Cold Starts and Hard Cases: A Two-Stage SFT-RLVR Approach for Legal Machine Translation (Just-NLP L-MT shared task)

Pawitsapak Akarajaradwong[†]
VISAI AI, Thailand
pawitsapaka_visai@vistec.ac.th

Chompakorn Chaksangchaichot[†]
VISAI AI, Thailand
chompakornc_pro@vistec.ac.th

Abstract

This paper details our system for the JUST-NLP 2025 Shared Task on English-to-Hindi Legal Machine Translation. We propose a novel two-stage, data-centric approach. First, we annotate the training data by translation difficulty and create easy and hard subsets. We perform SFT on the easier subset to establish a robust "cold start". Then, we apply RLVR exclusively on the harder subset, using machine translation metrics as reward signals. This strategy allowed our system to significantly outperform strong baselines, demonstrating the capability of our systems for machine translation tasks. Source code and model weights are available at https://github.com/ppaolong/FourCorners-JustNLP-MT-Shared-Task

1 Introduction

The Indian legal system presents a compelling machine translation challenge, requiring the translation of complex, jargon-heavy English into accessible Hindi to ensure judicial transparency and access to justice. The standard approach, Supervised Fine-Tuning (SFT), is suboptimal for this task as it tends to overfit by memorizing reference translations and inefficiently treats all training examples equally.

To overcome these challenges, we propose a hybrid SFT-RLVR pipeline guided by a data curriculum. We first employ an external LLM to annotate the training data by translation difficulty. A robust baseline model is then established via SFT on the "easy-to-medium" subset. Finally, we apply Reinforcement Learning with Verifiable Rewards (RLVR) exclusively on the "hard" subset, using the competition's MT evaluation metrics (BLEU, ROUGE, ChrF++) as direct, low-cost reward signals.

Our contributions are threefold:

1. We present a top-performing system for the JUST-NLP 2025 L-MT shared task.
2. We introduce a practical and effective data curriculum strategy that uses an LLM to segment data by difficulty for a hybrid SFT-RLVR training pipeline.
3. We provide empirical evidence that this approach leads to superior performance compared to standard SFT baselines in the specialized legal domain.

2 Related Work

2.1 Reinforcement Learning with Verifiable Rewards

While Supervised Fine-Tuning (SFT) is a standard baseline, reinforcement learning (RL) offers a compelling alternative to move beyond the limitations of token-level mimicry and improve model generalization. The traditional RLHF pipeline, often using Proximal Policy Optimization (PPO) (Schulman et al., 2017), is bottlenecked by its reliance on a separately trained value model. This has motivated the development of simpler, value-model-free alternatives like Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which computes advantages by comparing rewards across multiple samples. Subsequent refinements like Dr. GRPO (Liu et al., 2025), GSPO (Zheng et al., 2025), and DAPO (Yu et al., 2025) have further improved the stability and efficiency of this paradigm.

2.2 Reference-Based Metrics as Rewards for Text Generation

These RLVR frameworks make it practical to use automated reference-based metrics directly as reward signals. A significant advance was demonstrated by Chang et al. (2025), who showed that using BLEU (Papineni et al., 2002) as a direct reward for GRPO can match the performance of complex, human-trained reward models for general instruction following. This "metric-as-reward" principle has been effective in specialized domains like legal

[†]These authors contributed equally as co-first authors.

Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025), pages 101–106
December 24, 2025 ©2025 Association for Computational Linguistics

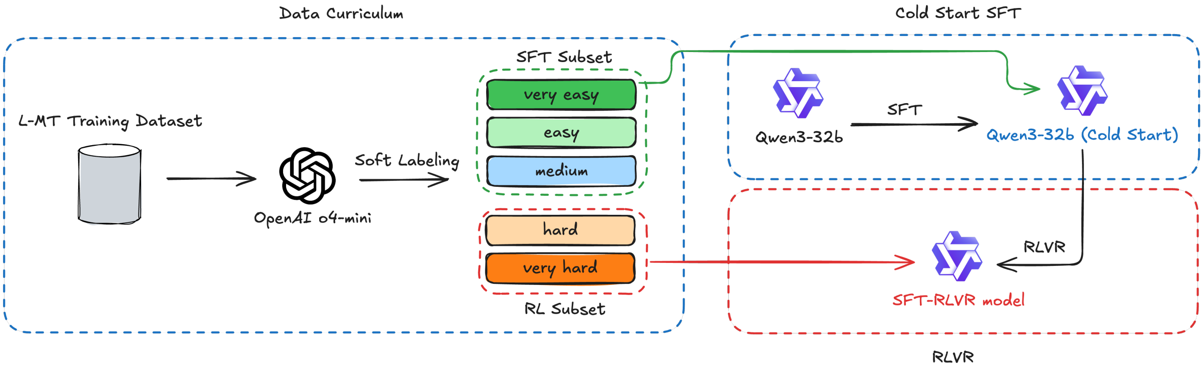

Figure 1: Overview of our proposed system.

