Knowledge is Still There: Why RL Beats SFT for Sustainable Continual Machine Translation

Anonymous ACL submission

Abstract

Traditional research on Continual Machine Translation (CMT) has long been predicated on the assumption that catastrophic forgetting stems from the irreversible erasure of underlying linguistic knowledge. In this work, we challenge this conventional wisdom within the context of Large Language Models (LLMs). Through systematic behavioral and representational analyses, we demonstrate that the model's latent knowledge often remains intact, yet its ability to align this knowledge with instructions becomes compromised. We attribute this "**expressive drift**" to the ubiquitous Supervised Fine-Tuning (SFT) paradigm. By minimizing token-level loss against fixed references, SFT neglects the inherent plurality of valid translations for a single source, inadvertently coercing the model toward narrow external distributions that conflict with its internal linguistic priors. To rectify this, we advocate for a transition from SFT to a Reinforcement Learning (RL) paradigm, which optimizes performance directly within the model's own generative distribution. Our results show that RL significantly mitigates forgetting in both translation-specific and general tasks by preserving the model's expressive stability. This work redefines the fundamental challenge of CMT and positions RL as a more principled approach for sustainable multilingual alignment in the era of LLMs.

1 Introduction

Continual Machine Translation (CMT) has long been studied as a challenging instance of continual learning (CL) (Wang et al., 2023a), where models must sequentially adapt to new languages, or domains without degrading performance on previously learned ones. Prior research has largely focused on small neural models (Gu et al., 2021; Cao et al., 2021; Wu et al., 2024a). With the recent emergence of LLM-based MT (Xu et al., 2024), the challenges have extended beyond the forgetting of specific translation tasks to include the degradation of in-context learning capabilities (Stap et al., 2024) and broader general abilities (Wu et al., 2025).

To mitigate performance degradation, prior studies have predominantly relied on classical CL approaches (Stap et al., 2024; Wu et al., 2025). These approaches typically assume that catastrophic forgetting equates to the irreversible erasure of underlying knowledge(Shi et al., 2026). This perspective, inherited from scratch-trained small models, overlooks the LLM paradigm where translation is framed as instruction alignment atop a pre-trained backbone (Stap et al., 2024). Consequently, observed forgetting may reflect a misalignment in task execution rather than a structural erasure of knowledge. However, existing CMT research has yet to mechanistically distinguish whether this degradation stems from erased knowledge storage or impaired expressive accessibility.

To resolve this ambiguity, we conduct an empirical study on multilingual CMT with LLMs. We first observe a "generative-discriminative gap": while generative performance (e.g., translation, IFEval) degrades, discriminative benchmarks (e.g., ARC-C, TruthfulQA) remain stable. This observation leads us to hypothesize that forgetting stems from "expressive drift," defined as a failure to align intact internal knowledge with external instructions rather than structural knowledge loss. We verify this hypothesis through a series of multi-faceted analyses. First, few-shot prompting effectively recovers performance, confirming that latent knowledge remains preserved. Second, layer-wise Centered Kernel Alignment (CKA)(Kornblith et al., 2019) and parameter update consistency analysis reveal that representational drift and gradient conflicts are localized in the final layers. These results demonstrate that core representations remain consistent throughout the continual learning process, validating that forgetting is primarily a misalignment in the generative interface.

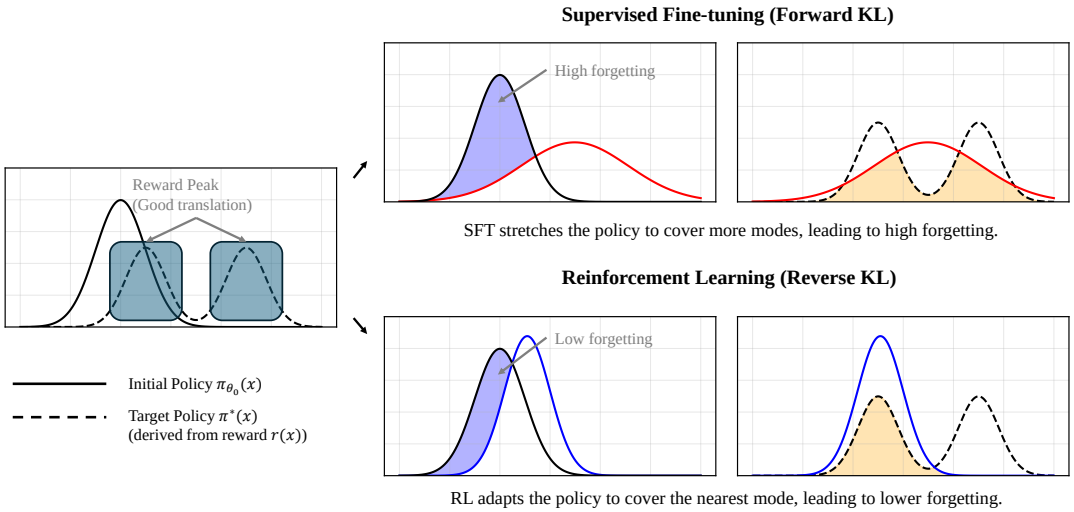


Figure 1: Illustration of different optimization behavior of SFT and RL, under the assumption that multiple high-quality translations (reward peaks) co-exist.

We argue that this phenomenon is closely tied to the dominant supervised fine-tuning (SFT) paradigm adopted by most existing continual learning methods for LLMs. Prior work suggests that SFT mainly teaches models to conform to a desired style and format, rather than fundamentally reshaping their underlying knowledge representations (Li and Kim, 2024). In the context of CMT, SFT minimizes the forward KL divergence, which encourages a "mode-covering" behavior by forcing the model to align with specific human-written references. However, translation inherently allows for multiple semantically valid outputs. These human-written references often diverge from the model's intrinsic linguistic distribution. This relentless coercion to external modes causes a severe but unnecessary distributional drift.

To rectify this, we advocate for a transition from SFT to a Reinforcement Learning (RL) paradigm. As shown in Figure 1, RL with automatic translation quality metrics as rewards (Feng et al., 2025) offers a more principled alternative for adapting LLMs. Unlike SFT, RL is an on-policy optimization process that operates within the model's own generative distribution. By minimizing the reverse KL divergence, RL facilitates a "mode-seeking" process, encouraging the model to discover and reinforce high-quality translations that are naturally compatible with its existing expressive patterns (Lai et al., 2025; Shenfeld et al., 2025). This ensures that the model adapts to new translation tasks without abandoning the linguistic priors that govern its learned abilities.

Our experimental results validate this shift in perspective. RL-based training not only enhances translation quality on new tasks but also significantly curtails forgetting across both previously learned languages as well as general-purpose reasoning capabilities. Together, these findings redefine the fundamental challenge of CMT in the era of LLMs: it is less a struggle for knowledge retention and more a requirement for maintaining expressive stability. By repositioning RL as a mechanism for stable alignment, we provide a new path forward for sustainable multilingual development in large-scale generative models.

In summary, this work makes three main contributions:

- we provide the first systematic evaluation of continual MT in the era of LLMs and a detailed analysis of the forgetting phenomenon;
- we identify expressive drift as the dominant form of degradation induced by continual MT training in LLMs, localizing the degradation to the model's final layers and attributing it to the distributional mismatch induced by SFT;
- we demonstrate that RL offers a more principled and effective framework for sustainable alignment by performing on-policy optimization, which maintains expressive stability, offering new insights into designing continual adaptation strategies for generation tasks in the era of LLMs.

