QuantJudger: Boosting Reasoning Capability of LLMs for Legal Quantitative Prediction via Power-Ratio Reward

Anonymous ACL submission

Abstract

Legal quantitative prediction involves estimations on fines, compensation, or criminal amounts for illegal cases based on regulations. It plays a critical role in regulatory compliance, judicial assistance, financial risk assessment, and legal technology applications. However, boosting the performance of such numerical reasoning remains challenging for two key reasons: (1) legal amounts are typically determined within a discretionary range rather than a single precise ground truth, causing conventional accuracy-based rewards to fail in guiding model training; and (2) the task demands multi-step procedural reasoning that maps heterogeneous case facts to numerical values under implicit legal standards, which current LLMs struggle to learn. To address these issues, we construct a comprehensive dataset specifically for LLM legal quantitative prediction, covering GDPR enforcement, EU financial regulations, and Chinese data security and personal information laws. Then, we train a reinforcement learning model using GRPO with Power-Ratio Reward specifically designed for legal quantitative prediction. Compared to the base model Qwen3-8B, our method improves prediction accuracy by +20.2% and our defined "*ratio*" metric by +35.9%. Furthermore, the model generalizes well to LawBench and CAIL-2018, which involve fine judgments and criminal amounts in court cases, achieving improvements of +14.4% and +16.3% in accuracy, respectively.

1 Introduction

The rapid advancement of legal large language models (LLM) (Shu et al., 2024; Fei et al., 2025) has progressively shifted from foundational tasks such as court judgment prediction and legal document retrieval (Luo et al., 2025) toward more complex and practically impactful endeavors. Among these, legal quantitative prediction—the automatic estimation of fines, compensation amounts, or criminal sums directly from case fact descriptions—stands out as a particularly valuable yet underexplored problem. We demonstrate several case examples for this task in Figure 1. This task holds significant promise for a wide range of domains, including regulatory compliance monitoring, judicial decision support, financial risk assessment, and legal technology applications. By automatically inferring numerical outcomes from case narratives, a LLM quantitative judge can assist regulators in prioritizing enforcement actions, help companies to estimate the compliance costs, and enable compliance officers to assess potential exposure more efficiently.

GDPR Case: The Italian DPA investigated Fastweb S.p.A. following hundreds of complaints about aggressive telemarketing practices. The investigation revealed systematic violations involving call centers using unregistered phone numbers and processing contact data obtained from external partners without valid consent.

Fine: 4,500,000 €

Chinese Data Law Case: Meituan, a food delivery and lifestyle services platform, collects real-time location, order histories, and merchant data from millions of daily transactions. A 2021 antitrust probe revealed that Meituan's exclusive merchant policies led to concentrated data holdings, potentially stifling competition and affecting service pricing for consumers.

Fine: 3,440,000,000 ¥

EU Finance Law Case: In October 2020, Coinbase's e-money arm, CBPL, agreed to stop onboarding high-risk customers due to severe anti-money laundering failures. However, during the 2020–2021 crypto boom, the company prioritized rapid growth. Flawed automated systems failed to block these restrictions, allowing over 226,000 high-risk users to open accounts and transact billions unmonitored.

Fine: 4,100,000 €

Figure 1: Case examples of legal quantitative prediction.

Despite its practical importance, legal quantitative prediction using LLM poses substantial challenges. First, legal amounts are typically determined within a discretionary range rather than as a single incontestable figure. Regulatory fines and judicial penalties arise from a combination of statutory guidelines, aggravating and mitigating fac-

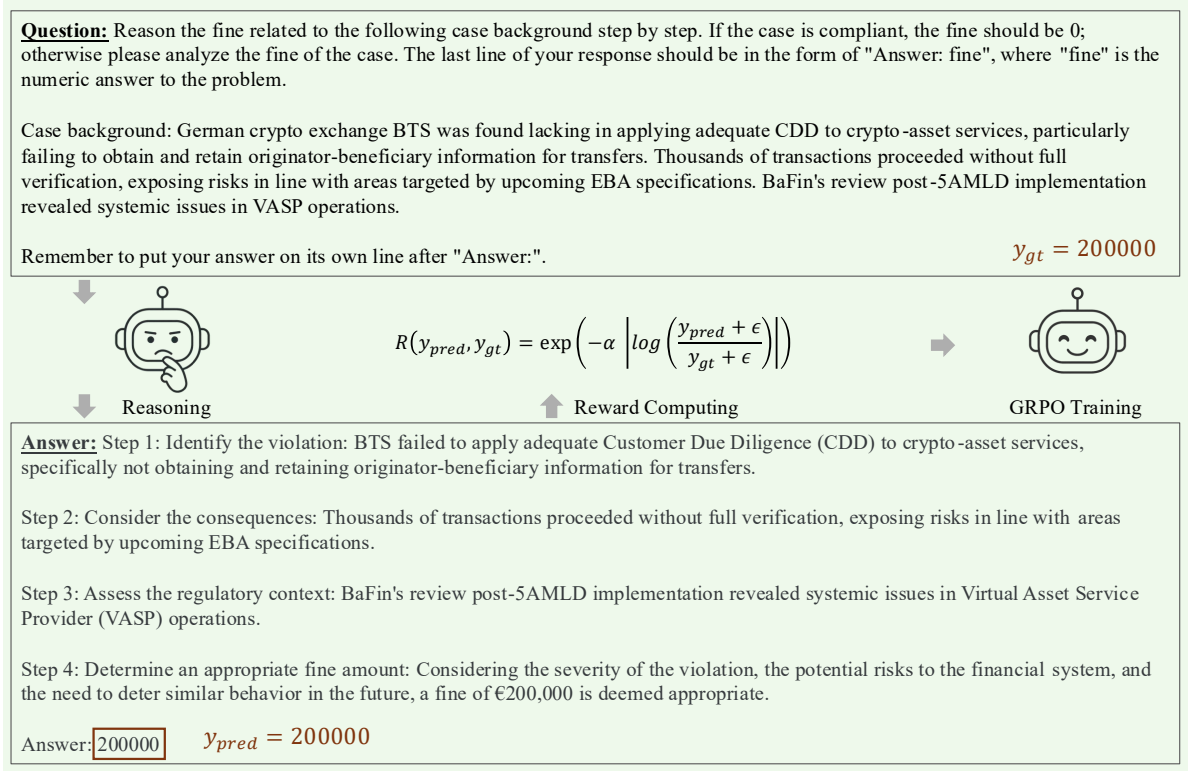


Figure 2: Overview of our training pipeline for enhancing legal quantitative prediction. It begins with our model reasoning over a given question and generating response rollouts. Next, we extract predicted numerical values from these rollouts and feed them into our designed reward model, which calculates rewards to guide GRPO training.

tors, and institutional discretion, meaning that the ground truth in any given case represents only one plausible point within a zone of reasonable outcomes. This inherent ambiguity causes conventional accuracy-based reward signals to fail during training: a model that predicts a fine of 180,000 euros may be just as legally defensible as the recorded 200,000 euro ground truth, yet a standard accuracy loss would penalize it indiscriminately.