question answering (Akarajaradwong et al., 2025). Concurrent to our work, Wang et al. (2025) also employ a two-stage SFT-RL pipeline for literature translation. However, their approach differs significantly, targeting subjective "free translation" and relying on a complex, LLM-as-a-judge (DeepSeek-v3 (DeepSeek-AI et al., 2025)) for its reward signal. Our work is situated within this context, but we apply a data-centric curriculum and use simple, verifiable MT metrics (including ROUGE (Lin, 2004) and ChrF++ (Popović, 2017)) to achieve high-fidelity translation in the precise legal domain.

3 Methodology

Our system employs a two-stage, data-centric pipeline designed to directly optimize a Qwen3-32B model (Yang et al., 2025) for the translation task, as depicted in Figure 1. The process involves meticulous data preparation followed by a hybrid training strategy of Supervised Fine-Tuning (SFT) and Reinforcement Learning with Verifiable Rewards (RLVR).

3.1 Data Preparation and Curriculum

First, we subjected the provided corpus to a comprehensive preprocessing pipeline. This involved normalizing the text by decoding HTML entities, standardizing quotation marks, and replacing ligatures. We then standardized the structure by correcting punctuation spacing, unifying list markers, fixing hyphenation, and collapsing all whitespace.

Inspired by the idea that training should focus on progressively harder examples (Ji et al., 2025), we created a data curriculum. Using o4-mini*, we annotated the 50,000 sentence pairs with five difficulty levels (detailed in Table 1). We then designated the 43,181 "easy-to-medium" pairs as the **SFT Subset** for initial fine-tuning, and the remaining 6,819 "hard" pairs as the **RL Subset** for subsequent reinforcement learning.

Difficulty Level	Count
SFT Subset	
very_low	4167
low	18860
medium	20154
RL Subset	
high	6812
very_high	7

Table 1: Distribution of difficulty labels in the training data.

3.2 Hybrid Training Pipeline

Our training process unfolds in two distinct stages. First, we perform a Cold-Start SFT on the SFT Subset. This efficiently adapts the base model to the legal domain's vocabulary and style using a standard cross-entropy objective, creating a strong foundation. Subsequently, we apply Metric-Driven RLVR exclusively on the RL Subset, using the model from Stage 1 as our policy. In this phase, we update the model to directly maximize rewards derived from standard MT evaluation metrics.

3.3 Reward Function Formulation

The reward signal for the RLVR stage is a weighted sum of primary and auxiliary components.

Primary Metric Rewards: The core of our reward signal was derived from a combination of MT evaluation metrics:

- **BLEU** (Papineni et al., 2002) [0.0-1.0]: Measures n-gram precision for fluency and ade-

*o4-mini-2025-04-16

quacy.

- **Composite ROUGE** (Lin, 2004) [0.0-1.0]: The average F1-score of ROUGE-1, ROUGE-2, and ROUGE-L, providing a comprehensive recall-oriented signal.
- **ChrF++** (Popović, 2017) [0.0-1.0]: A character-level metric robust to morphological variation.[†]

Auxiliary Quality Rewards: To ensure the model produced well-formed outputs, we included two auxiliary rewards:

- **Format Check Reward** [0, 1]: A binary reward that penalizes outputs that do not follow the expected format.
- **Allowed Character Reward** [0, 1]: A binary reward that penalizes the generation of invalid characters in the target Hindi script.

4 Experimental Setup

Statistic	English (Source)	Hindi (Target)
Number of Sentences	50,000	50,000
Total Tokens	1,492,721	1,560,783
Average Sentence Length (Tokens)	29.9	31.2
Median Sentence Length (Tokens)	28.0	29.0
Max Sentence Length (Tokens)	79	71
Type-Token Ratio (TTR)	0.02	0.02

Table 2: Key statistics of the L-MT training dataset.