2 Revisiting Continual Machine Translation with LLMs

Early research on CMT has primarily focused on NMT models with limited capacity, in which forgetting is often attributed to knowledge erasure during sequential fine-tuning (Gu et al., 2021; Cao et al., 2021; Wu et al., 2024a). With the emergence of LLMs pretrained on massive and diverse multilingual corpora, the training paradigm has shifted from pre-training from scratch to continual fine-tuning (Xu et al., 2024). This have introduced new challenges, such as forgetting the strong in-context learning (Stap et al., 2024) and general instruction following abilities (Wu et al., 2025).

To mitigate such degradation, existing studies have predominantly relied on replay-based methods. However, these approaches are largely predicated on the classical assumption that 'catastrophic forgetting' stems from the irreversible erasure of underlying knowledge. While this premise holds for NMT models trained from scratch, it overlooks the fundamental paradigm shift in LLM-based translation, where proficiency is achieved through the alignment of pre-trained knowledge with specific instructions (Zhou et al., 2023a). This distinction implies a potential decoupling between internal knowledge and its external expression. Consequently, current CMT frameworks have yet to profoundly investigate whether the essence of forgetting lies in knowledge erasure or a failure in expressive accessibility.

To better understand the forgetting dynamic of CMT in the era of LLMs, we conduct a comprehensive empirical study. Specifically, we fine-tune an LLM with multiple translation tasks sequentially and analyze the nature of forgetting.

2.1 Experiment Protocol

Datasets and Benchmarks Our experiments involve fine-tuning LLMs using general machine translation data and evaluating their performance across both translation and broader capabilities. The datasets and benchmarks used for fine-tuning and evaluation are listed below:

- **Multi-stage Multilingual Machine Translation.** For parallel training data, we adopt the human-written datasets collected by Xu et al. (2025), which consists of a mixture of the Flores-200 dev set, NTREX test data, and WMT'15-22 test data. It covers four English-centric language pairs in both translation directions: Chinese (`zh`), German (`de`), Russian (`ru`), and Spanish (`es`. For evaluation, we use the WMT'24 test sets corresponding to the same four language pairs and directions.
- **General Ability of LLMs** To assess the impact on general-purpose capabilities, we evaluate models on two representative benchmarks from the Open LLM Leaderboard (Fourrier et al., 2024): IFEval (Zhou et al., 2023b), MMLU-Pro (Wang et al., 2024b), ARC-C (Clark et al., 2018), HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), and WinoGrande (Sakaguchi et al., 2020).

[1]

Implementation detail We fine-tune Qwen2.5-3B-Instruct (Yang et al., 2024) for machine translation from English to other languages in a fixed task order: Chinese (`zh`), German (`de`), Russian (`ru`), and Spanish (`es`). Across all experiments, we use the same training configuration, with 5 epochs for each task and a batch size of 32. For full-parameter fine-tuning, the learning rate is set to 1×10^{-6}. For PEFT method, we use a learning rate of 5×10^{-5} and employ LoRA adapters with rank 8. For evaluation, we employ reference-based XCOMET-XL(Guerreiro et al., 2024) to assess translation quality. In addition, we evaluate the general capabilities of LLMs using `lm-evaluation-harness`(Gao et al., 2024).

Metrics Let a_i^j denote the prediction performance (such as the BLEU score, the COMET score, etc.) on task i after learning task j. The following metrics are introduced:

- **Average Performance (AP)** measures the average performance of all tasks after training on the last task: **AP** $=: \frac{1}{n}\sum_{i=1}^{n} a_i^n$
- **Forgetting Rate (FR)** measures how much knowledge has been forgotten across all the learned tasks at the end of learning: **FR** $=: \frac{1}{n-1}\sum_{i=1}^{n-1}[\max_{j=1}^{n-1}(a_i^j - a_i^n, 0)]$

These metrics help better describe the multi-stage sequential training process of CMT.

2.2 Results

Translation Performance and Forgetting. Preliminary results (Table 1) reveal distinct forgetting

[1]Details of these benchmarks could be found in Appendix C.

Table 1: Results of continual machine translation experiment. Forgetting quantifies the performance degradation relative to the peak performance achieved on each task.

Method	Training Stage	EN-ZH	EN-DE	EN-RU	EN-ES	Avg
Qwen2.5-3B-Instruct	–	67.21	78.63	64.68	74.00	71.13
Full parameter fine-tuning	EN-ZH	71.82	–	–	–	–
	EN-DE	71.25	80.46	–	–	–
	EN-RU	70.22	79.13	67.82	–	–
	EN-ES	68.92	77.44	65.65	75.20	71.80
	Forgetting	2.90	3.02	2.17	–	2.69
PEFT(LoRA)	EN-ZH	73.08	–	–	–	–
	EN-DE	71.27	81.65	–	–	–
	EN-RU	71.01	81.31	70.05	–	–
	EN-ES	62.80	81.53	71.79	79.41	73.91
	Forgetting	10.28	0.12	0.00	–	3.47

patterns between tuning strategies. Full-parameter fine-tuning triggers consistent performance decay across all previous stages, averaging a -2.69 drop. Conversely, LoRA's robustness is contingent on linguistic distance. Although low-rank constraints are theorized to mitigate forgetting (Biderman et al., 2024), we find this stability holds only for typologically similar tasks (e.g., the Indo-European cluster), where performance drops remain negligible. However, LoRA fails significantly on the initial EN-ZH task, which suffers a catastrophic -10.28 collapse. This suggests that LoRA's efficacy diminishes when the model adapts to tasks that diverge significantly from the initial linguistic distribution, such as the typological gap between Sino-Tibetan and Indo-European languages.

Impact of Continual MT on General Abilities. We next examine how continual MT affects the general abilities of LLMs. As shown in Figure 2, the impact of continual learning on general abilities is task-dependent rather than uniformly negative. Pronounced performance degradation is consistently observed on IFEval and MMLU-Pro. In contrast, performance on other benchmarks, including ARC-Challenge, HellaSwag, TruthfulQA, and WinoGrande, remains largely stable, with several methods exhibiting comparable or even slightly improved results relative to the original model.

3 Diagnosing the Nature of Forgetting

The results presented above demonstrate that LLMs exhibit clear forgetting under CMT. However, beyond quantifying forgetting, an important open question remains: *what is the underlying source of the observed degradation?*

3.1 The "Expressive Drift" Hypothesis

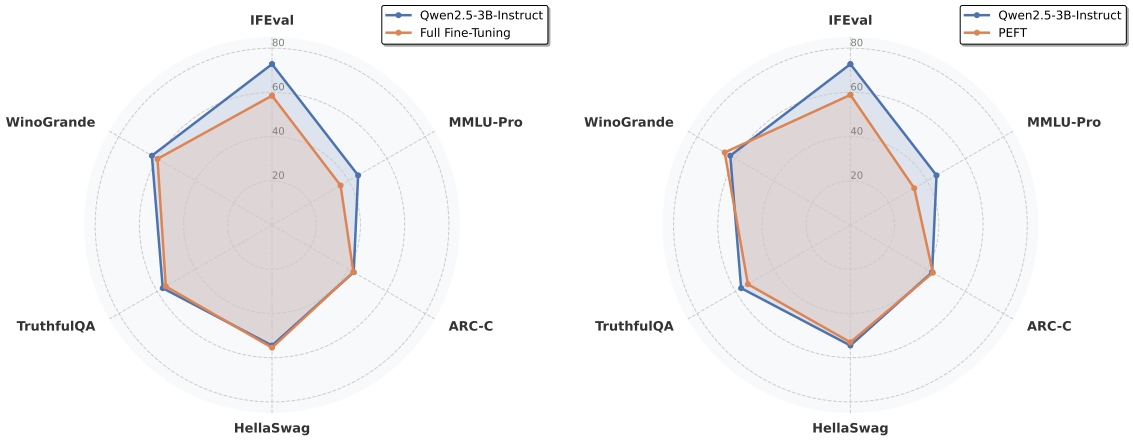
A notable clue emerges from the behavioral discrepancy across different evaluation protocols. Discriminative benchmarks (e.g., ARC-C, HellaSwag, WinoGrande), which only necessitate maintaining relative logit rankings among fixed options, remain stable. In contrast, generative tasks (e.g., IFEval, MMLU-Pro), which require the precise absolute calibration for successful decoding, suffer severe degradation. This contrast prompts the "expressive drift" hypothesis, which characterizes the observed forgetting as an expressive failure rather than a structural erasure of knowledge.

