Conditions for Catastrophic Forgetting in Multilingual Translation

Danni Liu **Jan Niehues**
Karlsruhe Institute of Technology, Germany
{danni.liu, jan.niehues}@kit.edu

Abstract

Fine-tuning multilingual foundation models on specific languages often induces catastrophic forgetting, degrading performance on languages unseen in fine-tuning. While this phenomenon is widely-documented, the literature presents fragmented results about when forgetting occurs. To address this ambiguity, we conduct a systematic empirical study using machine translation as a testbed to identify the conditions that trigger catastrophic forgetting in multilingual fine-tuning. Through controlled experiments across different model architectures, data scales, and fine-tuning approaches, we reveal that the relative scale between model and data size is a primary determinant of forgetting. Moreover, we demonstrate that a model's instruction-following ability is more critical for retaining multilingual knowledge than its architecture. Contrary to assumptions, parameter-efficient fine-tuning offers no clear advantage over full fine-tuning in mitigating forgetting. Lastly, we show that cross-lingual alignment can mitigate forgetting while also facilitating positive transfer to unseen target languages.

1 Introduction

Foundation models pretrained on vast amounts of multilingual data have become the standard backbone for modern natural language processing systems. To achieve optimal performance, however, these models typically require fine-tuning on downstream tasks. This specialization introduces a critical trade-off: while performance on the target task improves, the model may suffer from *catastrophic forgetting* (McCloskey and Cohen, 1989), a substantial degradation of capabilities on tasks or languages not present in the fine-tuning data.

A common use case is to fine-tuning multilingual models to focus on specific languages or language pairs. Ideally, this process would not harm, and might even improve, performance on unseen languages through positive transfer, as illustrated on the left of Figure 1. However, empirical evidence often shows the opposite. Models frequently lose proficiency in languages they were not fine-tuned on (Vu et al., 2022b; Sun et al., 2023; Winata et al., 2023), as shown on right of Figure 1.

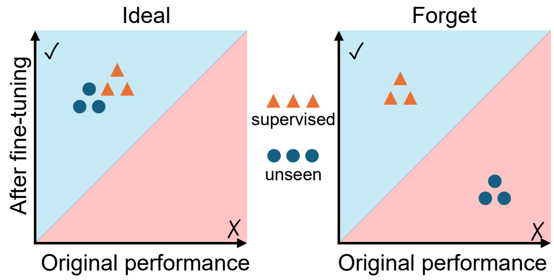


Figure 1: Selectively fine-tuning on some languages or translation directions may lead to positive transfer (left) or catastrophic forgetting (right).

Machine translation (MT) serves as a compelling testbed for studying multilingual catastrophic forgetting. First, a model supporting n languages encompasses $n(n-1)$ directed translation pairs, offering a large and structured space to analyze forgetting patterns. Second, languages can be "unseen" in different roles. For example, a language may be present only as a source language, only as a target language, or in specific source-target pairs that were never explicitly trained. This enables fine-grained analysis of how different types of exposure during fine-tuning affect retention. Moreover, forgetting can occur asymmetrically, where a model may retain the ability to translate from language A to B while losing the reverse direction.

Despite its practical importance, the literature presents a fragmented and sometimes contradictory picture of when catastrophic forgetting occurs in MT. On one hand, studies on traditional NMT models trained from scratch (Berard, 2021) and some large pretrained models (Vu et al., 2022a; Liu and Niehues, 2022; Liu et al., 2023; Lai et al., 2023) report severe forgetting after standard fine-tuning,

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 347–359

where the ability to translate unseen directions is almost entirely lost. These findings suggest that catastrophic forgetting is an inevitable consequence of selective specialization. On the other hand, recent works on large language models (LLMs) provided mixed evidence. Richburg and Carpuat (2024) demonstrated that fine-tuning Llama 2 (Touvron et al., 2023) and Tower (Alves et al., 2024) models on specific language pairs could improve performance on unseen pairs, indicating positive transfer. Conversely, Zan et al. (2024) found that their fine-tuned Llama 2 models performed very poorly on unseen directions, again indicating issues with forgetting.