Besides, the task demands multi-step procedural reasoning of a kind that current LLMs struggle to learn reliably. A proper fine determination requires sequentially identifying the specific legal provision violated, assessing the severity and scope of the infraction, considering contextual factors such as the entity's cooperation or remedial measures, and only then arriving at a numerical conclusion.

Existing works on legal mathematical reasoning have begun to explore the intersection of numerical computation and legal text (Zhang et al., 2025). This work has applied two-stage reinforcement learning pipelines combining supervised fine-tuning to improve performance on compensation calculation tasks. However, these studies have largely focused on relatively structured scenarios where the mathematical operation is well-defined and the ground truth is deterministic, such as summing predefined damages or applying statutory multipliers. The more challenging setting of discretionary fine prediction, where the target value is inherently approximate and the reasoning path is procedurally complex, remains largely unaddressed. Directly applying existing methods to this domain often leads to training failure, as the reward functions designed for exact-match or near-exact settings prove unsuitable for discretionary numerical targets.

To address these challenges, we innovate a new reward strategy to boost the reasoning capability on legal quantitative prediction. First, we construct a multi-domain legal quantitative prediction dataset spanning General Data Protection Regulations (GDPR), EU financial regulations (including Cross-boarder Transaction Law, Crypto-asset Law and Electronic money Law), and Chinese data law (Data Security Law, Cyber-security Law and Personal Information Protection Law), capturing diverse regulatory language, infraction types, and fine magnitudes. Second, we leverage Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to train a reinforcement learning model, with Power-Ratio Reward. We specifically design this reward model for the discretionary, ratio-sensitive nature of legal numerical prediction. Instead of penaliz-

Dataset	# Train ($y_{\text{gt}} \neq 0$ / $y_{\text{gt}} = 0$)	# Test ($y_{\text{gt}} \neq 0$ / $y_{\text{gt}} = 0$)	Task	Language
GDPR	2346 / 0	585 / 0	Fine Prediction	English
EU Finance Law	1166 / 1306	282 / 337	Fine Prediction	English
Chinese Data Law	209 / 830	64 / 197	Fine Prediction	English
LawBench	—	500 / 0	Criminal Amount Prediction	Chinese
CAIL-2018	—	323 / 277	Fine Prediction	Chinese
LexNum	—	743 / 0	Compensation Prediction	Chinese

Table 1: Data statistics details of our collected dataset. $y_{\text{gt}} \neq 0$ represents data samples with the numerical ground truth of the fine that is not equal to zero. $y_{\text{gt}} = 0$ means case data samples that are compliant and not punished.

ing all deviations equally, it evaluates predictions by their proportional relationship to ground truth, aligning training with the reasonable variation tolerated in legal decision-making. Our contributions can be summarized as:

- We first focus on the important task of legal quantitative prediction and build a multi-domain dataset for training and evaluation, spanning EU GDPR enforcement, EU financial regulations, and Chinese data laws.
- We propose power-ratio reward, a reward function suited to the ratio-sensitive nature of the task. Our method, built on Qwen3-8B, improves prediction accuracy by +20.2% and our metric *"ratio"* by +35.9% over base model Qwen3-8B.
- Our model generalizes well to out-domain benchmarks, with improvements of +14.4% on LawBench (Fei et al., 2024) and +16.3% on CAIL-2018 (Xiao et al., 2018) in accuracy.

2 Power-Ratio Reward

We first present a preliminary introduction to the reinforcement learning (RL) framework used in this work, and then describe the reward function designed explicitly for legal quantitative prediction tasks.

2.1 Preliminary: PPO and GRPO

Proximal Policy Optimization (PPO) (Schulman et al., 2017). PPO is a frequently-used reinforcement learning algorithm for fine-tuning models under contemporary LLM training paradigm. Let the probability ratio between the current policy π_θ and the old policy $\pi_{\theta_{\text{old}}}$ for a given prompt-response pair (x, o) be defined as $r_\theta = \frac{\pi_\theta(o|x)}{\pi_{\theta_{\text{old}}}(o|x)}$. The PPO objective is then given by

$$\mathcal{J}_{\text{ppo}}(\theta) = \mathbb{E}\left[\min\left(r_\theta A,\ \text{clip}(r_\theta, 1-\epsilon, 1+\epsilon)A\right)\right], \quad (1)$$

where A is an estimator of the advantage function which incorporates a value model V_ϕ and a reward function $R(x, o)$, and ϵ is a hyperparameter that defines the clipping range. This clipped objective discourages large changes to the policy by removing the incentive for moving the ratio $r(\theta)$ outside of the interval $[1-\epsilon, 1+\epsilon]$, thereby striking a balance between exploration and learning stability.

Group Relative Policy Optimization (GRPO). GRPO (Shao et al., 2024) further extends PPO by estimating advantages from the relative performance of a group of outputs. Given an input x and a group of G sampled outputs $\{o_i\}_{i=1}^G$ from the old policy $\pi_{\theta_{\text{old}}}$, each output is assigned a scalar reward r_i by a reward model. The advantage for the i-th output is then computed purely from within this group by normalizing the rewards:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}. \quad (2)$$

The policy is subsequently optimized by maximizing the following clipped objective, averaged over all outputs in the group:

$$\mathcal{J}_{\text{grpo}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \min\left(r_\theta A_i,\ \text{clip}\left(r_\theta, 1-\epsilon, 1+\epsilon\right) A_i\right). \quad (3)$$

This formulation allows the model to learn directly from comparative group feedback without training a value model, making it a particularly efficient and stable approach for aligning LLMs with reward signals.

2.2 Power-Ratio Reward

Accuracy-based rewards are commonly used to incentivize reasoning in language models, but their binary (0 or 1) nature provides a sparse supervision signal that is too hard for legal quantitative prediction tasks, where exact matches are rare and no effective gradient propagates. To address this, we propose Power-Ratio Reward, which offers a soft and smooth reward for near-correct predictions. The reward is defined as:

$$R(y_{\text{pred}}, y_{\text{true}}) = \exp\left(-\alpha \cdot \left|\log\left(\frac{y_{\text{pred}}+\epsilon}{y_{\text{gt}}+\epsilon}\right)\right|\right), \quad (4)$$