3.2 Disentangling Knowledge Erasure from Expressive Drift

To investigate this "expressive drift", it is necessary to disentangle *knowledge-level forgetting* from *expression-level failure*. To this end, we conduct a multi-layered empirical analysis spanning behavioral, representational, and parametric dimensions.

At the **behavioral level**, we utilize few-shot prompting to evaluate whether the lost capabilities can be reactivated through contextual anchoring. At the **representational level**, we perform a layer-wise analysis using Centered Kernel Alignment (CKA)(Kornblith et al., 2019) to quantify the similarity of internal states across training stages. Finally, at the **parameter level**, we examine the consistency of update directions across different model components. Together, these analyses provide a comprehensive lens through which to observe the preservation and expression of knowledge during continual adaptation.

Behavior Level: Few-shot Performance Recover Figure 3 compares zero-shot and few-shot



(a) Full-parameter Fine-tuning (b) PEFT

Figure 2: Performance retention on general ability benchmarks.

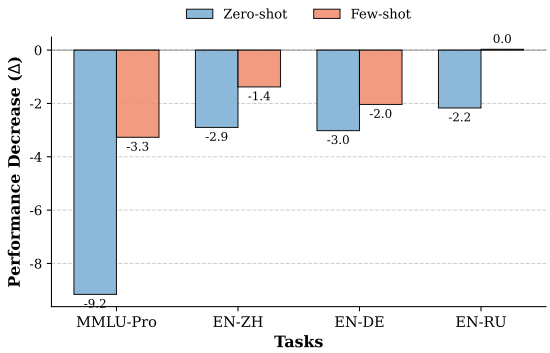


Figure 3: Forgetting under Zero-shot and Few-shot Evaluation on MMLU-Pro and Translation tasks.

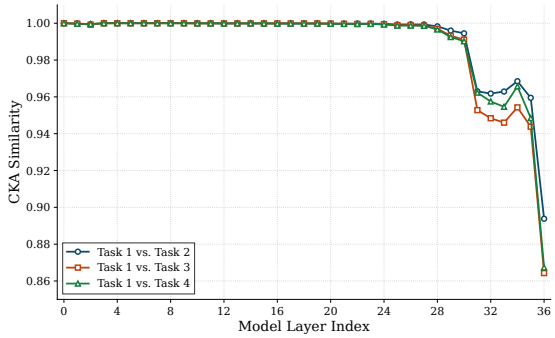


Figure 4: Layer-wise CKA similarity of Task 1 representations across sequentially trained model states.

(5 demonstrations) forgetting across MMLU-Pro and machine translation. As shown in the figure, few-shot prompting significantly reduces performance degradation for MMLU-Pro and translation. Both MMLU-Pro and translation require the retrieval and composition of concrete semantic or factual knowledge. The strong few-shot recovery suggests that their zero-shot forgetting is largely not due to the loss of underlying knowledge, but to failures in activating or expressing that knowledge. Few-shot exemplars help re-establish task semantics and generation behavior, thereby mitigating this form of forgetting.

Representation Level: Layer-wise CKA Similarity Figure 4 characterizes the layer-wise CKA similarity between the model state after the initial task and its subsequent states following the completion of Tasks 2, 3, and 4, utilizing the Task 1 test dataset as the evaluation probe. The analysis reveals a striking level of representational stability across the initial segments of the model, where CKA scores remain consistently near 1.00 through the first 28 layers. However, a pronounced divergence becomes evident in the deeper layers, particularly from layer 32 to 36, where similarity scores decline across all model comparisons toward a minimum near 0.86.

Research in large language model interpretability generally suggests that lower layers are primarily responsible for storing fundamental knowledge while upper layers govern the expression and task-specific mapping of that knowledge(Tenney et al., 2019; Geva et al., 2021; Meng et al., 2022). The sustained high similarity in the model backbone indicates that the fundamental knowledge is preserved within the latent space. In contrast, the observed reduction in CKA similarity in the final layers reflects a realignment of the output projections, implying that the expression of knowledge has drifted to align with the evolving task requirements.

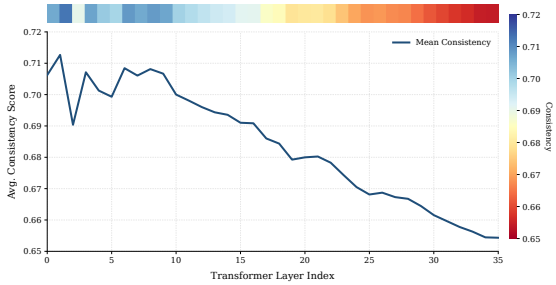


Figure 5: Layer-wise average consistency of parameter updates across the transformer architecture.

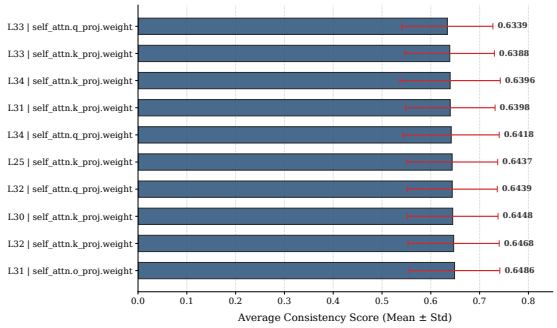


Figure 6: Top 10 modules with the lowest update consistency.

Parameter Level: Update Conflict Across Modules To locate the forgetting in specific modules, we compute the parameter updates for each learned task and calculate sign consistency. As illustrated in Figure 5, parameter update consistency exhibits a clear monotonic decline as model depth increases. Detailed analysis in Figure 6 confirms that update conflicts are exclusively localized within the self-attention projections of the deeper layers (L25-L34), specifically affecting the q_proj, k_proj, and o_proj weights. This spatial concentration of gradient conflict demonstrates that forgetting are a product of impaired knowledge expression within the deep attention blocks.

Together, the collective evidence across behavioral, representation, and parameter levels confirms our hypothesis that forgetting in sequential learning is driven by expressive drift rather than genuine knowledge erasure.

4 Towards Reinforcement Learning for Continual MT.

The identification of representation drift as the primary driver of forgetting necessitates a critical re-evaluation of the SFT paradigm for MT in the era of LLM. Although human-authored references provide high-quality signals, they are often out-of-distribution relative to the model's internal generative space. Rigid cross-entropy optimization against a single gold standard neglects the inherent semantic multiplicity of translation and imposes sharp distributional shifts on terminal attention layers, thereby exacerbating drift by overwriting existing priors with external targets.

To mitigate this, we propose a reinforcement learning framework that utilizes translation quality estimation as a reward signal. Unlike SFT, the RL paradigm facilitates the exploration of semantically appropriate candidates already aligned with the model's internal distribution. By optimizing within this generative space, RL preserves the stability of the underlying knowledge base while refining expressive capacity. This behavior-guided approach provides a principled mechanism to balance stability and plasticity, ensuring the acquisition of new capabilities without the expressive misalignment induced by traditional supervised methods. A formal theoretical explanation could be found in Appendix B.