These conflicting results raise fundamental questions about the factors leading to catastrophic forgetting. Has the emergence of large language models altered the dynamics of catastrophic forgetting? To what extent do model architecture (encoder-decoder versus decoder-only), scale, or fine-tuning methodology determine whether a model forgets or generalizes? How do factors like the volume of fine-tuning data, the use of parameter-efficient fine-tuning (PEFT), or instruction-following capabilities influence the retention of multilingual abilities? To resolve these ambiguities, we conduct a systematic study to identify the conditions that trigger catastrophic forgetting in multilingual MT. We systematically control for key variables, including model architecture and size, fine-tuning data composition and scale, full-parameter vs. parameter-efficient fine-tuning, and instruction-following versus standard fine-tuning approaches. With a series of controlled experiments, we demonstrate that:

- The relative scale between pre-trained model parameters and fine-tuning data volume is a critical factor in catastrophic forgetting, with smaller models fine-tuned on larger datasets being most vulnerable (§4.1).
- Whether a model supports instruction-following, rather than its underlying architecture (encoder-decoder versus decoder-only), is a primary factor impacting catastrophic forgetting (§4.2).
- Contrary to common assumptions, parameter-efficient fine-tuning with LoRA (Hu et al., 2022) provides no significant advantage over full fine-tuning in preventing catastrophic forgetting under our experimental conditions (§4.4).
- Besides mitigating forgetting, cross-lingual alignment methods may facilitate positive transfer, with improvements observed on translation directions with unseen target languages (§5).

2 Related Work

Catastrophic Language Forgetting in MT Catastrophic forgetting in machine translation has been extensively studied. Dakwale and Monz (2017); Thompson et al. (2018, 2019) established that domain-specific fine-tuning degrades performance on previously learned domains with specific subject areas or text styles. Many subsequent works have investigated the underlying mechanisms and mitigation strategies for domain forgetting in MT, e.g., Gu and Feng (2020); Saunders and DeNeefe (2024); Eschbach-Dymanus et al. (2024); Wu et al. (2024); Hu et al. (2024). Compared to domain forgetting, the multilingual dimension of forgetting has received less attention. Berard (2021) demonstrated severe language forgetting in conventional encoder-decoder-based MT models during standard fine-tuning on selected languages, while Vu et al. (2022a) showed that domain-specific fine-tuning compounds forgetting across both domains and languages. Liu and Niehues (2022); Liu et al. (2023) confirmed that standard fine-tuning consistently triggers catastrophic forgetting of unseen language pairs, even in pre-trained models with large language coverage, such as M2M-124 (Goyal et al., 2022) and mBART-50 (Tang et al., 2021). Besides language and domain forgetting, models also lose in-context learning abilities after fine-tuning (Alves et al., 2023).

Model Factors Influencing Forgetting The scale of both the model and its pretraining data has been identified as a key factor in mitigating catastrophic forgetting (Ramasesh et al., 2022). Our study extends this analysis by examining the relative scale between the model and the fine-tuning data. The choice of fine-tuning methodology is another contested factor. Kalajdzievski (2024) suggests that LoRA does not resolve catastrophic forgetting, while Biderman et al. (2024) suggest that LoRA "learns less and forgets less". The finding by Zhang et al. (2024) that the optimal fine-tuning method is highly task-dependent warrants a specific investigation for the task of multilingual MT.

3 Controlled Setting to Study Catastrophic Forgetting in MT

Our controlled experiments are structured along two dimensions, namely the choice of base model and the characteristics of the training dataset.

Model	Size	Model Type
M2M-124-0.2B	175M	Translation-specific
M2M-124-0.6B	615M	
Qwen2.5-0.5B-Instruct	494M	Instruction-following
Qwen2.5-7B-Instruct	7B	
Llama-3-8B-Instruct	8B	

Table 1: Base models and their configurations.

3.1 Base Models

An overview of all base models and their configurations is provided in Table 1.

Translation-Specific Models We choose M2M-124[1] (Goyal et al., 2022) with two sizes:

- M2M-124-0.2B: smallest-scale baseline
- M2M-124-0.6B: larger-scale comparison to isolate model size effects

Instruction-Following Models We evaluate and fine-tune models from two prominent families, Qwen 2.5 (Qwen Team et al., 2025) and Llama 3 (AI @ Meta et al., 2024):

- Qwen2.5-0.5B-Instruct : similar to M2M-124-0.6B in size for comparison between translation-specific and instruction-following models[2]
- Qwen2.5-7B-Instruct: larger-scale instruction-following baseline
- Qwen2.5-7B-Instruct (LoRA): identical to full fine-tuning but using LoRA as a PEFT approach
- Llama-3-8B-Instruct (LoRA): similar scale to above but from another family

