaction is subject to VAT under Sec. 77/2.
Gold Answer	**Yes**, the payment is subject to tax. The MOU is effectively a 'contract for work', which constitutes a 'provision of service' and is therefore subject to VAT as per Sec. 77/2.
Semantic Reward Model	**No. [Failure: Fails the abstract leap.]** MOUs are considered non-binding agreements and payments under them do not constitute taxable income.
LLM-Based Reward Model	**No. [Failure: Focuses on surface mismatch, misses the underlying problem.]** The question is about income tax, but the context is about VAT. Therefore, the context is irrelevant.

Table 2: Qualitative error analysis of OpenThaiGPT1.5-7B-Instruct with different reward signals. We highlight **correct reasoning** and **flawed reasoning**.

expertise. This points to a clear area for future work in developing reward mechanisms that incentivize deeper conceptual reasoning.

8 Efficiency of Reward Signal Proxies

To investigate the efficiency gains between semantic and LLM-based rewards, we compared the training cost in GPU-hours of both methods. When using LLM-based judge, Qwen2.5-72B-Instruct served with optimized inference stack, the training time cost up to 264 GPU-hours ($216). Our proposed semantic reward, on the other hand, reduced the GPU-hours needed for training down to 104 GPU-hours ($85), achieving up to over 2.5× cost and time saving.Additionally, semantic reward is also more memory efficient as training only requires one GPU, while the LLM-based reward model requires one additional GPU for hosting the reward model. [10]

This disparity arises because the lightweight BGE-M3 calculation adds minimal latency to the RL loop, whereas the 72B judge model requires a dedicated GPU for inference, creating a significant training bottleneck. While a large judge may offer reward signals with higher fidelity to final evaluation metrics, its computational cost is a major barrier to online RL training. The strong in-domain performance of the BGE-M3 proxy confirms its value as a cost-effective and practical method for injecting answer quality signals during GRPO.

9 Conclusion

This work demonstrates that GRPO can be used to align LLMs toward better legal question answering and usually outperforms instruction tuning. The model aligned with GRPO yields significant citation accuracy gains on in-domain tasks and improves generalization on complex legal reasoning for Thai-aligned models, a setting where instruction tuning consistently degrades performance.

Ablations of reward functions reveal a key trade-off: an efficient semantic reward is cost-effective for in-domain tasks but loses effectiveness on complex generalization. While an LLM-based reward model provides a more accurate signal, it does so at over 2.5× the computational cost. Our ablations confirm that using both citation and answer quality rewards is necessary for the best outcomes.

Ultimately, our findings show that GRPO is a highly effective approach for specialized domains, but its success depends on the careful synergy of the RL algorithm, base model capabilities, and reward design. Future work should focus on creating reward mechanisms that are both accurate enough for complex reasoning and efficient enough for practical training.

[10] Cost in USD was estimated based on A100 80GB PCIE median rental cost of $0.82/hr via https://vast.ai/pricing/gpu/A100-PCIE accessed April 2025.

Limitations

Our study is subject to four key limitations, primarily stemming from computational constraints and the unique structure of our dataset.

First, our exploration of combining reward signals was restricted; while a naive combination of semantic and judge-based rewards proved suboptimal, we could not exhaustively explore calibrated weighting or normalization schemes that might yield synergistic benefits.

Second, we applied GRPO exclusively to instruction-tuned models, leaving the investigation of its direct application to base pre-trained models as an area for future work.

Third, our evaluation was confined to a single, specialized legal dataset. This limitation arises from the scarcity of publicly available legal corpora and, more significantly, the novelty of our data-framing approach, which organizes context around individual legal sections—the smallest reasonable unit of law. Replicating this fine-grained structure in other legal or non-legal domains (e.g., medicine) to enable cross-domain evaluation was beyond the scope of this work.

Finally, our experiments utilized the standard GRPO algorithm (Shao et al., 2024). We did not evaluate the more recent "Dr. GRPO" variant (Liu et al., 2025), which introduces improvements like length normalization to address known optimization biases. A direct comparison of these algorithms presents a valuable direction for future research.

References

Pawitsapak Akarajaradwong, Pirat Pothavorn, Chompakorn Chaksangchaichot, Panuthep Tasawong, Thitiwat Nopparatbundit, and Sarana Nutanong. 2025. Nitibench: A comprehensive study of llm framework capabilities for thai legal question answering. *Preprint*, arXiv:2502.10868.

Counsel AI Corporation. 2025. Harvey ai. Accessed: 2025-04-25.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *Preprint*, arXiv:2304.06767.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *Preprint*, arXiv:2405.07863.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Vivek Kumar and Arjun Subramaniam. 2019. Abstractive summarisation with bertscore reward. CS229 Project Report.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems. *Preprint*, arXiv:2407.01370.