4.1 Dataset and Evaluation Metrics

All experiments use the official dataset: WMT25 Legal Domain Test Suite (Singh et al., 2025), strictly adhering to the competition's rule of no external data. The training set of 50,000 English-Hindi pairs is characterized by long, complex sentences (avg. 30 tokens) and a highly specialized, repetitive vocabulary (TTR of 0.02), as detailed in Table 2. This dual challenge of syntactic complexity and lexical precision motivated our two-stage approach. Model performance was ranked on the held-out test set using a combined score of the official metrics: BLEU, ROUGE-L, and ChrF++.

4.2 Models

4.2.1 Baselines

We benchmarked our two-stage SFT-RLVR systems against two strong baselines, all using Qwen3-32B (Yang et al., 2025) as the backbone:

- **Base Model:** The pre-trained Qwen3-32B without any fine-tuning.
- **Full SFT:** A strong baseline fine-tuned on the entire 50,000-pair training set.

4.2.2 Our Proposed Systems

Our proposed systems follow the two-stage pipeline described in Section 2. Due to the no-external-data rule, we did not synthesize chain-of-thought and generated translations directly, unlike concurrent work (Wang et al., 2025). We experimented with four primary reward signals for the RLVR phase:

1. **SFT-RLVR-BLEU:** Uses only the BLEU score as a reward signal.
2. **SFT-RLVR-ROUGE:** Uses only the ROUGE composite score as the reward signal.
3. **SFT-RLVR-ChrF++:** Uses only the ChrF++ score as the reward signal.
4. **SFT-RLVR-Combined:** Uses an equally weighted average of BLEU, ROUGE, and ChrF++ scores as the primary reward signal.

Auxiliary Quality Rewards (from Section 3.3) were consistently applied on our proposed systems to ensure the generation of well-formed outputs.

Parameter	SFT	RLVR
Base Model	Qwen3-32b	Qwen3-32b (cold start)
Quantization	4-bit	4-bit
Optimizer	AdamW	AdamW
Learning Rate	1e-4	1e-4
Batch Size	32	1
Rollout	-	32
Numbers of Epochs	3	1
Scheduler	cosine	cosine
LoRA Rank	256	1
LoRA α	256	1
Loss Type	cross-entropy	DAPO

Table 3: Key hyperparameters for the Supervised Fine-Tuning (SFT) and Reinforcement Learning with Verifiable Rewards (RLVR) stages. Other parameters use its default value from trl (von Werra et al., 2020).

4.3 Implementation Details

We implemented our pipeline using the unsloth library (Daniel Han and team, 2023) for efficient training, with stanza (Qi et al., 2020) for tokenization and trl's (von Werra et al., 2020) implementation of the DAPO algorithm (Yu et al., 2025) for the RL phase. All models were fine-tuned using LoRA (Hu et al., 2021), with key hyperparameters detailed in Table 3. All models were trained on A100 80GB GPU.

[†] The standard ChrF++ score (0-100) is normalized to a [0, 1] range.

Model	BLEU	ROUGE-L	ChrF++	Joint Score
Qwen3-32b (baseline)	0.114	0.346	0.405	0.865
Qwen3-32b Full SFT (strong baseline)	0.446	0.624	0.708	1.778
Qwen3-32b Cold-Start SFT	0.440	0.618	0.702	1.760
+RLVR-BLEU (best system)	**0.501**	**0.657**	**0.742**	**1.900**
+RLVR-ROUGE	0.475	0.639	0.720	1.834
+RLVR-ChrF++	0.492	0.654	**0.742**	1.888
+RLVR-Combined	0.495	0.654	0.741	1.890

Table 4: Performance of all models on the test set from codabench submission system. Joint Score is the sum of BLEU, ROUGE-L, and Chrf++ scores. Our best system is highlighted in bold.

For the RLVR stage, we set the LoRA rank and α to 1. This choice is informed by Schulman and Lab (2025) findings that the sparse, per-sequence reward signal from policy gradient methods requires significantly less adapter capacity than the dense, per-token signal of SFT .

5 Results and Discussion

5.1 Overall Performance

The results, presented in Table 4, confirm the superiority of our two-stage SFT-RLVR pipeline (see final official leaderboard in Table 5). All RLVR variants significantly outperform the Full SFT baseline, validating our hybrid approach. Our top-performing system, SFT-RLVR-BLEU, demonstrating that a data-centric curriculum followed by direct metric optimization is a highly effective strategy.