Model	GDPR (Ratio/Accuracy)	Chinese (Ratio/Accuracy)	Finance (Ratio/Accuracy)	Avg. (Ratio/Accuracy)
GPT-4o-mini	19.79 / 2.72	74.89 / 72.62	62.54 / 56.31	52.41 / 43.88
GPT-5-mini	27.81 / 4.93	78.22 / 73.00	72.30 / 59.60	59.44 / 45.85
GPT-5.4	44.52 / 16.84	79.64 / 76.17	68.14 / 60.90	64.10 / 51.30
GPT-5.5	48.82 / 19.22	**82.54** / 78.07	70.52 / 60.28	67.29 / **52.52**
Claude-Opus-4.7	39.86 / 9.86	80.46 / 76.80	71.82 / 60.62	64.05 / 49.10
DeepSeek-V4-Pro	37.31 / 11.05	81.16 / **79.04**	67.38 / 60.96	61.95 / 50.35
Gemini-3.1-Pro	**49.54** / **19.90**	74.19 / 71.44	39.80 / 36.32	54.51 / 42.55
Qwen3-1.7B	4.59 / 0.68	37.77 / 37.41	25.84 / 25.08	22.73 / 21.06
+ GRPO	$36.35_{(+31.8)}/6.29_{(+5.6)}$	$81.84_{(+44.1)}/76.51_{(+39.1)}$	$70.04_{(+44.2)}/59.11_{(+34.0)}$	$62.74_{(+40.0)}/47.30_{(+26.2)}$
Qwen3-4B	2.85 / 0.17	28.24 / 26.91	25.43 / 23.15	18.84 / 16.74
+ GRPO	$43.34_{(+40.5)}/6.80_{(+6.6)}$	$82.10_{(+53.9)}/74.27_{(+47.4)}$	$75.18_{(+49.8)}/59.37_{(+36.2)}$	$66.87_{(+48.0)}/46.82_{(+30.1)}$
Qwen2.5-7B	6.27 / 0.68	48.52 / 47.95	38.35 / 35.63	31.05 / 28.09
+ GRPO	$42.56_{(+36.3)}/6.46_{(+5.8)}$	$80.66_{(+32.1)}/73.01_{(+25.1)}$	$75.21_{(+36.9)}/61.65_{(+26.0)}$	$66.15_{(+35.1)}/47.04_{(+19.0)}$
Qwen3-8B	5.59 / 0.34	50.06 / 48.54	40.11 / 37.71	31.92 / 28.86
+ GRPO	$44.80_{(+39.2)}/7.82_{(+7.5)}$	$82.52_{(+32.5)}/77.17_{(+28.6)}$	$\mathbf{76.12}_{(+36.0)}/\mathbf{62.14}_{(+24.4)}$	$\mathbf{67.81}_{(+35.9)}/49.04_{(+20.2)}$

Table 2: Fine prediction results across in-domain datasets (Ratio ↑ / Accuracy ↑).

where $y_{\text{pred}}, y_{\text{gt}}$ represent the numerical value of prediction and ground truth, ϵ is a small constant for numerical stability, and $\alpha > 0$ controls the decay rate. This reward lies in $(0, 1]$, equals 1 only for perfect predictions, and decays as a power law with fractional error, providing effective supervision for legal quantitative reasoning.

3 Dataset Collection

In this section, we introduce our training datasets and data for generalization evaluations. The dataset statistics details are demonstrated in Table 1.

3.1 Training Datasets.

- **GDPR Cases**: We leverage GDPR enforcement records from the GDPR Enforcement Tracker[1], which documents real-world penalties and compliance failures under European data protection law.

- **EU Financial Regulation Dataset**: We incorporate the EU financial regulation subset from OmniCompliance-100K (Hu et al., 2026), a real-world, multi-domain legal case dataset. Our sampled subsets include Cross-border Transaction Law, Crypto-asset Law, and Electronic Money Law.

- **Chinese Data Law Dataset**: We further include the Chinese data law subset from OmniCompliance-100K (Hu et al., 2026), covering the Data Security Law, the Cybersecurity Law, and the Personal Information Protection Law.

[1]https://www.enforcementtracker.com/

3.2 Out-domain Evaluation Datasets.

To assess our model's generalization in legal quantitative prediction, we use an out-of-domain evaluation set drawn from several datasets:

- **LawBench** (Fei et al., 2024). A comprehensive Chinese legal benchmark that assesses model performance across diverse legal tasks such as judgment prediction, charge classification, and penalty term estimation. We specifically select the criminal amount estimation subset, which requires predicting the monetary sums involved in criminal cases.

- **CAIL-2018** (Xiao et al., 2018): A dataset containing 2.6 million real criminal court cases from an official Chinese website. We sample 600 data instances from its test set.

- **LexNum** (Zhang et al., 2025): A dataset of compensation cases grounded in Chinese law, specifically economic compensation from the Labor Contract Law of the People's Republic of China, work injury compensation from the Work Injury Insurance Regulations, and traffic accident compensation.

4 Experimental Settings

In this section, we describe our experimental setup, covering the LLMs employed, training configurations, and evaluation settings.

4.1 Models

Models for Training. We train GRPO (Shao et al., 2024) on Qwen model family, including Qwen2.5-7B (Qwen et al., 2025), Qwen3-1.7B, Qwen3-4B, and Qwen3-8B (Yang et al., 2025).

Model	Chinese (Ratio/Accuracy)	Finance (Ratio/Accuracy)	Avg. (Ratio/Accuracy)
DeepSeek-V4	30.39 / 21.08	22.46 / 8.22	26.42 / 14.65
GPT-4o-mini	10.92 / 1.45	17.19 / 2.59	14.06 / 2.02
GPT-5-mini	33.98 / 12.50	38.08 / 8.72	36.03 / 10.61
GPT-5.4	26.89 / 13.14	24.04 / 7.59	25.47 / 10.36
GPT-5.5	36.67 / 18.39	29.43 / 6.39	33.05 / 12.39
Claude-Opus-4.7	**39.24** / **24.45**	33.90 / 7.90	36.57 / **16.18**
Gemini-3.1-Pro	26.06 / 13.91	11.12 / 4.36	18.59 / 9.13
Qwen3-1.7B	1.61 / 0.00	2.05 / 0.19	1.83 / 0.09
+ GRPO	$25.97_{(+24.36)}/4.01_{(+4.01)}$	$29.94_{(+27.89)}/4.80_{(+4.61)}$	$27.95_{(+26.12)}/4.40_{(+4.31)}$
Qwen3-4B	5.60 / 0.00	5.54 / 0.19	5.57 / 0.09
+ GRPO	$35.10_{(+29.50)}/2.90_{(+2.90)}$	$42.88_{(+37.34)}/6.60_{(+6.41)}$	$38.99_{(+33.42)}/4.75_{(+4.66)}$
Qwen2.5-7B	5.50 / 2.90	7.02 / 0.38	6.26 / 1.64
+ GRPO	$34.99_{(+29.49)}/2.56_{(-0.34)}$	$43.67_{(+36.65)}/\mathbf{11.98}_{(+11.60)}$	$39.33_{(+33.07)}/7.27_{(+5.63)}$
Qwen3-8B	7.89 / 1.45	7.11 / 1.53	7.50 / 1.49
+ GRPO	$35.79_{(+27.90)}/14.42_{(+12.97)}$	$\mathbf{44.70}_{(+37.59)}/11.34_{(+9.81)}$	$\mathbf{40.25}_{(+32.74)}/12.88_{(+11.39)}$

Table 3: Fine prediction results on $y_{\text{gt}} \neq 0$ data across in-domain datasets.