LexisNexis. 2023. LexisNexis Launches Lexis+ AI, a Generative AI Solution with Hallucination-Free Linked Legal Citations. lexisnexis.com. [Accessed 13-08-2024].

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *Preprint*, arXiv:2503.20783.

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. *Preprint*, arXiv:2405.20362.

Kunat Pipatanakul, Potsawee Manakul, Natapong Nitarach, Warit Sirichotedumrong, Surapon Nonesung, Teetouch Jaknamon, Parinthapat Pengpun, Pittawat

Taveekitworachai, Adisai Na-Thalang, Sittipong Sripaisarnmongkol, Krisanapong Jirayoot, and Kasima Tharnpipitchai. 2024. Typhoon 2: A family of open text and multimodal thai large language models. *Preprint*, arXiv:2412.13702.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *Preprint*, arXiv:2408.10343.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Ehsan Shareghi, Jiuzhou Han, and Paul Burgess. 2024. Methods for legal citation prediction in the age of llms: An australian law case study. *Preprint*, arXiv:2412.06272.

Hao Sun, Yunyi Shen, Jean-Francois Ton, and Mihaela van der Schaar. 2025. Reusing embeddings: Reproducible reward model research in large language model alignment without gpus. *Preprint*, arXiv:2502.04357.

Akash Takyar. 2024. AI agents for legal: Applications, benefits, implementation and future trends — leewayhertz.com. leewayhertz.com. [Accessed 13-08-2024].

Kobkrit Viriyayudhakorn. 2024. Thanoy AI Chatbot - genius AI lawyer. iapp.co.th. [Accessed 13-08-2024].

Go Yasui, Yoshimasa Tsuruoka, and Masaaki Nagata. 2019. Using semantic similarity as reward for reinforcement learning in sentence generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 400–406, Florence, Italy. Association for Computational Linguistics.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Preprint*, arXiv:2504.13837.

Sumeth Yuenyong, Kobkrit Viriyayudhakorn, Apivadee Piyatumrong, and Jillaphat Jaroenkantasima. 2025. Openthaigpt 1.5: A thai-centric open source large language model. *Preprint*, arXiv:2411.07238.

Kepu Zhang, Weijie Yu, Sunhao Dai, and Jun Xu. 2025. Citalaw: Enhancing llm with citations in legal domain. *Preprint*, arXiv:2412.14556.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Hyperparameters

A.1 Training Hyperparameters

This section details the key hyperparameters used for Instruction Tuning Fine-tuning (IT) and Group Relative Policy Optimization (GRPO) training procedures described in Section 4. Common parameters related to LoRA configuration, precision, optimizer betas, and data handling were kept consistent where applicable.

Hyperparameter	GRPO Value	IT Value
Learning Rate (lr)	5.00E-06	1.00E-05
LR Scheduler Type	`constant_with_warmup`	`cosine`
Max Gradient Norm	0.2	1.0
Epochs	1	3
Rollout Batch Size	10	N/A
SFT Batch Size	N/A	4
Max Prompt Length	8192	8192
Max Completion Length	2048	2048
LoRA Rank (r)	256	256
Precision	bfloat16	bfloat16
Retrieval Top-k	10	10
Gradient Accumulation Steps	1	1
Weight Decay	0.1	0.1
Warmup Ratio	0.1	0.1
Adam Beta1	0.9	0.9
Adam Beta2	0.99	0.99

Table 3: Comparison of Key Hyperparameters for SFT and GRPO Training.

A.2 Inferencing Hyperparameters

This section details the hyperparameters used during the inference phase to generate the model outputs for the final evaluation presented in Section 5. These settings were applied consistently across all model configurations (Baseline, SFT, GRPO) when evaluating on the Nitibench-CCL and Nitibench-Tax test sets using vLLM (Kwon et al., 2023). The following parameters were used for text generation:

Generation Seeds: Inference was repeated three times for each model configuration using the following distinct random seeds: `69420`, `69421`, and `69422`. The final reported metrics are the mean and standard deviation across these three runs (as detailed in Section 4.4).

Retrieval Top-k: Set to `10`, same as the `Retrieval Top-k` in the training hyperparameter.

Temperature: Set to `1.0` for standard diversity in the output.

B Input and Output Formats

This section provides concrete examples of the input prompt structure fed to the models and the target output format used during fine-tuning (both SFT and GRPO), complementing the description in Section 4.