3.2 Data

Dataset Overview As shown in Table 2, we experiment on datasets of different scales:

- SMALL: training dataset in ALMA (Xu et al., 2024), covering five languages paired with English: Czech (cs), German (de), Icelandic (is), Russian (ru), and Chinese (zh). The unseen languages include Hebrew (he), Japanese (ja), and Ukranian (uk).
- LARGE: from the WMT 21 Shared Task on Large-Scale Multilingual Machine Translation (Wenzek et al., 2021), focusing on three related Austronesian languages paired with English: Javanese (jv), Malay (ms), and Tagalog (tl). The unseen language is Indonesian (id).

[1]We choose M2M-124 over NLLB-200 models of similar sizes (NLLB Team, 2024) as the former showed stronger performance in our preliminary experiments.

[2]We note that this is not fully controlled setup contrasting M2M-124-0.6B due to different pre-training data.

Dataset	Details
SMALL	**Training Data:** ALMA (117K sentence pairs) **Test** (supervised): WMT23 (Kocmi et al., 2023) **Test** (unseen pair): WMT24 (Kocmi et al., 2024) **Test** (unseen source): WMT23 **Test** (unseen target): WMT23 **Training directions:** {cs, de, is, ru, zh}↔en **Testing directions:** - Unseen pair (20): {cs, de, is, ru, zh}↔{cs, de, is, ru, zh} - Unseen source (3): {he, ja, uk} → en - Unseen target (3): en → {he, ja, uk}
LARGE	**Training Data:** WMT21 large-scale multilingual track (54M sentence pairs) **Test** (unseen pair): FLoRes (Goyal et al., 2022) **Test** (unseen source): FLoRes **Test** (unseen target): FLoRes **Training directions:** {jv, ms, tl}↔en **Unseen testing directions:** - Unseen pair (6): {jv, ms, tl}↔{jv, ms, tl} - Unseen source (4): id → {en, jv, ms, tl} - Unseen target (4): {en, jv, ms, tl} → id

Table 2: Dataset overview for training and testing configurations for both small and large-scale experiments.

- subsampled LARGE: sampled from the LARGE dataset with 12K, 120K, and 1.2M sentences per language pair respectively.

Unseen Language Pairs We evaluate catastrophic forgetting on three types of unseen language pairs. Our analysis focuses on pairs where at least one language was seen during fine-tuning, as pairs with two unseen languages consistently showed severe performance degradation in preliminary experiments. The three categories are:

- **Unseen Pair**: Both the source and target languages are present in the fine-tuning data, but not in combination. This is the most challenging category as explained next.
- **Unseen Source**: The source language has not been seen during fine-tuning, but the target language has.
- **Unseen Target**: The target language has not been seen during fine-tuning, but the source language has.

Among the three evaluated categories, the "unseen pair" scenario presents a unique challenge. While counterintuitive, this case is often more difficult than scenarios involving languages completely unseen during fine-tuning. The primary reason for this difficulty lies in the English-centric nature of the fine-tuning dataset. Because all training examples are paired with English, the model learns an implicit association that a specific source language uniquely predicts English as the target language, which represents a spurious correlation (Gu et al., 2019).[3] In contrast, the other two conditions do not present this conflict to the same level. For an unseen source language, the model has not formed any directional association during fine-tuning. Therefore, there is no learned association to be overridden. The unseen target language scenario is also comparatively less difficult. Specifically, as long as the source language has to translate into multiple different target languages during training or has not been seen in training, the model does not learn a one-to-one mapping to a single output. This condition applies to four of the seven unseen target language scenarios (en →{he, ja, uk, id}), where the source language was part of a multi-target translation setup. This configuration discourages over-specialization toward a single output language, reducing the overall difficulty of translating into a unseen target language for this category.

Language Control Mechanisms Following the original models, we use different language specification methods. For M2M-124, we follow their token-based control, prepending source and target sentences with their respective language tokens:

```
<source_lang_token> source sentence
<target_lang_token> target sentence
```

For instruction-following models, we use the system prompt "Translate the given sentence from [source language] to [target language]" followed by the source sentence. In ablations, we also test instructions in the target language[4].