4.1 RL Training Setup

RL Algorithm. We utilize Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as our RL algorithm. GRPO is a simplified PPO-based algorithm that significantly reduces training costs by eliminating the need for a value model. It operates by sampling G rollouts $\{o_1, ..., o_G\}$ from the current policy for a given input, calculating their cumulative rewards $R = \{R_1, ..., R_G\}$, and then using these rewards to estimate advantages $\hat{A}_{i,t}$ to guide policy updates. The optimization objective for GRPO is defined as follows:

$$L_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\mathcal{L}_{i,t}^{\text{clip}}(\theta) - \beta D_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\right)\right] \tag{1}$$

where

$$\mathcal{L}_{i,t}^{\text{clip}}(\theta) = \min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\Big)\hat{A}_{i,t}\right)$$
$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})},\ \hat{A}_{i,t} = \frac{R_i - \text{mean}(R)}{\text{std}(R)} \tag{2}$$

Reward Design. To mitigate reward hacking, we adopt COMET-22 (Rei et al., 2022), a reference-based metric trained on large-scale human judgments, as the reward model. Unlike surface-level n-gram metrics, COMET-22 provides a semantic-level assessment of translation adequacy and flu-

Table 2: Results of continual machine translation and zero-shot general ability evaluation. Forgetting is computed based on translation performance degradation across tasks. Best results in each column are highlighted in bold.

	Continual Machine Translation						General Ability (Zero-shot)	
Method	EN-ZH	EN-DE	EN-RU	EN-ES	AP↑	FR↓	IFEval	MMLU-pro
Qwen2.5-3B-Instruct	67.21	78.63	64.68	74.00	71.13	–	72.78	44.97
Seq-SFT	68.92	77.44	65.65	75.20	71.80	2.71	58.51	35.81
Seq-RL	**76.93**	**84.36**	**73.98**	**81.53**	**79.20**	**0.28**	**67.75**	**43.44**
Llama-3.2-3B-Instruct	72.89	70.51	61.68	74.93	70.00	–	82.37	37.41
Seq-SFT	60.26	77.39	59.08	74.24	67.74	3.64	77.10	30.15
Seq-RL	**66.47**	**86.54**	**61.47**	**81.36**	**73.96**	**0.22**	**81.29**	**38.77**

ency, allowing flexible evaluation beyond exact lexical overlap.

Implementation Details. We conduct RL experiments on both Qwen2.5-3B-Instruct (Yang et al., 2024) and Llama-3.2-3B-Instruct (Dubey et al., 2024). To ensure a fair comparison, we adopt the same hyperparameter configuration as in the previous experiments. Specifically, the learning rate is fixed to 1×10^{-6}. The model is trained with a batch size of 32 for 5 epochs on each language. For each training example, we sample $G = 8$ rollouts using a temperature of 1.0. The coefficient β for the KL term is set to 0.01.

4.2 Results&Analysis

Table 2 reports the results of the proposed RL-based approach under the continual MT setting across two base models. The RL policy yields consistent and substantial improvements in translation performance over all language pairs. Notably, these improvements are achieved with near-zero forgetting, indicating that translation quality is enhanced without destabilizing previously acquired behaviors. In addition to MT performance, RL preserves zero-shot general abilities as measured by IFEval and MMLU-pro. Compared with the original models, RL maintains comparable or improved performance on both benchmarks, suggesting that optimizing translation quality via RL does not interfere with general reasoning capabilities.

Figure 7 illustrates the *temporal dynamics* of forgetting during continual learning. Under SFT, general abilities drop sharply at the onset of the first MT task before stabilizing into an oscillatory decline, while translation retention decreases gradually. This trend indicates that the primary interference in SFT arises from cross-paradigm conflicts between translation-specific supervision and general-purpose behaviors rather than conflicts among similar MT tasks. In contrast, RL exhibits smooth and synchronized retention across both general and translation capabilities without abrupt stage-wise performance drops.

Taken together, these results demonstrate that RL can effectively improve translation performance while maintaining behavioral stability across both task-specific and general evaluation settings.

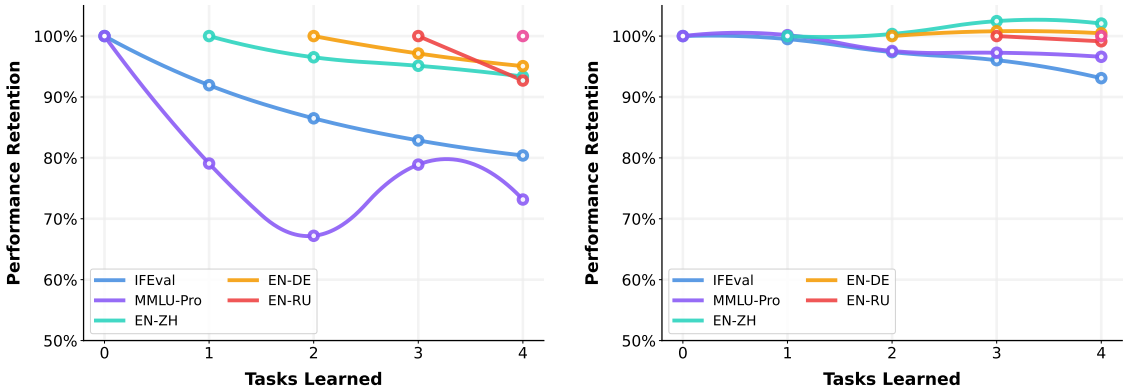
To verify that RL mitigates forgetting by addressing expressive drift, we conduct CKA analysis on RL-trained models. As shown in Figure 8, RL preserves representational structures across all layers with similarity exceeding 0.99. This global stability contrasts with the localized drift observed in SFT final layers. These results confirm that RL avoids the distribution shifts typical of teacher forcing, thereby maintaining a stable generative interface.

To provide a comprehensive assessment of the proposed framework, we offer additional empirical results in Appendix F, including comparative studies with diverse CL baselines, robustness analyses regarding translation directions and task sequences, and human evaluation.

5 Related Works

5.1 Continual Machine Translation

Continual learning for neural machine translation has mainly been studied in settings with relatively small models, where the central challenge is mitigating catastrophic forgetting during sequential fine-tuning. Existing methods largely fall into regularization-based approaches, which constrain parameter updates via techniques such as Elastic Weight Consolidation or knowledge distillation (Thompson et al., 2019; Saunders et al., 2019; Dakwale and Monz, 2017; Khayrallah et al., 2018; Cao et al., 2021), and architecture-based approaches



(a) Supervised Fine-tuning (b) Reinforcement Learning

Figure 7: Comparison of performance retention between SFT and RFT during CMT. Performance on each task is normalized by its initial peak and tracked as the model sequentially learns multiple translation tasks.

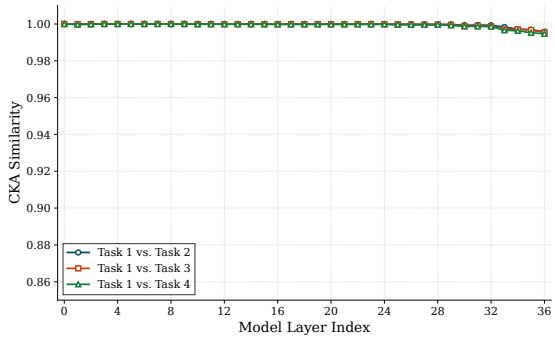


Figure 8: Layer-wise CKA similarity of Task 1 representations across sequentially RL-trained model states.

that isolate knowledge through pruning, parameter freezing, or auxiliary components (Gu et al., 2021; Wu et al., 2024a; Liang et al., 2021; Huang et al., 2023). In the context of LLMs, prior work primarily focused on continual pre-training over multilingual or parallel corpora (Yang et al., 2023; Xu et al., 2024; Zheng et al., 2024; Yu et al., 2025; Lu et al., 2024), leaving continual adaptation in post-training stages comparatively underexplored.