Model	Chinese	Finance	Avg.
DeepSeek-V4	97.05	99.14	98.09
GPT-4o-mini	94.82	95.56	95.19
GPT-5-mini	92.08	96.91	94.49
GPT-5.4	95.69	99.35	97.52
GPT-5.5	96.56	99.57	98.06
Claude-Opus-4.7	93.36	98.80	96.08
Gemini-3.1-Pro	88.63	62.06	75.34
Qwen3-1.7B	48.97	43.56	46.27
+ GRPO	$\mathbf{99.11}_{(+50.14)}$	$\mathbf{99.78}_{(+56.23)}$	$\mathbf{99.45}_{(+53.18)}$
Qwen3-4B	35.30	40.39	37.85
+ GRPO	$96.63_{(+61.32)}$	$98.54_{(+58.15)}$	$97.58_{(+59.74)}$
Qwen2.5-7B	62.03	62.54	62.28
+ GRPO	$94.96_{(+32.93)}$	$97.90_{(+35.37)}$	$96.43_{(+34.15)}$
Qwen3-8B	63.29	64.91	64.10
+ GRPO	$97.12_{(+33.82)}$	$99.01_{(+34.10)}$	$98.07_{(+33.96)}$

Table 4: Fine prediction accuracy results on $y_{\text{gt}} = 0$ data across in-domain datasets.

Models for Evaluation. Besides, we compare our models with wide range of closed-source and open-source LLMs, including DeepSeek-V4-Pro (DeepSeek-AI, 2026), GPT-5.5 (OpenAI, 2025b), GPT-5-mini (OpenAI, 2025a), GPT-4o-mini (OpenAI et al., 2024), Claude-Opus-4.7 (Anthropic, 2025), and Gemini-3.1-Pro (Google, 2025).

4.2 Training Settings

RL Training Setting. For the reward model Power-Ratio Reward, as shown in Equation 4, we set the decay rate control parameter $\alpha = 2.0$ and the stability constant $\epsilon = 10^{-3}$. For rollout group size in GRPO, as defined in Equation 3, we set $G = 5$.

Hyper-parameters for LLM Training. The training batch size is 1024, with a per-GPU batch size of 8. The maximum response length is set to 1024 tokens. The learning rate is 1×10^{-6}, and the KL coefficient is 1×10^{-3}.

4.3 Evaluation Tasks and Metrics

For the datasets shown in Table 1, the evaluation tasks span *"fine prediction"*, *"criminal amount prediction"*, and *"compensation prediction"*. For all tasks, the goal is to predict a numerical amount with a given data sample with case background. Particularly, $y_{\text{gt}} = 0$ data represents case data samples that are compliant and not punished. We shown several examples in Figure 1, with additional examples for each of the collected datasets demonstrated in Table 5 and Table 6, which is located at the Appendix section.

For evaluation metrics, we leverage the accuracy of extract match for numerical prediction and ground truth, which can be formulated as

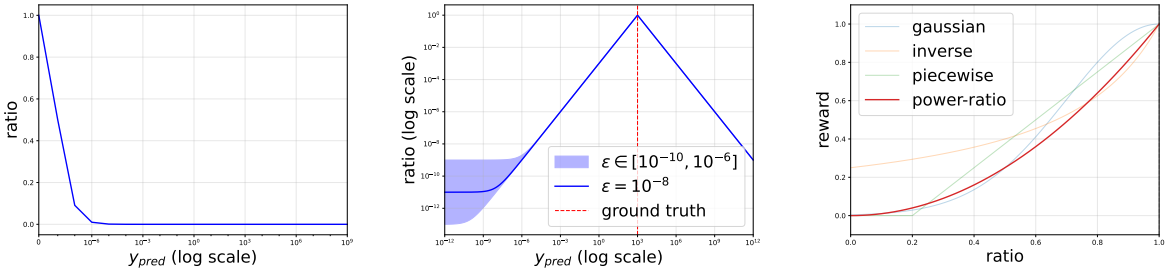
$$\text{accuracy} = \mathbb{I}\{y_{\text{pred}} = y_{\text{gt}}\}. \tag{5}$$

Besides, we specifically design an evaluation metric for legal numerical prediction, named *"ratio"*:

$$\text{ratio} = \frac{\min(y_{\text{pred}}, y_{\text{gt}}) + \epsilon}{\max(y_{\text{pred}}, y_{\text{gt}}) + \epsilon}, \tag{6}$$

where ϵ prevents division by zero and is set to 1×10^{8}. This new metric measure the relative distance between ground truth and prediction, ranging in $(0, 1]$.

Effectiveness of the Metric Ratio. We further analyze the effectiveness of the proposed metric, "ratio." First, in Figure 3a, we illustrate the relationship between the ratio and y_{pred} when $y_{\text{gt}} = 0$. The ratio equals 1 only when $y_{\text{pred}} = 0$; otherwise, it decreases monotonically and sharply as y_{pred} increases. In particular, the ratio drops to 10^{-8} when $y_{\text{pred}} = 10^{-1}$. Besides, in Figure 3b, we shows the y_{pred}-ratio curve when $y_{\text{gt}} \neq 0$. The ratio decreases as the discrepancy between y_{pred} and y_{gt} becomes larger.



(a) y_{pred}-ratio curves when $y_{gt} = 0$. (b) y_{pred}-ratio curves when $y_{gt} \neq 0$. (c) ratio-reward curves.

Figure 3: y_{pred}-ratio curves and ratio-reward curves.

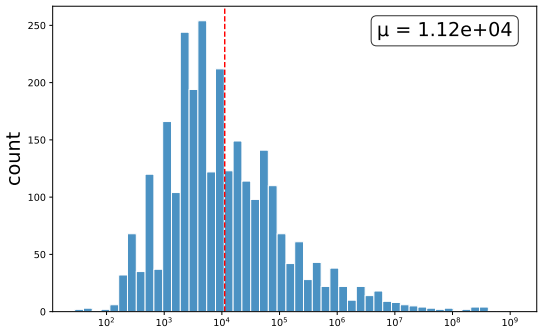


Figure 4: The fine distribution over GDPR cases.

5 Experiments

In this section, we demonstrate our experimental results and draw several key findings. Additionally, we show GRPO training curves in Figure 7

5.1 In-domain Results

(1) Our GRPO-based training framework leads to substantial and consistent performance improvements across all trained models and datasets. We demonstrate main results of our experiments in Table 2. Specifically, Qwen3-1.7B-GRPO improves from 22.73/21.06 to 62.74/47.30 in average ratio/accuracy, corresponding to gains of +40.0 and +26.2, respectively. On GDPR, its performance increases from 4.59/0.68 to 36.35/6.29 (+31.8/+5.6); on the Chinese dataset, from 37.77/37.41 to 81.84/76.51 (+44.1/+39.1); and on Finance, from 25.84/25.08 to 70.04/59.11 (+44.2/+34.0). Qwen3-4B-GRPO exhibits the largest overall improvement, increasing from 18.84/16.74 to 66.87/46.82 (+48.0/+30.1). Its GDPR performance rises from 2.85/0.17 to 43.34/6.80 (+40.5/+6.6), Chinese performance from 28.24/26.91 to 82.10/74.27 (+53.9/+47.4), and Finance performance from 25.43/23.15 to 75.18/59.37 (+49.8/+36.2). Similarly, Qwen2.5-7B-GRPO improves from 31.05/28.09 to 66.15/47.04 (+35.1/+19.0), including gains from 6.27/0.68 to 42.56/6.46 (+36.3/+5.8) on GDPR, from 48.52/47.95 to 80.66/73.01 (+32.1/+25.1) on Chinese, and from 38.35/35.63 to 75.21/61.65 (+36.9/+26.0) on Finance. Finally, Qwen3-8B-GRPO improves from 31.92/28.86 to 67.81/49.04 (+35.9/+20.2). Its GDPR results increase from 5.59/0.34 to 44.80/7.82 (+39.2/+7.5), Chinese results from 50.06/48.54 to 82.52/77.17 (+32.5/+28.6), and Finance results from 40.11/37.71 to 76.12/62.14 (+36.0/+24.4). These results demonstrate that GRPO consistently delivers large gains across all datasets, particularly on the Chinese and Finance benchmarks.