3.3 Training and Inference

For full fine-tuning, we update all model parameters. For LoRA, we adopt a rank of 8 and α of 16, applying adapters to all components within self-attention (Query, Key, Value, Output, Gate) and linear projections. This LoRA configuration was chosen after initial experiments applying LoRA to fewer components showed weaker supervised performance. It also creates conditions more analogous to full fine-tuning than selective adapter application, minimizing potential confounding factors related to parameter coverage. More training and inference details are available in Appendix A.

[3] For instance, when translating from German-Czech after fine-tuning on English-Czech and German-English, the model has been implicitly trained to associate German inputs with English outputs. Direct German-Czech translation requires the model to override this spurious correlation.

[4] We translate English instructions with DeepL. For languages not supported by DeepL, we use Google Translate.

3.4 Metrics

For evaluation, we primarily use COMET-22 (Rei et al., 2022) as our main quality metric due to its strong correlation with human judgments (Freitag et al., 2022). However, COMET has known limitations when models generate unintended languages (Zouhar et al., 2024), which is particularly relevant for catastrophic forgetting. Therefore, we include BLEU[5] (Papineni et al., 2002) as a complementary string-matched metric. When appropriate, we also report language accuracy using the language identification tool by Lui and Baldwin (2011).

4 Gain-Forgetting Analyses

We investigate the trade-off between performance gains on fine-tuned language pairs and potential catastrophic forgetting on those unseen during fine-tuning. To visualize this relationship, we create scatter plots (Figure 2 and Figure 3) where each point represents a language pair's performance before (x-axis) and after (y-axis) fine-tuning. The diagonal line ($y = x$) is a reference boundary, where points below indicate catastrophic forgetting, while those above indicate performance improvement.

4.1 Model Scale and Fine-Tuning Data Size

Impact of Model Size Larger model variants consistently exhibit greater resistance to catastrophic forgetting. For M2M-124 models, the 0.6B parameter variant shows fewer language pairs in the forgetting zone compared to its 0.2B counterpart. Similarly for Qwen2.5, the 7B model demonstrates substantially less forgetting than the 0.5B model across all language pairs. This confirms the finding from Ramasesh et al. (2022) that the base model scale helps mitigate forgetting.

Impact of Fine-Tuning Data Volume We additionally observe that the amount of fine-tuning data plays a crucial role in forgetting. By contrasting Figure 2 ($\sim$100K sentences FT data) and Figure 3 ($\sim$ 54M sentences FT data), it becomes clear that higher-data-volume fine-tuning leads to stronger forgetting across all model variants. This observation extends the findings of Ramasesh et al. (2022), by demonstrating that catastrophic forgetting is impacted not only by base model scale, but also by the intensity of task-specific training.

[5] with default tokenizer "13a" in sacreBLEU (Post, 2018), and the dedicated tokenizers for Chinese and Japanese.