B.1 Example Input Prompt Structure

The following illustrates the format of the input provided to the models. This example assumes the context retrieval resulted in $k = 5$ relevant sections after length management. The `{context}` placeholder represents the actual text content of the corresponding legal section. The `<law_code>` tags contain unique integer identifiers assigned to each distinct legal section within our corpus; these identifiers are used as keys and do not necessarily correspond to official statutory section numbers.

```
What is the difference between financial
    institution business and financial
    business?

Relevant sections
<law_code>1</law_code><context>...</context>
<law_code>2</law_code><context>...</context>
<law_code>3</law_code><context>...</context>
<law_code>4</law_code><context>...</context>
<law_code>5</law_code><context>...</context>
```

B.2 Example Target Output Structure

The models were trained to generate outputs adhering to the following XML-like structure. This format separates the reasoning process, the final answer, and the cited sources.

```
<reasoning>
The laws related to the method for director
    resignation are ...
</reasoning>
<answer>
According to Section 1153/1 of the Civil and
    Commercial Code and ...
</answer>
<citation>
<law_code>2</law_code>
<law_code>5</law_code>
</citation>
```

Note: The `<reasoning>` block contains the model's generated explanation or thought process. The `<answer>` block contains the final synthesized answer to the query. The `<citation>` block lists the `<law_code>` identifiers that the model cites as sources for its answer. During IT, this structure represents the target output. During GRPO, adherence to this format and the correctness of the content within the tags (`<answer>` and `<citation>`) are evaluated by the reward functions.

C Evaluation of Qwen-72B as an Automated Judge

To assess the viability of using Qwen2.5-72B-Instruct as an online judge for generating Coverage and Consistency rewards in GRPO (Section 3.1), we compared its judgment reliability against `gpt-4o-2024-08-06` on the Nitibench-CCL dataset, as it achieved the highest performance among judges evaluated in the original Nitibench paper (Akarajaradwong et al., 2025). We follow Nitibench's decoding hyperparameters: `temperature = 0.5`, `seed = 69420`, and `max_completion_tokens = 2048`.

As shown in Table 4, Qwen-72B achieved high reliability, closely matching GPT-4o. For **Coverage**, Qwen-72B reached an F1-score of 0.84 (vs. 0.88 for GPT-4o), and for **Consistency**, it scored 0.97 (vs. 0.98 for GPT-4o). These results demonstrate that Qwen2.5-72B-Instruct functions as a reliable automated judge for these metrics on this dataset, validating its use for providing sufficiently accurate reward signals during GRPO training as an alternative to external API calls.

Model	Metric	Precision	Recall	F1-score	Support
		Nitibench-CCL			
gpt-4o-2024-08-06	Coverage	.88	.88	.88	200
	Consistency	.98	.97	.98	150
Qwen2.5-72B-Instruct	Coverage	.85	.83	.84	200
	Consistency	.98	.97	.97	150

Table 4: Performance comparison of GPT-4o (`gpt-4o-2024-08-06`) and Qwen2.5-72B-Instruct as automated judges for Coverage and Consistency metrics on the Nitibench-CCL dataset.

D Complexity of Nitibench-Tax over Nitibench-CCL

While both Nitibench-CCL and Nitibench-Tax evaluate Thai Legal QA, the Nitibench-Tax dataset presents a significantly more complex challenge, designed specifically to test model generalization and deeper reasoning capabilities (see Figure 4 for answer length and section per answer comparison).

model	Citation F1 ↑	SD	gains (%)	Coverage ↑	SD	gains (%)	Consistency ↑	SD	gains (%)	Joint score	gains (%)
Nitibench-CCL											
openthaigpt1.5-qwen2.5-7b-instruct	0.4299	0.0048		0.5556	0.0010		0.8234	0.0048		0.6030	
+LoRA GRPO (semantic reward)	**0.7017**	0.0016	63.23	**0.7214**	0.0041	29.84	**0.8554**	0.0021	3.89	**0.7595**	25.96
+LoRA GRPO (semantic reward, citation first)	0.6545	0.0044	52.25	0.7065	0.0053	27.16	0.8528	0.0028	3.57	0.7379	22.39
Nitibench-Tax											
openthaigpt1.5-qwen2.5-7b-instruct	0.1850	0.0247		**0.3367**	0.0519		0.5400	0.0849		0.3539	
+LoRA GRPO (semantic reward)	**0.2482**	0.0054	34.16	0.2500	0.0424	-25.74	**0.6000**	0.0490	11.11	0.3661	3.44
+LoRA GRPO (semantic reward, citation first)	0.2172	0.0146	17.43	0.2768	0.0026	-17.79	0.5333	0.0411	-1.24	0.3424	-3.24

Table 5: Comparison of GRPO (semantic reward) performance on OpenThaiGPT1.5-7B using the default output format (reasoning->answer->citation) versus a modified format placing citations before the answer (reasoning->citation->answer).