5.2 Continual Learning for LLMs

Although LLMs exhibit strong instruction-following abilities, continual fine-tuning on downstream tasks can still lead to catastrophic forgetting. To address this issue, classic continual learning techniques have been adapted to LLMs (Wu et al., 2024b; Shi et al., 2024). Prior work has primarily focused on replay-based methods and their variants, which mitigate forgetting by revisiting data from previous tasks (Scialom et al., 2022; Yin et al., 2022; Mok et al., 2023; He et al., 2024b; Wang et al., 2024a; Wu et al., 2025; He et al., 2024a). In parallel, architecture-based approaches, often implemented via parameter-efficient tuning such as LoRA, have been explored to isolate task-specific updates (Biderman et al., 2024; Wang et al., 2023b; Dou et al., 2023). However, these methods largely rely on supervised fine-tuning. More recent studies have begun to investigate reinforcement learning as an alternative paradigm, suggesting that its on-policy property can help alleviate forgetting (Shenfeld et al., 2025; Chen et al., 2025) and reduce inter-task interference.

6 Conclusion

This study revisits CMT for LLMs and redefines the fundamental challenge as maintaining expressive stability rather than preventing structural knowledge erasure. Through comprehensive analyses, we demonstrate that forgetting stems from expressive drift, where preserved latent knowledge fails to align with task instructions. This degradation is primarily localized within terminal attention layers because the mode covering behavior of SFT forces the model toward external distributions that conflict with internal linguistic priors. Conversely, RL provides a more principled alternative by facilitating a mode seeking alignment within the model's own generative distribution. Our results confirm that reinforcement learning effectively mitigates forgetting across both translation tasks and general capabilities. These findings position optimization driven by model policies as a sustainable framework for the sequential adaptation of generative models at scale.

Limitations

Our study is subject to certain limitations. First, prior studies have shown that RL can be less effective at inducing fundamentally new behaviors or overcoming strong model priors. Consistent with this observation, we find limited improvements in certain multilingual settings, such as enhancing Chinese translation performance for LLaMA-based models. This may hinder the effectiveness of RL when trying to learn complex out-of-distribution tasks in CL settings. Second, RL is susceptible to reward hacking. Although human evaluations in Appendix F.5 validate our findings, generalizing this framework to domains lacking trustworthy reward models remains a challenge.

Ethical Considerations

This work investigates continual machine translation with large language models and does not introduce new architectures or involve real-world deployment. All experiments are conducted on publicly available datasets and standard benchmarks, and we do not identify any direct ethical risks.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *CoRR*, abs/1606.06565.

Dan Biderman, Jacob P. Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. 2024. Lora learns less and forgets less. *Trans. Mach. Learn. Res.*, 2024.

Yue Cao, Hao-Ran Wei, Boxing Chen, and Xiaojun Wan. 2021. Continual learning for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3964–3974. Association for Computational Linguistics.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. 2025. Retaining by doing: The role of on-policy data in mitigating forgetting. *CoRR*, abs/2510.18874.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Praveen Dakwale and Christof Monz. 2017. Fine-tuning for neural machine translation with limited degradation across in- and out-of-domain data. In *Proceedings of Machine Translation Summit XVI, Volume 1: Research Track, MTSummit 2017, September 18-22, 2017, Nagoya, Aichi, Japan*, pages 156–169.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *CoRR*, abs/2312.09979.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *CoRR*, abs/2504.10160.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. Assoc. Comput. Linguistics*, 10:522–538.

Shuhao Gu, Yang Feng, and Wanying Xie. 2021. Pruning-then-expanding model for domain adaptation of neural machine translation. In *Proceedings*

of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pages 3942–3952. Association for Computational Linguistics.

Nuno Miguel Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet : Transparent machine translation evaluation through fine-grained error detection. *Trans. Assoc. Comput. Linguistics*, 12:979–995.

Jinghan He, Haiyun Guo, Kuan Zhu, Zihan Zhao, Ming Tang, and Jinqiao Wang. 2024a. SEEKR: selective attention-guided knowledge retention for continual learning of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 3254–3266. Association for Computational Linguistics.

Yongquan He, Xuancheng Huang, Minghao Tang, Lingxun Meng, Xiang Li, Wei Lin, Wenyuan Zhang, and Yifu Gao. 2024b. Don't half-listen: Capturing key-part information in continual instruction tuning. *CoRR*, abs/2403.10056.

Kaiyu Huang, Peng Li, Jin Ma, Ting Yao, and Yang Liu. 2023. Knowledge transfer in incremental learning for multilingual neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15286–15304. Association for Computational Linguistics.

Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn. 2018. Regularized training objective for continued training for domain adaptation in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, NMT@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 36–44. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Song Lai, Haohan Zhao, Rong Feng, Changyi Ma, Wenzhuo Liu, Hongbo Zhao, Xi Lin, Dong Yi, Min Xie, Qingfu Zhang, Hongbin Liu, Gaofeng Meng, and Fei Zhu. 2025. Reinforcement fine-tuning naturally mitigates forgetting in continual post-training. *CoRR*, abs/2507.05386.

Jianwei Li and Jung-Eun Kim. 2024. Superficial safety alignment hypothesis. *CoRR*, abs/2410.10862.

Jianze Liang, Chengqi Zhao, Mingxuan Wang, Xipeng Qiu, and Lei Li. 2021. Finding sparse structures for domain specific neural machine translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13333–13342. AAAI Press.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6467–6476.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 10748–10772. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jisoo Mok, Jaeyoung Do, Sungjin Lee, Tara Taghavi, Seunghak Yu, and Sungroh Yoon. 2023. Large-scale lifelong learning of in-context instructions and how to tackle it. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12573–12589. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural*

Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. 2022. COMET-22: unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 578–585. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.

Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2019. Domain adaptive inference for neural machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 222–228. Association for Computational Linguistics.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6107–6122. Association for Computational Linguistics.

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

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. 2025. RI's razor: Why online reinforcement learning forgets less. *CoRR*, abs/2509.04259.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *CoRR*, abs/2404.16789.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2026. Continual learning of large language models: A comprehensive survey. *ACM Comput. Surv.*, 58(5):120:1–120:42.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6189–6206. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4593–4601. Association for Computational Linguistics.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2062–2068. Association for Computational Linguistics.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2023a. A comprehensive survey of continual learning: Theory, method and application. *CoRR*, abs/2302.00487.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023b. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10658–10671. Association for Computational Linguistics.

Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, Haonan Lu, and Yujiu Yang. 2024a. Inscl: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 663–677. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Junhong Wu, Yuchen Liu, and Chengqing Zong. 2024a. F-MALLOC: feed-forward memory allocation for continual learning in neural machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

(Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 7180–7192. Association for Computational Linguistics.

Junhong Wu, Yang Zhao, Yangyifan Xu, Bing Liu, and Chengqing Zong. 2025. Boosting LLM translation skills without general ability loss via rationale distillation. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 12217–12236. Association for Computational Linguistics.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024b. Continual learning for large language models: A survey. *CoRR*, abs/2402.01364.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: plug & play modules and adaptive rejection for quality translation at scale. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *CoRR*, abs/2305.18098.

Wenpeng Yin, Jia Li, and Caiming Xiong. 2022. Contintin: Continual learning from task instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3062–3072. Association for Computational Linguistics.