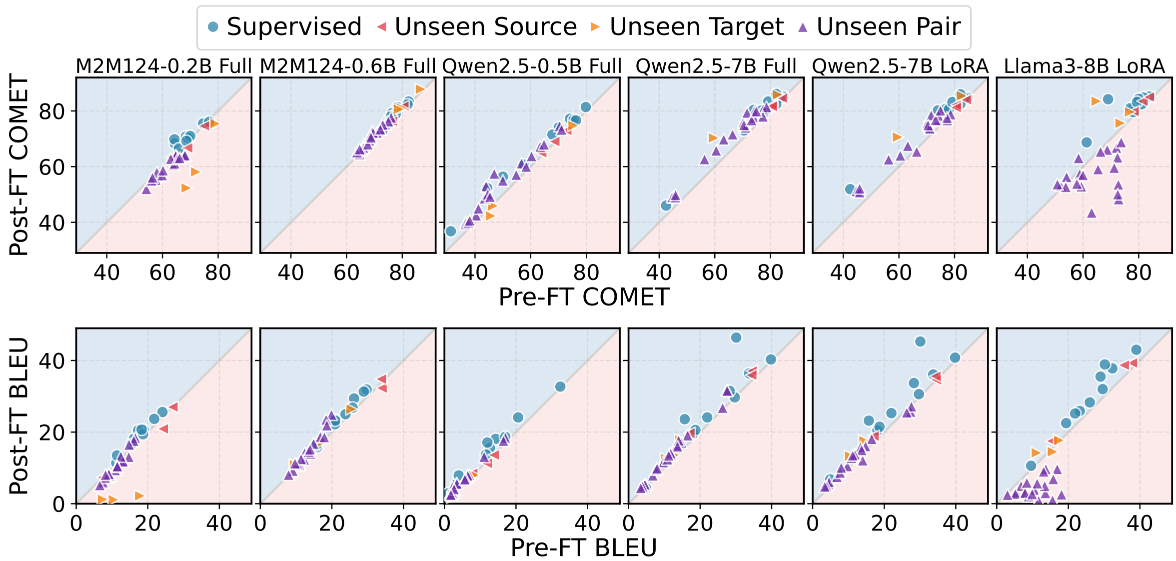


Figure 2: Gain-forgetting plots on the SMALL dataset (117K sentence pairs). Catastrophic forgetting is minimal, except unseen language pairs on Llama (addressed later in Table 4).

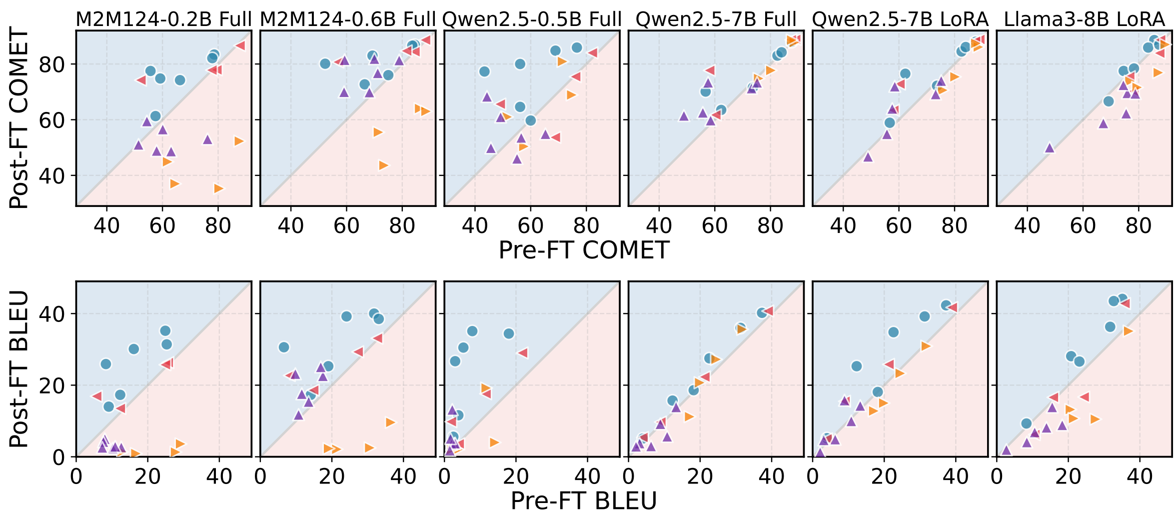


Figure 3: Gain-forgetting plots on the LARGE dataset (54M sentence pairs). Catastrophic forgetting is more severe, especially with translation-specific models where they show performance collapse approaching 0 BLEU.

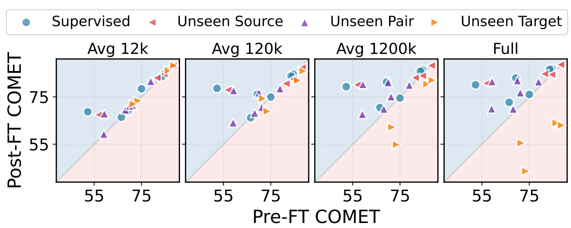


Figure 4: Controlled analysis of data volume effects by subsampled portions of the LARGE dataset. Forgetting becomes severe as fine-tuning data amount increases.

Controlled Analysis of Data Volume Effects To isolate the impact of data volume from dataset-specific factors (e.g., the ALMA dataset has higher-quality data), we conduct controlled experiments using subsampled portions of the WMT21 dataset. We systematically vary the amount of fine-tuning data while maintaining a consistent data source. Starting with 12K sentences per language pair (matching ALMA), we increase the volume by an order of magnitude at each step: 12K $\rightarrow$ 120K $\rightarrow$ 1,200K sentences per language pair. Figure 4 demonstrates the progressive increase in catastrophic forgetting as training volume grows. At 12K sentences per language pair, the gain-forgetting pattern resembles ALMA results, with most language pairs clustered near the diagonal line and minimal performance changes. At 120K sentences, a shift toward forgetting emerges, particularly for target languages unseen in fine-tuning.