Precision-Oriented Rewards Excel: The primary factor differentiating the performance of our RLVR models was the choice of reward signal. We find a clear advantage for precision-oriented rewards. The top results were achieved by models trained on BLEU and ChrF++, which directly penalize deviations in specific terminology and structure, a critical requirement for maintaining fidelity in the legal domain.

Recall-Oriented Rewards Are Less Effective: On the other hand, the model rewarded with the ROUGE score was our least effective. While this model still beat the baseline, its focus on the overall "gist" of the translation is a poor fit for legal text, where exact wording is critical.

Combining Rewards May Be Suboptimal: Interestingly, using only the BLEU score as a reward was also more effective than combining all three metrics. This suggests that giving the model a single, clear goal for precision works better than a mixed signal.

Our key takeaway is that for a high-stakes field like law, the winning strategy is to directly reward the model for getting the exact words right. We further observe the relative stability of each reward during RLVR phase in Appendix A.

5.2 Ablation Study

Furthermore, an ablation study validates the efficiency of our two-stage design. In Table 4, SFT on Easy Data model alone performs competitively with, though slightly below, the Full SFT baseline. This demonstrates that our "Cold-Start" SFT phase effectively creates a strong foundation. The subsequent targeted RLVR phase not only recovers this minor deficit but elevates the model's performance considerably across all metrics.

6 Conclusion

This paper presents a top-performing system for the JUST-NLP 2025 English-to-Hindi Legal MT shared task. Our approach overcomes the limitations of standard Supervised Fine-Tuning with a two-stage, data-centric pipeline: a Cold-Start SFT on an automatically curated easy subset, followed by Reinforcement Learning with Verifiable Rewards (RLVR) on the harder subset. By directly optimizing for MT metrics as rewards, notably the precision-oriented BLEU, our system significantly outperformed a strong SFT baseline.

Our work validates direct metric optimization, guided by a data curriculum, as a powerful and efficient strategy for developing state-of-the-art systems in specialized domains. Future directions include exploring adaptive reward schemes and applying this methodology to other high-stakes NLP tasks.

Limitation

Our approach has several limitations. First, due to computational and time constraints, we were

Rank Possition	Team Name	Affiliation	Country	BLEU↑	METEOR↑	TER↓	CHRF++↑	BERTScore↑	COMET↑	AutoRank↑	Leaderbord Name
1	Team-SVNIT	Sardar Vallabhbhai National Institute of Technology, Surat	India	51.61	75.8	37.09	73.29	92.61	76.36	61.62	rupeshdhakad06
2	FourCorners	VISAI AI	Thailand	50.19	69.54	42.32	73.67	92.7	75.74	60.31	pawitsapak
3	goodmen	Sardar Vallabhbhai National Institute of Technology, Surat	India	48.56	67.15	41.63	73.07	92.38	75.16	59.39	skdrj123
4	JUNLP	Jadavpur University	India	46.03	71.84	42.08	70.59	91.19	73.72	58.90	iamamit
5	JUST-MEI	SOA University	India	46.67	72.86	44.63	70.03	90.86	72.12	58.79	lsmeetei
6	Lawgorithms	Thangal Kunju Musaliyar College of Engineering	India	46.27	71.8	43.06	68.32	91.03	72.14	58.26	sreehari_saji
7	Tokenizers	Sardar Vallabhbhai National Institute of Technology, Surat	India	34.08	61.78	55.25	56.75	87.39	65.2	50.87	tokenizers

Table 5: Official Leaderboard of JustNLP MT Shared Task (https://exploration-lab.github.io/JUST-NLP/JustNLP25_L-MT_Result.pdf).

unable to explore the potential of other metrics to use as a reward for RLVR such as BERTScore. Our reported results are based on a limited exploration of model hyperparameters and data combinations. Second, our data curriculum's reliance on a proprietary LLM for difficulty annotation impacts the full reproducibility and transparency of our pipeline. Finally, the competition's evaluation protocol, which provided only aggregate scores via the CodaBench platform, precluded a qualitative, example-by-example analysis of our model's improvements over the baselines.

References

Pawitsapak Akarajaradwong, Chompakorn Chaksangchaichot, Pirat Pothavorn, Ekapol Chuangsuwanich, Attapol Rutherford, and Sarana Nutanong. 2025. Aligning LLMs for Thai legal question answering with efficient semantic-similarity rewards. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 304–316, Suzhou, China. Association for Computational Linguistics.