Donglei Yu, Yang Zhao, Jie Zhu, Yangyifan Xu, Yu Zhou, and Chengqing Zong. 2025. Simulpl: Aligning human preferences in simultaneous machine translation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *CoRR*, abs/2402.15061.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911.

A Definition of Continual Machine Translation

Continual Machine Translation (CMT) frames the incremental acquisition of new translation capabilities as a sequential optimization problem across tasks $\{T_1, \ldots, T_n\}$. Unlike conventional NMT, CMT assumes that data arrives in stages with restricted access to previous samples, necessitating a balance between current task performance and the preservation of prior knowledge. Consequently, CMT evaluation moves beyond aggregate translation quality to incorporate specific CL metrics that quantify catastrophic forgetting, knowledge retention, and cross-task generalization. Specifically, let a_i^j denote the prediction performance (such as the BLEU score, the COMET score, etc.) on task i after learning task j. The following metrics are introduced:

- **Average Performance (AP)** measures the average performance of all tasks after training on the last task: **AP** $=: \frac{1}{n}\sum_{i=1}^{n} a_i^n$

- **Forgetting Rate (FR)** measures how much knowledge has been forgotten across all the learned tasks at the end of learning: **FR** $=: \frac{1}{n-1}\sum_{i=1}^{n-1}(\max_{j=1}^{n-1} a_i^j - a_i^n)$

B Formal Mathematical Equivalence between KL Divergence and Training Paradigms

Limitations of Supervised Fine-Tuning: Forward KL Divergence and Forced Distribution Fitting Given training data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$ and its defined distribution P_{data}, the optimization objective of SFT is to minimize the negative log-likelihood (NLL):

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{x\sim P_{\text{data}}}[\log P_\theta(x)] \tag{3}$$

where θ denotes the model's parameter space. To elucidate the statistical implications of this loss function, we introduce Kullback-Leibler (KL) divergence to measure the informational discrepancy between the empirical distribution P_{data} and the model distribution P_θ. According to the definition of KL divergence:

$$\begin{aligned} & D_{\text{KL}}(P_{\text{data}} \parallel P_\theta) \\ =& \mathbb{E}_{x\sim P_{\text{data}}}\left[\log \frac{P_{\text{data}}(x)}{P_\theta(x)}\right] \\ =& \mathbb{E}_{x\sim P_{\text{data}}}[\log P_{\text{data}}(x)] - \mathbb{E}_{x\sim P_{\text{data}}}[\log P_\theta(x)] \\ =& -H(P_{\text{data}}) + \mathcal{L}_{\text{SFT}}(\theta) \end{aligned} \tag{4}$$

Since the entropy of the ground-truth data distribution $H(P\text{data})$ is constant relative to the model parameters θ, minimizing $\mathcal{L}_{\text{SFT}}(\theta)$ is mathematically equivalent to minimizing the forward KL divergence.

Forward KL divergence $D_{\text{KL}}(P_{\text{data}} \parallel P_\theta)$ is characterized by a "mean-seeking" or "mode-covering" behavior. Specifically, the model distribution P_θ must allocate probability density wherever $P_{\text{data}}(x) > 0$ to prevent the divergence from approaching infinity. In the context of continual translation training, this mechanism introduces two critical issues. First, it penalizes valid expressions because if an autonomously generated sequence does not match the reference verbatim, SFT exerts significant gradient pressure to suppress that expression despite its semantic correctness. Second, it forces a distribution shift whenever the style or distribution of the reference translation lies outside the model's pre-trained preference manifold, as forward KL compels model parameters to shift toward these outliers. This rigid fitting via "teacher forcing" pressures the model to abandon its robust pre-trained representations in favor of imitating specific surface tokens. Consequently, this process induces deep attention routing disorders and expressive drift, ultimately precipitating catastrophic forgetting.

Reinforcement Learning: Flexible Alignment via Semantic Rewards Instead of requiring the model to memorize specific reference samples at the token level, RL optimizes global semantic utility by introducing semantic-aware evaluation metrics, such as COMET, as the reward function $R(x)$.In standard RL optimization, the objective function aims to maximize the expected reward without significantly deviating from the reference model P_{ref}:

$$\mathcal{J}RL(\theta) = \mathbb{E}x \sim P_\theta[R(x)] - \beta D_{\text{KL}}(P_\theta \parallel P_{\text{ref}}) \tag{5}$$

where β is the temperature parameter controlling regularization strength. To theoretically analyze the

convergence properties of this objective, we define an optimal target distribution $P^*(x)$ following the Gibbs distribution:

$$P^*(x) = \frac{1}{Z} P_{\text{ref}}(x) \exp\left(\frac{R(x)}{\beta}\right) \tag{6}$$

where $Z = \int P_{\text{ref}}(x) \exp(R(x)/\beta) dx$ is the partition function that ensures the normalization of the probability density.By substituting $P^*(x)$ into the definition of KL divergence, it can be proven that maximizing $\mathcal{J}_{RL}(\theta)$ is equivalent to minimizing the reverse KL divergence between the model distribution P_θ and the target distribution P^*. The derivation is as follows:

$$\begin{aligned}
&D_{\text{KL}}(P_\theta \parallel P^*) \\
=&\mathbb{E}_{x\sim P_\theta}\left[\log \frac{P_\theta(x)}{P^*(x)}\right] \\
=&\mathbb{E}_{x\sim P_\theta}\left[\log P_\theta(x) - \log\left(\frac{1}{Z} P_{\text{ref}}(x) e^{R(x)/\beta}\right)\right] \\
=&\mathbb{E}_{x\sim P_\theta}\left[\log \frac{P_\theta(x)}{P_{\text{ref}}(x)} - \frac{R(x)}{\beta} + \log Z\right] \\
=&D_{\text{KL}}(P_\theta \parallel P_{\text{ref}}) - \frac{1}{\beta}\mathbb{E}_{x\sim P_\theta}[R(x)] + \log Z
\end{aligned} \tag{7}$$

Rearranging the terms yields:

$$\mathcal{J}_{RL}(\theta) = -\beta D_{\text{KL}}(P_\theta \parallel P^*) + \beta \log Z \tag{8}$$

Since $\beta \log Z$ depends solely on the reward and reference models and remains constant relative to the parameters θ, the RL training objective is mathematically equivalent to minimizing the reverse KL divergence:

$$\arg\max_\theta \mathcal{J}_{RL}(\theta) = \arg\min_\theta D_{\text{KL}}(P_\theta \parallel P^*) \tag{9}$$

Reverse KL divergence is characterized by "mode-seeking" behavior. Unlike SFT, which passively covers external samples, RL drives the model to actively explore regions within its own distribution space that yield high semantic rewards. This paradigm offers significant flexibility. First, it acknowledges diversity of expression, as the model receives positive reinforcement for any sequence that aligns semantically with the source sentence without being forced to replicate the specific lexical choices of a reference. Second, it maintains generative stability through on-policy sampling, allowing the model to fine-tune based on its existing knowledge base. This approach avoids the destructive distribution stretching induced by SFT and mitigates exposure bias. Consequently, this shift from "surface imitation" to "objective alignment" enables the model to acquire translation skills while preserving the structured capabilities developed during pre-training, thereby fundamentally mitigating performance degradation in continual learning.

C Details of Datasets and Benchmarks

Language Pair	Train	Test
EN–ZH	6565	997
EN–DE	6529	997
EN–RU	6568	997
EN–ES	2909	997

Table 3: Dataset statistics of the multilingual translation benchmarks used in our experiments.