(2) Our GRPO-trained models also become highly competitive with, and in several cases surpass, SOTA models. In terms of average ratio, Qwen3-8B-GRPO achieves the best overall result of 67.81, outperforming GPT-5.5 (67.29), GPT-5.4 (64.10), Claude-Opus-4.7 (64.05), and DeepSeek-V4-Pro (61.95). On the Finance benchmark, Qwen3-8B-GRPO achieves the highest overall performance among all compared methods, reaching 76.12 ratio and 62.14 accuracy, surpassing all proprietary baselines. On the Chinese benchmark, Qwen3-8B-GRPO achieves 82.52 ratio and 77.17 accuracy, nearly matching or exceeding the strongest closed-source models, including GPT-5.5 (82.54/78.07) and Claude-Opus-4.7 (80.46/76.80). Furthermore, Qwen3-4B-GRPO and Qwen2.5-7B-GRPO also achieve highly competitive average ratio scores of 66.87 and 66.15, respectively, approaching or surpassing several frontier proprietary systems. These findings demonstrate that our GRPO training strategy can effectively unlock strong reasoning and compliance capabilities in relatively small open-source models, enabling them to rival or even exceed much larger LLMs.

5.2 Analysis on $y_{gt} = 0$ and $y_{gt} \neq 0$ Cases

Additionally, we present detailed results for cases where the ground truth equals 0 ($y_{gt} = 0$), as shown

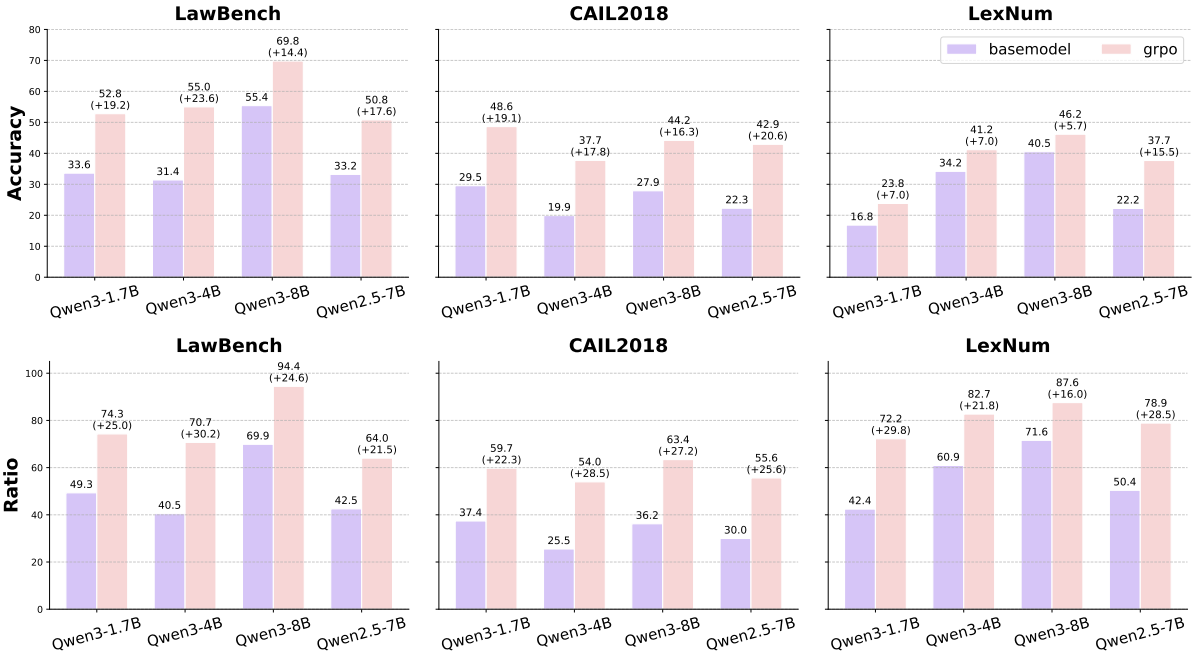


Figure 5: Experimental results across 3 out-domain datasets. Our models maintain improvements over all settings.

in Table 4, and where the ground truth is nonzero ($y_{gt} \neq 0$), as shown in Table 3, from which we draw the following findings.

(3) On $y_{gt} = 0$ data, our GRPO-trained models can achieve nearly perfect prediction accuracy. Specifically, all GRPO variants obtain average accuracies above 96%, substantially outperforming their corresponding base models. Among them, Qwen3-1.7B-GRPO achieves the best overall performance with an average accuracy of 99.45, improving significantly from the original 46.27. Similar gains are observed for Qwen3-4B-GRPO (37.85 $\rightarrow$ 97.58), Qwen2.5-7B-GRPO (62.28 $\rightarrow$ 96.43), and Qwen3-8B-GRPO (64.10 $\rightarrow$ 98.07). In addition, proprietary models such as DeepSeek-V4 and GPT-5.5 also achieve strong results with average accuracies of 98.09 and 98.06, respectively. These results suggest that the $y_{gt} = 0$ subset is comparatively easy to fit, and GRPO training enables the models to almost perfectly capture the underlying patterns in such cases.

(4) The 1.7B small model tends to overfit to the $y_{gt} = 0$ subset, yielding relatively low scores on the more challenging $y_{gt} \neq 0$ subset. Although Qwen3-1.7B-GRPO achieves near-perfect accuracy of 99.45 on $y_{gt} = 0$, its performance on $y_{gt} \neq 0$ remains limited, reaching only 27.95/4.40 of ratio/accuracy on average. In contrast, larger models exhibit substantially stronger generalization ability on the harder non-zero cases. For example, Qwen3-8B-GRPO achieves the best open-source performance with 40.25/12.88, while Qwen2.5-7B-GRPO reaches 39.33/7.27. For comparison, Claude-Opus-4.7 attains the strongest overall result with 36.57/16.18, followed by GPT-5.5 with 33.05/12.39. These observations suggest that the smallest 1.7B model may memorize simpler $y_{gt} = 0$ patterns during GRPO training, but lacks sufficient capacity to generalize effectively to more complex $y_{gt} \neq 0$ instances.

5.3 Out-domain Generalization

To examine the generalization capability of our method, we additionally evaluate our models on three out-domain datasets, including LawBench, CAIL-2018, and LexNum, where their tasks span criminal amount prediction, fine prediction, and compensation prediction, respectively.