Yapei Chang, Yekyung Kim, Michael Krumdick, Amir Zadeh, Chuan Li, Chris Tanner, and Mohit Iyyer. 2025. Bleuberi: Bleu is a surprisingly effective reward for instruction following. *Preprint*, arXiv:2505.11080.

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

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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.

Yunjie Ji, Sitong Zhao, Xiaoyu Tian, Haotian Wang, Shuaiting Chen, Yiping Peng, Han Zhao, and Xiangang Li. 2025. How difficulty-aware staged reinforcement learning enhances llms' reasoning capabilities: A preliminary experimental study. *Preprint*, arXiv:2504.00829.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

John Schulman and Thinking Machines Lab. 2025. Lora without regret. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/lora/.

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.

Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025. Evaluation of llm for english to hindi legal domain machine translation systems. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 823–833, Suzhou, China. Association for Computational Linguistics.

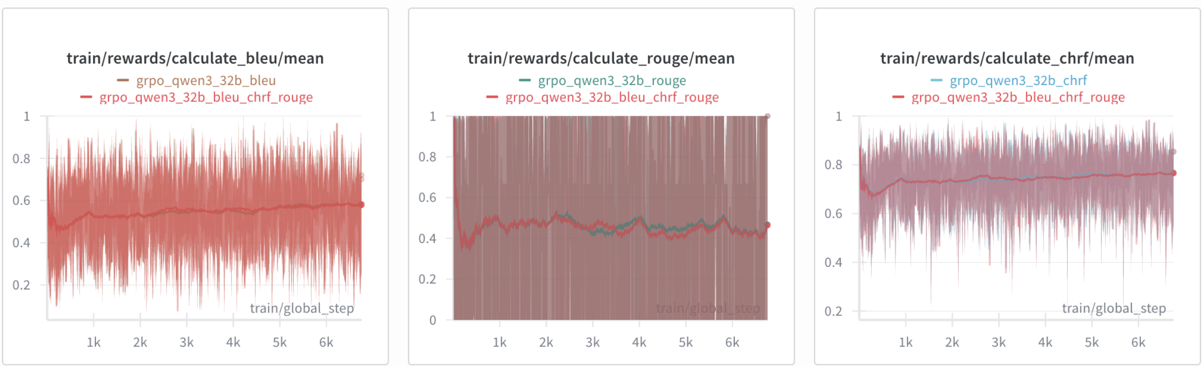

Figure 2: Mean reward signals during the RLVR phase for models trained with single-metric and combined rewards. (Left) The BLEU reward shows a stable, gradual increase. (Center) The composite ROUGE reward is highly erratic and unstable, with no clear upward trajectory. (Right) The ChrF++ reward, similar to BLEU, exhibits a strong and consistent increasing trend. The x-axis represents the training steps, and the y-axis represents the normalized reward score.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025. Deep-trans: Deep reasoning translation via reinforcement learning. *Preprint*, arXiv:2504.10187.

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.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

A Analysis of Reward Signal Stability During RLVR

To provide further insight into the training dynamics of our RLVR models, we plot the mean reward signals for BLEU, our composite ROUGE score, and ChrF++ over the course of training. Figure 2 illustrates the stability and progression of these metrics as rewards.

A key observation from these learning curves is the difference in the stability of the reward signals. As seen in the left and right panels of Figure 2, the mean rewards for BLEU and ChrF++ exhibit a clear and stable upward trend throughout the training process. Although individual batch rewards are noisy (indicated by the wide, faint bands), the smoothed average consistently improves, demonstrating that the model is successfully learning a policy that optimizes for these precision-oriented metrics.

In contrast, the middle panel shows that the composite ROUGE reward is highly unstable. The learning curve is erratic and jagged, with no sustained upward trajectory. This instability suggests that the optimization landscape for a recall-oriented metric like ROUGE is less smooth for this task. The model struggles to find a consistent policy that reliably increases the ROUGE score, possibly because the reward signal is less sensitive to the incremental, precision-focused improvements that are easier for the model to learn.

This empirical observation further supports our main finding in Section 5: that precision-oriented metrics like BLEU and ChrF++ not only lead to better final evaluation scores but also provide a more stable and effective training signal for the RL agent in the high-fidelity legal translation domain.