Model	Average		en→id	
	COMET	BLEU	COMET	BLEU
Qwen2.5-0.5B-Instruct	64.0	**8.0**	71.6	11.8
+ Instruction-based FT	**65.3**	6.8	**80.9**	**19.2**
+ Token-based FT	60.9	5.2	72.1	14.0

Table 3: Effectiveness of different language control mechanisms on unseen target languages compared to the base model without fine-tuning. Instruction-based language control outperforms token-based control.

At 1,200K sentences, severe catastrophic forgetting occurs, though not to the full extent observed in the complete dataset. The progressive degradation suggests that the intensity of fine-tuning on specific language pairs impacts the forgetting patterns.

4.2 Architecture and Language Control Mechanism

To isolate the impact of model architecture and pretraining objectives on catastrophic forgetting, we compare two models of comparable scale but different designs: M2M-124-0.6B, a translation-specific encoder-decoder model, and Qwen2.5-0.5B, a general-purpose decoder-only model pre-trained for instruction following. While these two models differ in their pre-training data, which precludes a fully controlled comparison, we also replicate M2M-124's language control mechanism on Qwen2.5 to reduce potential confounding factors.

Forgetting Patterns Across Architectures In Figure 3 on the LARGE dataset, where forgetting effects are strongest due to large-scale fine-tuning data, both models exhibit catastrophic forgetting with multiple language pairs falling below the diagonal. However, they differ in their forgetting patterns: M2M-124-0.6B exhibits severe performance degradation on unseen target languages, while Qwen2.5-0.5B shows modest forgetting. We hypothesize that this is related to the target language control mechanisms by the models. As discussed in §3.2, M2M-124 relies on language-specific tokens prepended to both source and target sentences, with the target-side token determining the output language. In contrast, with Qwen, we use natural language instructions to specify the target language, leveraging its existing instruction-following capabilities. This instruction-based mechanism may support more generalizable language control and help mitigate catastrophic forgetting on unseen target languages. We examine this hypothesis next.

Isolating Language Control Mechanisms To test the previous hypothesis that natural language instructions facilitate language control, we conduct a controlled experiment by fine-tuning Qwen2.5-0.5B using the same token-based language specification format as M2M-124, as described in §3.2. This format eliminates natural language instructions entirely, allowing fairer comparisons between models while holding the language control method unchanged. The results support our hypothesis that instruction-following paradigms provide superior language control. As shown in Table 3, when trained with token-based language control, Qwen2.5's performance on unseen target languages drops substantially from 65.3 to 60.9 COMET over 4 unseen target language pairs. To account for low initial performance in some non-English language pairs, we specifically examine the English-Indonesian pair, which has a stronger baseline. In this case, performance still degrades substantially from 80.9 to 72.1 COMET and from 19.2 to 14.0 BLEU. These results on Qwen show that it is the instruction-following ability, rather than the decoder-only architecture, that provides stronger protection against target language forgetting.

Impact of In-Language Instructions Building on our previous findings regarding instruction-following for language control, we investigate whether using instructions in the target language (in-language instructions) can mitigate catastrophic forgetting on unseen language pairs. While prior work on in-language instructions for multilingual LLMs shows mixed results (Marchisio et al., 2024; Mondshine et al., 2025; Liu et al., 2025; Romanou et al., 2025; Enomoto et al., 2025), these studies primarily evaluate models out-of-the-box. In contrast, we focus specifically on the training effects of in-language instructions.

We focus on the Llama3-8B trained on the SMALL dataset, which exhibits strong catastrophic forgetting (rightmost plots in Figure 2). As the results in Table 4 suggest, for unseen language pairs affected by forgetting, in-language instructions substantially outperform English instructions. Specifically, average language accuracy improves dramatically from 22.1% to 82.0%, with corresponding translation quality gains as measured by COMET increasing from 57.2 to 70.9. It is worth noting that this does not impact performance on supervised language pairs, and slightly improves performance on unseen target languages (COMET 79.6→80.3).

	Metric	Supervised	Unseen Pair	Unseen Source	Unseen Target
Original	COMET	76.9	63.8	81.1	71.8
	BLEU	26.0	10.9	29.5	14.8
	LangID	97.9	85.1	97.9	93.5
English instruction	COMET	81.4	57.2	82.8	79.6
	BLEU	30.0	4.4	31.8	15.5
	LangID	96.7	22.1	98.4	94.1
In-language instruction	COMET	81.7	70.9	82.9	80.3
	BLEU	30.3	14.4	32.4	16.1
	LangID	97.5	82.0	98.5	95.1

Table 4: With Llama3 on the SMALL dataset, in-language instructions recover catastrophic forgetting on unseen pairs, reversing a 6.6 COMET loss (63.8→57.2) into a 7.1 COMET gain (63.8→70.9).