(5) Our models exhibit strong generalization ability on out-of-domain legal benchmarks, consistently improving performance across all three datasets. On LawBench, the GRPO-trained models achieve substantial gains over the corresponding base models, with ratio improvements ranging from +22.3 to +28.5 points and accuracy improvements ranging from +16.3 to +20.6 points across different model scales. Similar performance trends are observed on CAIL2018, where the ratio metric is consistently improved by more than +22 points, accompanied by significant accuracy gains of up to +20.6 points, indicating the robustness of the proposed approach when transferred to external legal reason-

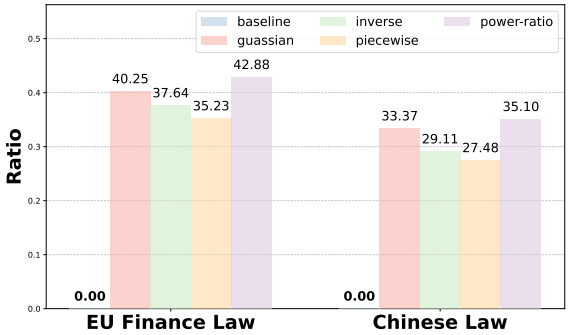


Figure 6: Ratio results on $y_{gt} \neq 0$ data over Qwen3-4B-GRPO models trained with different reward functions.

ing tasks. On LexNum, although several backbone models already exhibit relatively strong baseline performance, our method still yields notable improvements, increasing the ratio metric by +16.0 $\sim$ +29.8 points and the accuracy metric by +5.7 $\sim$ +15.5 points. Overall, these results suggest that the proposed training strategy effectively enhances the cross-domain generalization capability of LLMs and enables consistent performance improvements across diverse multilingual legal datasets.

5.4 Alternative Reward Function Designs

To further evaluate the advancement of the power-ratio reward design, we compare it with several alternative reward functions. These reward designs are introduced below, and their corresponding reward curves are illustrated in Figure 3c.

- **Baseline Accuracy Reward.** This reward corresponds to standard accuracy and provides no partial credit for approximately correct predictions. The accuracy reward can be formulated as $R_{\text{acc}}(y_{\text{pred}}, y_{\text{gt}}) = \mathbb{I}\{y_{\text{pred}} = y_{\text{gt}}\}$.

- **Gaussian Reward.** This reward adopts a Gaussian-shaped similarity function. The reward is: $R_{\text{gaussian}}(y_{\text{pred}}, y_{\text{gt}}) = \exp\left(-\frac{(y_{\text{pred}}-y_{\text{gt}})^2}{2\sigma^2}\right)$, $\sigma = \rho(y_{\text{gt}} + \varepsilon)$. The parameter $\rho = 0.2$ determines the relative spread of the Gaussian kernel through $\sigma = \rho(y_{\text{gt}} + \varepsilon)$, where $\varepsilon = 10^{-8}$ ensures computation stability.

- **Inverse Relative-Error Reward.** This reward uses an inverse proportional mapping. The reward $R_{\text{inverse}}(y_{\text{pred}}, y_{\text{gt}}) = \frac{1}{1+\alpha\left|\frac{y_{\text{pred}}-y_{\text{gt}}}{y_{\text{gt}}+\varepsilon}\right|}$, where $\alpha = 3.0$ controls the penalty strength and $\varepsilon = 10^{-8}$ ensures computation stability.

- **Piecewise Linear Reward.** This reward introduces a linearly decreasing score with a hard cutoff threshold. The formula is $R_{\text{piecewise}}(y_{\text{pred}}, y_{\text{gt}}) = \max\left(0, 1 - \frac{1}{\tau}\left|\frac{y_{\text{pred}}-y_{\text{gt}}}{y_{\text{gt}}+\varepsilon}\right|\right)$, where $\tau = 0.8$ denotes the normalized error threshold and $\varepsilon = 10^{-8}$ ensures computation stability.

We train our models on the Eu Finance Law and Chinese Law dataset with base model Qwen3-4B. We illustrate the ratio results on $y_{gt} \neq 0$ data in Figure 6 and draw the following observations.

(6) The baseline model, which takes a standard accuracy reward, consistently fails on data points where the ground truth $y_{gt} \neq 0$. This is evidenced by its zero ratio across EU Finance Law and Chinese Law datasets, as shown in Figure 6.

(7) In contrast, the power-ratio reward function achieves the best overall results, outperforming all other rewards. Specifically, under the EU Finance Law dataset, the power-ratio yields the highest ratio at 42.88, significantly surpassing the baseline (0.00), Gaussian (35.23), inverse (29.11), and piecewise (37.64) approaches. Similarly, in the Chinese Law dataset, the power-ratio again achieves the top performance with a ratio of 35.10, compared to the baseline (0.00), Gaussian (27.48), inverse (33.37), and piecewise (35.10). These results demonstrate that the power-ratio reward is robust and highly effective in handling non-zero ground truth labels.

6 Conclusion

We present a reinforcement learning (RL) approach for legal quantitative prediction that significantly improves the accuracy and reliability of numerical estimates in regulatory and judicial contexts. Specifically, we introduce the Power-Ratio Reward, a reward model that computes the ratio between the predicted and ground-truth values, providing a soft measure of their relative differences. Our RL model achieves substantial gains over the baseline and generalizes effectively to three out-domain benchmarks. These results demonstrate the potential of specialized legal LLMs to support regulatory compliance, judicial decision-making, and financial risk assessment, marking an important step toward more trustworthy and practical legal AI systems.

Limitation

Due to computational resource limitations, we have not scaled our method to larger models. Nevertheless, we achieve significant improvements across various legal quantitative prediction tasks. We encourage researchers and the community to scale up the experiments to explore even more promising results.

Ethical Considerations

We affirm that all authors of this paper acknowledge the ACL Code of Conduct.

Potential Risks. Our collected datasets pertain to data security, financial crime, and personal information protection laws, and may contain sensitive information. All original dataset sources have already filtered out sensitive data and performed anonymization. We further reviewed the datasets for any remaining personally identifiable information (PII) and removed several cases as necessary. As a result, it is safe to release our collected datasets.

References

Anthropic. 2025. Claude opus 4.7. https://www.anthropic.com/claude. Large language model developed by Anthropic.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

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.

Zhiwei Fei, Songyang Zhang, Xiaoyu Shen, Dawei Zhu, Xiao Wang, Jidong Ge, and Vincent Ng. 2025. InternLM-law: An open-sourced Chinese legal large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9376–9392, Abu Dhabi, UAE. Association for Computational Linguistics.

Google. 2025. Gemini 3 pro. https://gemini.google.com/. Large language model developed by Google DeepMind.

Wenbin Hu, Huihao Jing, Haochen Shi, Changxuan Fan, Haoran Li, and Yangqiu Song. 2026. Omnicompliance-100k: A multi-domain, rule-grounded, real-world safety compliance dataset. *Preprint*, arXiv:2603.13933.

Kangcheng Luo, Quzhe Huang, Cong Jiang, and Yansong Feng. 2025. Automating legal interpretation with LLMs: Retrieval, generation, and evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4015–4047, Vienna, Austria. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025a. Gpt-5 mini. https://openai.com/chatgpt. Large language model developed by OpenAI.

OpenAI. 2025b. Gpt-5.5. https://openai.com/chatgpt. Large language model developed by OpenAI.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4882–4889, New York, NY, USA. Association for Computing Machinery.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.