This difference stems from several key aspects of their origin and structure:

1. **Dataset Origin and Curation:**
   - **Nitibench-CCL:** This dataset was curated manually by legal experts who crafted question-answer pairs primarily based on single, specific legal sections from a defined corpus of 35 financial laws. The process involved a two-tiered expert review to ensure quality. While its corresponding training data (from WangchanX-Legal-ThaiCCL-RAG[11]) could be multi-label due to semi-automated generation, the test set used for evaluation predominantly consists of single-label instances.
   - **Nitibench-Tax:** This dataset originates from real-world tax rulings scraped directly from the Thai Revenue Department's official website[12] (cases from 2021 onwards). These represent authentic inquiries and official responses, reflecting the complexity of actual tax law application. The curation involved extracting relevant cited sections and condensing the official responses using an LLM, after filtering out non-interpretive cases.

   The use of real, official rulings in Nitibench-Tax inherently introduces more complex scenarios and language compared to the expert-crafted, typically single-provision-focused questions in the Nitibench-CCL test set.

2. **Answer Length and Complexity:** The complexity difference is reflected in the average length of the ground-truth answers (after condensation). The average answer length in **Nitibench-CCL is approximately 75 characters**, whereas in **Nitibench-Tax, it is roughly 606 characters** - over eight times longer on average. This suggests that Tax answers inherently require significantly more detail and potentially cover more sub-points derived from the underlying complex rulings.

3. **Multi-Label Nature (Sections per Answer):** This is a critical quantitative differentiator. The Nitibench-CCL test set is explicitly single-label, with an average of **1 ground-truth relevant legal section** per question. In contrast, Nitibench-Tax is inherently multi-label, with an average of **2.62 relevant sections** per case. This requires models not just to identify relevant sections but to synthesize information and reason across multiple legal provisions simultaneously, significantly increasing the reasoning complexity compared to the single-label focus of CCL.

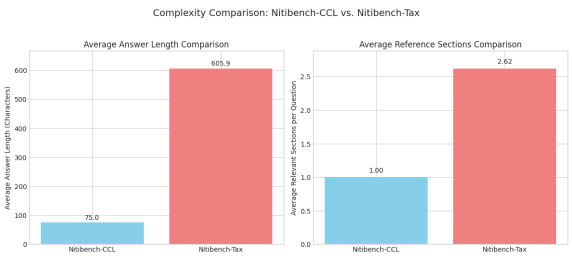

Figure 4: Complexity Comparison of Nitibench-CCL vs. Nitibench-Tax.

In summary, the combination of using real-world, complex tax rulings as source material and their inherent multi-label requirement (demanding reasoning across multiple sections) makes Nitibench-Tax a substantially harder benchmark than Nitibench-CCL for evaluating advanced legal reasoning and generalization abilities.

[11] https://huggingface.co/datasets/airesearch/WangchanX-Legal-ThaiCCL-RAG

[12] https://www.rd.go.th

E Impact of Citation and Answer Position in Output Format

The standard output format used in our experiments follows the structure: reasoning -> answer -> citation (as in Appendix B.2), where the model first provides its reasoning, then the synthesized answer, and finally the supporting citations. To investigate whether the position of the citation block relative to the answer block influences performance, we conducted an additional experiment.

We modified the target output structure to: reasoning -> citation -> answer, placing the citation block immediately after the reasoning and before the final answer. We then retrained the OpenThaiGPT1.5-7B-Instruct model using the GRPO (semantic reward) configuration with this modified "citation-first" target format. All other training parameters remained identical to the corresponding main experiment run.

The results of this comparison are presented in Table 5. The data clearly indicates that altering the standard format to place citations before the answer consistently resulted in **lower performance across nearly all metrics** on both the Nitibench-CCL and Nitibench-Tax datasets compared to the default format, where citations appear last. Notably, Citation F1, Coverage, and the overall Joint Score decreased in the "citation-first" configuration. On the challenging Nitibench-Tax set, this format led to performance even worse than the baseline in terms of Joint Score (-3.24% gain).

While the exact reasons require deeper analysis, this finding suggests that the default structure (reasoning -> answer -> citation) may provide a more natural or effective flow for the model during generation and training. It's possible that generating the answer text first helps the model consolidate the information needed before explicitly listing the supporting citations. Regardless, based on these results, maintaining the structure with the citation block at the end appears preferable for achieving optimal performance with our GRPO approach.