Our experiments involve fine-tuning LLMs using general machine translation data and evaluating their performance across both translation and broader capabilities. The datasets and benchmarks used for fine-tuning and evaluation are listed below:

Multi-stage Multilingual Machine Translation. For parallel training data, we adopt the human-written datasets collected by Xu et al. (2025). This collection comprises human-authored test sets from previous WMT, along with the development and test sets from Flores-200 (Goyal et al., 2022). It covers four English-centric language pairs in both translation directions: Chinese (zh), German (de), Russian (ru), and Spanish (es). For evaluation, we use the WMT'24 test sets corresponding to the same four language pairs and directions. Here, we present detailed statistics for the datasets used in our experiments in Table **??**.

General Ability of LLMs To assess the impact of continual MT on general-purpose capabilities, we evaluate models on two representative benchmarks from the Open LLM Leaderboard (Fourrier et al., 2024):

- **IFEval** (Zhou et al., 2023b) evaluates a model's ability to strictly follow explicit and verifiable instructions, such as enforcing required keywords, output formats, or structural constraints. It focuses on instruction compliance rather than semantic correctness, enabling precise and deterministic evaluation.
- **MMLU-Pro** (Wang et al., 2024b) is a refined and more challenging version of the original

MMLU benchmark, designed to assess broad knowledge and reasoning capabilities across diverse domains. It reduces ambiguity and data contamination present in earlier versions, providing a more reliable measure of general reasoning ability.

- **ARC-C** (Clark et al., 2018) consists of multiple-choice science questions that require non-trivial reasoning and inference beyond surface-level pattern matching. It is commonly used to evaluate core scientific reasoning and common-sense understanding.

- **HellaSwag** (Zellers et al., 2019) is a commonsense reasoning benchmark that requires selecting the most plausible continuation of a given situation. The dataset is adversarially constructed to minimize annotation artifacts, emphasizing robust contextual and commonsense understanding.

- **TruthfulQA** (Lin et al., 2022) evaluates a model's tendency to generate truthful answers to questions that commonly elicit false or misleading responses. It is designed to measure robustness against hallucination and the propagation of widely held misconceptions.

  WinoGrande (Sakaguchi et al., 2020) is a large-scale commonsense reasoning benchmark based on pronoun resolution. It tests a model's ability to leverage contextual and world knowledge to resolve ambiguous references, while reducing dataset biases through careful adversarial filtering.

D Prompt Templates

The following prompt is utilized in all experiments:

The Template for Machine Translation

Prompt: Please translate the following sentence from `{Source Language Name}` to `{Target Language Name}`:
`{Source Sentence}`

Response: `{Reference Translation}`

E Details of Few-shot Forgetting Analysis

To disentangle *knowledge-level forgetting* from *expression-level failure*, we compare zero-shot and few-shot evaluation settings. For MMLU-Pro and IFEval, we adopt the `lm-evaluation-harness` framework, which allows setting the number of demonstrations. For machine translation, we randomly sample demonstrations for each language pair from the corresponding training data. In all few-shot settings, we use 5 demonstrations, which are organized into multi-turn chat histories following the model's default chat template.

F Additional Experiments

F.1 Comparison with CL Approaches

In our main paper, we mainly compare two different training paradigms, namely SFT and RL. To further validate the effectiveness of RL for CMT, we add a comparison with state-of-the-art CL approaches.

Baseline Approaches We evaluate a diverse set of representative CL approaches:

- **Seq-FT**, a baseline that sequentially fine-tunes the entire LLM on the data stream as new tasks arrive;

- **GEM** (Lopez-Paz and Ranzato, 2017), a rehearsal-based approach that stores gradients of past-task samples and projects the current task's gradient to avoid conflicts with previous ones;

- **EWC** (Kirkpatrick et al., 2016; Thompson et al., 2019; Saunders et al., 2019), a regularization-based method that leverages the Fisher Information Matrix to identify and constrain parameters that are critical for previously learned tasks;

- **LwF** (Khayrallah et al., 2018; Dakwale and Monz, 2017; Cao et al., 2021), another regularization-based approach that mitigates forgetting via knowledge distillation from earlier task models;

- **Seq-LoRA**, a parameter-efficient baseline that sequentially fine-tunes LoRA adapters as new tasks arrive; and

- **O-LoRA** (Wang et al., 2023b), a parameter-efficient CL method that assigns different tasks to orthogonal low-rank subspaces to explicitly reduce inter-task interference.

Implementation Details We adhere to the settings used in the main experiments, fine-tuning

Qwen2.5-3B-Instruct (Yang et al., 2024) sequentially for multilingual machine translation tasks. Across all experiments, we use the same training configuration, with 5 epochs for each task and a batch size of 32. For full-parameter fine-tuning based methods (Seq-FT, GEM, EWC, LwF), the learning rate is set to 1×10^{-6}. For PEFT methods (Seq-LoRA, O-LoRA), we use a learning rate of 5×10^{-5} and employ LoRA adapters with rank 8. Evaluation is conducted with XCOMET-XL and `lm-evaluation-harness`.

Results Table 4 summarizes the performance of the proposed Seq-RL framework compared to traditional CL baselines across MT and general capability benchmarks (Detailed results available in Table 5). The results demonstrate that Seq-RL consistently achieves superior plasticity while maintaining exceptional stability. In terms of translation performance, Seq-RL attains the highest average score of 79.20, outperforming the strongest SFT-based CL baseline, O-LoRA (75.31), by a significant margin. This improvement is consistent across all language pairs, with Seq-RL reaching peak scores in EN-ZH (76.93) and EN-DE (84.36).

Critically, Seq-RL effectively mitigates catastrophic forgetting without sacrificing the model's general reasoning and instruction-following abilities. While conventional sequential SFT (Seq-SFT) suffers from a high forgetting rate (2.71) and a sharp decline in general benchmarks (e.g., IFEval dropping from 72.78 to 58.51), Seq-RL maintains a low forgetting rate of 0.28 and preserves general capabilities (67.75 on IFEval and 43.44 on MMLU-pro). While LwF achieves a slightly lower forgetting rate (0.00), its translation performance (74.71) remains considerably lower than that of Seq-RL. These findings corroborate our previous representational analysis: by optimizing within the model's existing generation space rather than forcing alignment to out-of-distribution human references, RL avoids the disruptive representational drift characteristic of SFT. This allows the model to acquire new task-specific translation skills while keeping its underlying knowledge base and instruction-following behaviors intact.

F.2 Generalization to Many→EN Direction.

Prior work has shown that CMT in the Many→EN setting is substantially easier than in the EN→Many direction, largely due to the shared target language (English), which mitigates cross-lingual interference in multilingual MT. However, as shown in Table 6, sequential SFT still exhibits severe performance degradation on both translation and general-purpose evaluation tasks. This suggests that target-language consistency alone is insufficient to prevent forgetting, and that SFT with human-written references may even exacerbate performance degradation. In contrast, Seq-RL consistently achieves the best average performance with near-zero forgetting, demonstrating that RL remains highly effective in preserving translation quality and general abilities even in this comparatively easier continual learning setting.

F.3 Effects of Task Order

Since CL is known to be sensitive to task order, we performed an additional "Reverse Order" experiment on the Qwen-2.5 model (i.e., reversing the sequence of language pairs).

As shown in Table 7, we observed a significant performance collapse in the last training stage. We attribute this to the substantial typological distance between Chinese (Sino-Tibetan) and the other three European languages (Indo-European). This linguistic divergence triggers a severe Language Conflict, resulting in the "language dominance" phenomenon where models generated Chinese outputs even for non-Chinese translation prompts. These results further demonstrate that catastrophic forgetting is highly sensitive to the training order. When a task with a distinct data distribution (like en-zh) is introduced last, traditional CL methods like O-LoRA and Seq-SFT struggle to balance the integration of new knowledge with the preservation of existing linguistic structures.