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, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Kepu Zhang, Guofu Xie, Weijie Yu, Mingyue Xu, Xu Tang, Yaxin Li, and Jun Xu. 2025. Legal mathematical reasoning with LLMs: Procedural alignment through two-stage reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1586–1598, Suzhou, China. Association for Computational Linguistics.

A Data Examples

We provide sample data from each of the datasets used in our experiments, as presented in Table 5 for in-domain datasets and Table 6 for out-of-domain datasets.

B Training Curves

We demonstrate the training process by showing the training curves in Figure 7.

C Computation Resources

In our experiments, we use 8 NVIDIA A800 GPUs to train and evaluate both our models and the baseline models, requiring a total of two weeks of GPU time to complete all experiments. The total cost of API calls was approximately $200 USD.

D Licenses

We utilize the VeRL (Sheng et al., 2024) framework to train the GRPO models, which is licensed under Apache-2.0, along with datasets that are individually licensed: Omni-Compliance (Hu et al., 2026), CAIL-2018 (Xiao et al., 2018), and LexNum (Zhang et al., 2025) under MIT license, and LawBench (Fei et al., 2024) under Apache-2.0 license. We adhere to the respective license terms when using these resources in our work.

GDPR Example. "Case Background": "Scanshare s.r.l., as data processor for the Antonio Cardarelli hospital, managed a platform for a public competition. Due to a configuration error, candidate codes and personal documents were temporarily accessible online, leading to unlawful disclosure of participants' personal data." "Ground Truth": "60,000 €"
Chinese Cyber-security Law Example. "Case Background": "Alibaba Group, founded by Jack Ma, engaged in practices where it required merchants to sign exclusive agreements limiting their business on competing platforms. This occurred amid rapid growth in e-commerce, with Alibaba holding significant market share. The company's 'choose one of two' policy directed merchants to prioritize Alibaba over rivals like JD.com and Pinduoduo. Regulators noted this affected competition in online retail. The situation drew attention after Jack Ma's public criticism of financial regulators in October 2020, leading to scrutiny of Ant Group's planned IPO suspension. Alibaba's overall operations involved extensive data collection and market dominance in digital services across China." "Ground Truth": "18,230,000,000 ¥"
Chinese Data-security Law Example. "Case Background": "Beijing Shangmi Technology Co., Ltd., a provider of blockchain and data services, engaged in cross-border data transfers between 2018 and 2022. The company processed personal data of approximately 100,000 individuals and important data involving national security, exporting it to overseas entities without obtaining necessary security assessments or approvals. This involved data on user identities, transaction records, and network mappings, which were shared with foreign servers and partners for business operations and analytics." "Ground Truth": "11,700,000 ¥"
Chinese Personal Information Protection Law Example. "Case Background": "Meituan, a food delivery and lifestyle services app, collects user addresses, order histories, and payment details provided during service requests. A 2021 audit found the company processed this data for extensive merchant analytics and promotions beyond delivery fulfillment, involving data from hundreds of millions of transactions legally disclosed by users." "Ground Truth": "30,000,000 ¥"
EU Finance Cross Border Payment Law Example. "Case Background": "Euronet Worldwide, operating as a payment service provider in the EU, faced issues in 2018-2019 with its prepaid card and direct debit services under SEPA. The firm processed direct debits without consistently applying the scheme's standards for mandate signing, electronic validation, and creditor reference formatting. This affected services in countries like France and Italy, where customers reported unauthorized collections linked to Euronet's epay and money transfer products. The SEPA scheme mandates specific practices for inter-PSP communication and error handling, which were bypassed in Euronet's high-volume processing systems to accelerate transactions. "Ground Truth": "300,000 €"
EU Finance Cryptocurrency Law Example. "Case Background": "Bitstamp, a Luxembourg-based crypto exchange registered as a CASP under MiCA, introduced a verification protocol in late 2023 for withdrawals over €1,000 to self-hosted addresses. Users were required to demonstrate control via test transactions, signed messages, or integration with wallet ownership proofs before finalizing transfers. This measure addressed the TFR's requirement for adequate assessment of originator ownership, preventing unverified outflows that could facilitate illicit fund movements. The process involved real-time on-chain monitoring and risk scoring, applied across all EU jurisdictions where Bitstamp operated, handling over €10 billion in monthly volume. Bitstamp's approach was developed in response to EBA guidelines and pilot-tested with regulators to ensure alignment with Article 19b risk mitigators." "Ground Truth": "0"
EU Finance Cryptocurrency Law Example. "Case Background": "In 2019, the European Banking Authority issued a statement addressing several platforms offering 'stablecoins' marketed as electronic money without authorization. One prominent case involved a pan-EU operator issuing digital tokens backed by fiat, processing €100 million+ in loads/redemptions, bypassing e-money licensing by claiming non-e-money status despite fitting EU definitions." "Ground Truth": "5,000,000 €"

Table 5: Data examples sampled from in-domain datasets. All tasks are about fine prediction.

LexNum Data Example (Task: Compensation Prediction). "Case Background": " 叶某华在广西某某能源开发有限公司工作了11年4个月，2024年5月被告违法解除劳动合同，且未支付经济补偿。根据《劳动合同法》第四十六条和第四十七条，劳动者的经济补偿应按在单位工作的年限和最后12个月的平均工资计算。叶某华的最后12个月的平均工资为1248元。请计算叶某华应获得的经济补偿金额，包括因违法解除劳动合同而支付的赔偿金，工作年限为11.5个月。因为公司存在违法解除劳动关系的行为，所以赔偿要翻倍。 (Translation: Ye Mouhua worked at a certain energy development company in Guangxi for 11 years and 4 months. In May 2024, the defendant illegally terminated the labor contract and did not pay economic compensation. According to Articles 46 and 47 of the Labor Contract Law, the economic compensation for a worker should be calculated based on the number of years worked at the company and the average monthly wage over the last 12 months. Ye Mouhua's average monthly wage for the last 12 months was 1,248 yuan. Please calculate the amount of economic compensation Ye Mouhua should receive, including the damages for the illegal termination of the labor contract, with a work period of 11.5 years. Because the company illegally terminated the labor relationship, the compensation should be doubled.)" "Ground Truth": "16,000 CNY"
CAIL-2018 Data Example (Task: Fine Prediction). "Case Background": " 经审理查明: 2017年8月20日凌晨1时许,被告人李某伙同苏某(另案处理)至义乌市佛堂镇王宅村407号门前,盗得被害人杨某停放在该处的红色杭派牌电动车一辆(价值人民币2520元)。现被盗电动车已被被害人杨某自行找回。上述事实,被告人李某在庭审过程中亦无异议,且有被害人杨某的陈某,证人施某的证言,同案犯苏某的供述,现场辨认笔录及照片,接受证据材料清单,收款收据,价格认定结论书,行政处罚决定书,撤销行政处罚决定书,前科刑事判决书,抓获经过,被告人李某的供述,身份证明等证据证实,足以认定。 (Translation: After trial, it was established that: In the early morning of August 20, 2017, at approximately 1:00 a.m., the defendant Li, together with Su (handled in a separate case), went to No. 407, Wangzhai Village, Fotang Town, Yiwu City, and stole a red Hangpai-brand electric bicycle belonging to the victim Yang, which was parked there (valued at RMB 2,520). The stolen electric bicycle has since been recovered by the victim Yang on their own. The above facts were not disputed by the defendant Li during the trial. They are corroborated by the testimony of the victim Yang and witness Shi, the confession of the co-perpetrator Su, the on-site identification records and photos, the list of evidence materials, payment receipts, the price appraisal report, administrative penalty decision, revocation of administrative penalty decision, prior criminal judgment, the record of apprehension, the confession of the defendant Li, identity documents, and other evidence, which are sufficient to establish the facts.)" "Ground Truth": "1,000 CNY"
LawBench Data Example (Task: Criminal Amount Prediction). "Case Background": " 文书:经审理查明，2016年1月期间，被告人刘XX与宋XX、胡某（二人均已判刑）三人商定在瓯阳专线班车上实施盗窃，并约定谁动手便分得所盗财物的一半，其余两人平分另一半，具体犯罪事实如下：1、2016年1月2日11时许，被告人刘XX与宋XX、胡某三人乘坐瓯阳专线班车经过南平市建阳区某镇某路段后，由胡某、刘XX望X，宋XX用刀片将被害人刘某的外衣右边口袋割破，将袋内装有现金1500元的黄色小信封盗走。下车后，宋XX分得赃款750元，胡某、刘XX两人各分得375元赃款。2、2016年1月20日11时许，被告人刘XX与宋XX、胡某三人乘坐瓯阳专线班车至某路段时，见被害人张某1睡着，遂由胡某、刘XX望X，宋XX用刀片将张某1外衣口袋割破，将口袋内用牛筋捆着的现金7000元盗走。事后宋XX分得3500元，胡某、刘XX各分得1750元。 (Translation: Document: After trial, it was determined that in January 2016, the defendant Liu XX, together with Song XX and Hu (both of whom have already been sentenced), agreed to commit theft on the Ouyang express bus line. They arranged that whoever carried out the theft would receive half of the stolen property, and the remaining half would be equally divided between the other two. The specific criminal acts are as follows: Around 11:00 a.m. on January 2, 2016, the defendant Liu XX, together with Song XX and Hu, were on the Ouyang express bus. After passing through a certain section of a town in Jianyang District, Nanping City, Hu and Liu XX kept watch while Song XX used a blade to cut open the right pocket of the victim Liu's outer jacket and stole a yellow small envelope containing 1,500 yuan in cash. After getting off the bus, Song XX received 750 yuan of the stolen money, while Hu and Liu XX each received 375 yuan. Around 11:00 a.m. on January 20, 2016, the defendant Liu XX, together with Song XX and Hu, were again on the Ouyang express bus. Upon seeing the victim Zhang asleep at a certain section of the road, Hu and Liu XX kept watch while Song XX used a blade to cut open Zhang's jacket pocket and stole 7,000 yuan in cash, tied with beef tendon. Afterwards, Song XX received 3,500 yuan, while Hu and Liu XX each received 1,750 yuan.)" "Ground Truth": "8,500 CNY"

Table 6: Data examples sampled from out-domain datasets.

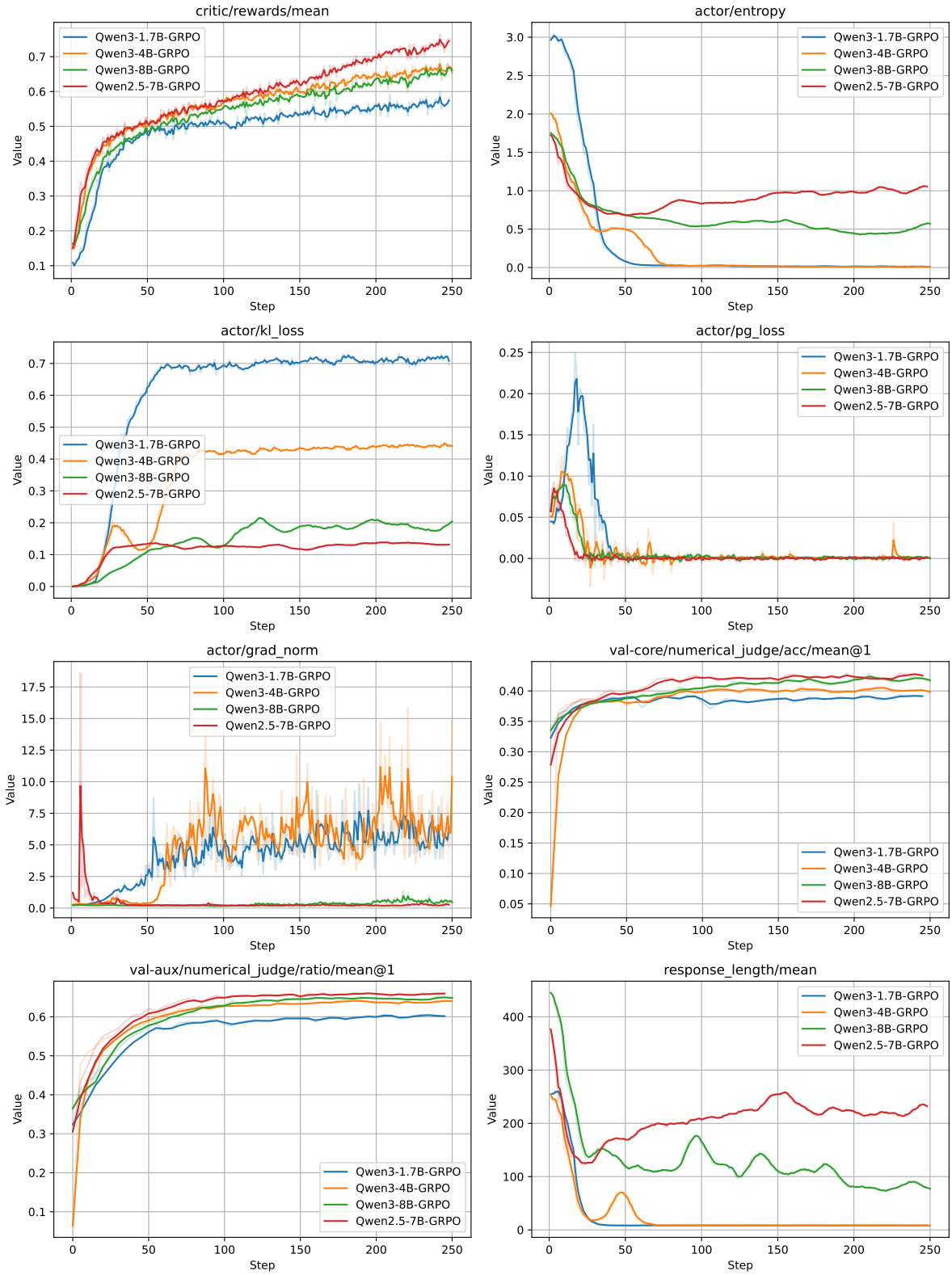


Figure 7: GRPO Training curves across models with difference scales.