Remarkably, despite these challenging conditions, Seq-RL maintained consistent high performance (AVG: 79.09, FR: -0.1575). Unlike SFT-based methods, Seq-RL preserves the model's instruction-following capabilities (IFEval: 69.38) and general knowledge (MMLU: 44.52). This reinforces our core argument: RL-based training is inherently more robust against task interference, effectively mitigating forgetting even when faced with significant linguistic shifts and adverse task sequences.

F.4 Failure of Language Control

We have conducted a granular forgetting analysis focusing on Target Language Interference (TLA), which we found to be the primary driver of the performance drop in our multilingual continual learn-

Table 4: Comparison with CL approaches

Method	Machine Translation						General Ability	
	EN-ZH	EN-DE	EN-RU	EN-ES	Avg↑	FR↓	IFEval	MMLU-pro
Qwen2.5-3B-Instruct	67.21	78.63	64.68	74.00	71.13	–	72.78	44.97
Seq-SFT	68.92	77.44	65.65	75.20	71.80	2.71	58.51	35.81
EWC	69.03	77.37	65.54	74.94	71.72	1.93	57.67	36.39
GEM	69.19	77.50	66.08	75.26	72.01	1.92	56.47	35.84
LwF	72.63	80.91	68.54	76.77	74.71	**0.00**	66.79	42.15
LoRA	62.80	81.53	71.79	79.41	73.88	3.47	58.87	33.34
O-LoRA	72.17	81.05	70.06	77.96	75.31	0.33	62.83	35.19
Seq-RL	**76.93**	**84.36**	**73.98**	**81.53**	**79.20**	0.28	**67.75**	**43.44**

ing setup.

Instead of a uniform degradation across linguistic features (like word spelling or grammar), we observed a catastrophic failure in language control. We quantified this using TLA—the percentage of generated tokens that match the intended target language (detected via a Language Identification tool).

As shown in Table 8, while the en-es XCOMET score drops significantly, the Target Language Accuracy (TLA) plummets from 90.10% to 3.16% after the en-zh stage. This near-total collapse of TLA suggests a catastrophic failure in language control (steering). The model does not "forget" how to translate the source semantics; instead, its "output gate" is overwritten by the most recent language (zh), leading to massive off-target generation.

F.5 Human Evaluation

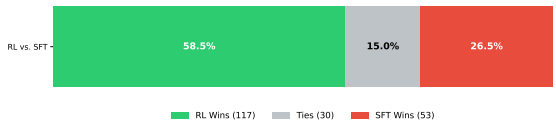


Figure 9: Human Evaluation: Win/Tie/Loss Comparison.

A central challenge in reinforcement learning is preventing "reward hacking," where a model exploits vulnerabilities in the reward function to achieve high scores at the expense of genuine generation quality or linguistic logic(Amodei et al., 2016; Ouyang et al., 2022). To verify whether the performance gains from RL align with authentic human preferences, we conducted a double-blind human evaluation. We randomly sampled 200 instances from the English-to-Chinese test set and compared the outputs generated by the SFT and RL models. Under a double-blind protocol, reviewers assessed faithfulness and fluency without knowledge of the models' identities, designating each pair as a Win, Loss, or Tie.

Figure 9 shows the results of human evaluation. Across the 200 samples, the RL model achieved a win rate 2.18 times higher than that of the SFT model. Reviewer feedback indicated that RL-trained models produced translations with superior terminological accuracy and syntactic naturalness, significantly reducing the over-literal translations and language-mixing issues common in SFT outputs. This substantial margin in win rates demonstrates that the improvements in XCOMET scores correspond to perceptible gains in translation quality rather than mere overfitting to neural evaluation metrics.

Table 5: Full performance matrices for continual machine translation under different training strategies. The Original row indicates the initial model performance. For each method, rows correspond to the model state after training on a specific language pair, yielding a lower-triangular matrix over language pairs. The Forgetting row reports the performance degradation on earlier tasks after the final training stage.

Method	Training Stage	EN-ZH	EN-DE	EN-RU	EN-ES
Original	–	67.21	78.63	64.68	74.00
Seq-SFT	EN-ZH	71.82	–	–	–
	EN-DE	71.25	80.46	–	–
	EN-RU	70.22	79.13	67.82	–
	EN-ES	68.92	77.44	65.65	75.20
	Forgetting	2.90	3.02	2.17	0.00
EWC	EN-ZH	71.90	–	–	–
	EN-DE	71.43	79.84	–	–
	EN-RU	70.27	78.63	65.98	–
	EN-ES	69.03	77.37	65.54	74.94
	Forgetting	2.87	2.47	0.44	0.00
GEM	EN-ZH	71.93	–	–	–
	EN-DE	71.30	79.95	–	–
	EN-RU	70.42	78.95	66.64	–
	EN-ES	69.19	77.50	66.08	75.26
	Forgetting	2.74	2.45	0.56	0.00
LwF	EN-ZH	71.85	–	–	–
	EN-DE	72.29	80.60	–	–
	EN-RU	72.46	80.32	68.31	–
	EN-ES	72.63	80.91	68.54	76.77
	Forgetting	0.00	0.00	0.00	0.00
LoRA	EN-ZH	73.08	–	–	–
	EN-DE	71.27	81.65	–	–
	EN-RU	71.01	81.31	70.05	–
	EN-ES	62.80	81.53	71.79	79.41
	Forgetting	10.28	0.12	0.00	0.00
O-LoRA	EN-ZH	72.51	–	–	–
	EN-DE	71.79	81.69	–	–
	EN-RU	71.21	80.97	69.56	–
	EN-ES	72.17	81.05	70.06	77.96
	Forgetting	0.34	0.64	0.00	0.00
RL	EN-ZH	75.38	–	–	–
	EN-DE	75.64	83.97	–	–
	EN-RU	77.23	84.64	74.62	–
	EN-ES	76.93	84.36	73.98	81.53
	Forgetting	0.00	0.00	0.64	0.00

Table 6: Results of continual machine translation on Many→EN direction and zero-shot general ability evaluation. Forgetting is computed based on translation performance degradation across tasks. Best results in each column are highlighted in bold.

	Continual Machine Translation						**General Ability (Zero-shot)**	
Method	ZH-EN	DE-EN	RU-EN	ES-EN	AP↑	FR↓	IFEval	MMLU-pro
Qwen2.5-3B-Instruct	80.85	87.23	81.06	85.83	83.74	–	72.78	44.97
Seq-SFT	72.09	79.80	71.78	81.36	77.45	3.17	51.92	28.70
Seq-RL	**82.82**	**89.14**	**83.28**	**88.14**	**85.85**	**0.02**	**71.58**	**44.51**

Table 7: Results of reverse task order setting

	Machine Translation						General Ability	
Method	**EN-ES**	**EN-RU**	**EN-DE**	**EN-ZH**	**Avg↑**	**FR↓**	IFEval	MMLU-pro
Qwen2.5-3B-Instruct	74.00	64.68	78.63	67.21	71.13	-	72.78	44.97
Seq-SFT	58.98	54.89	63.08	66.89	60.96	-7.85	51.44	36.10
Seq-RL	**80.36**	**73.35**	**83.27**	**79.37**	**79.09**	**-0.1575**	**69.38**	**44.52**

Table 8: Translation performance and target language accuracy

Training Stages	**EN-ES**	**EN-RU**	**EN-DE**	**EN-ZH**
EN-ES Performance (XCOMET Score)	73.81	71.04	68.95	58.98
TLA (%)	90.10	89.53	77.95	**3.16**