4.3 Analyses by Language Pair Types

The results in Figure 2 and Figure 3 also suggest that catastrophic forgetting patterns are strongly dependent on the language pair type. As shown in the previous section, a major issue for language pairs unseen during fine-tuning is generating incorrect output languages. Therefore, we separately discuss the two language control mechanisms.

Token-Based Control and Target Language Forgetting For translation-specific models (M2M-124 variants) which use specialized tokens for language control, performance degradation is most acute for unseen target languages. This is expected, as if the language token for a target language is never encountered during fine-tuning, the model's ability to interpret it and generate the correct language catastrophically degrades.

Unseen Pairs as Main Vulnerability for Instruction-Following Models In contrast, instruction-following models demonstrate greater resilience on unseen target languages, a capability we attribute to the generalizable nature of natural language prompts (§4.2). However, these models are not immune to forgetting and are most susceptible when handling unseen language pairs, where both source and target languages are absent from the fine-tuning set. This is particularly evident with the Llama3-8B model. We hypothesize this vulnerability is compounded by the fact that these unseen pairs are often non-English-centric. Base models typically possess weaker zero-shot capabilities for such translation directions due to the prevalence of English in their pre-training data. Fine-tuning on a different, often English-centric, set of pairs appears to accelerate the forgetting of these already fragile, non-English-centric translation abilities.

4.4 Comparing LoRA and Full Fine-Tuning

We observe that LoRA and full fine-tuning result in comparable levels of catastrophic forgetting (fourth and fifth columns of Figure 2 and Figure 3). Note that we applied LoRA adapters to all components of self-attention and linear projections, thereby minimizing differences in parameter coverage as a confounding factor. Our finding differs from that of Biderman et al. (2024), who observed that LoRA mitigates forgetting when adapting models to dissimilar domains like code and math. We hypothesize that this difference is because our fine-tuning task (translation) requires a smaller domain shift for the base models, which already exhibit strong zero-shot translation capabilities, whereas adapting to code or math requires a larger deviation.

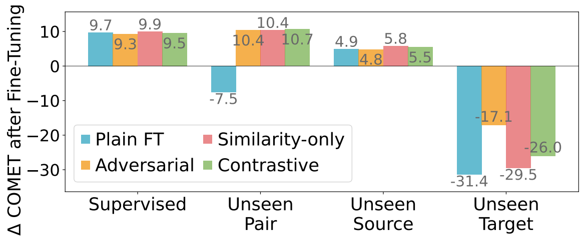
5 Evaluating Cross-Lingual Alignment for Forgetting Mitigation

Having identified the architectural and training factors that impact catastrophic forgetting, we pose a question about mitigation strategies: Do established forgetting mitigation methods primarily restore lost performance, or do they also improve cross-lingual transfer? We focus on cross-lingual alignment methods, as they encourage similar representations for semantically equivalent content across languages, which could mitigate forgetting.

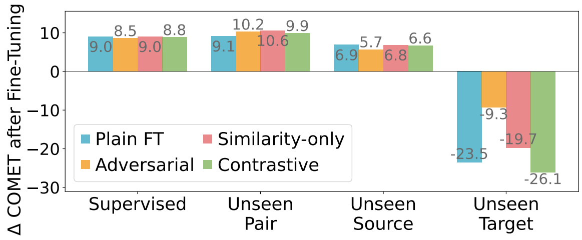
5.1 Evaluated Methods

We evaluate three prominent cross-lingual alignment techniques that encourage shared representations across languages:

- **Adversarial language identification** (Ganin et al., 2016; Arivazhagan et al., 2019): includes an adversarial language classifier that encourages language-agnostic representations by penalizing the model's ability to predict the source language from hidden states.
- **Similarity-only loss** (Arivazhagan et al., 2019; Pham et al., 2019): pulls together translation pairs without negative examples. While a naive implementation would lead to representation collapse, joint training with the translation loss mitigates this by maintaining discriminative power for the primary task (Duquenne et al., 2023).
- **Contrastive loss** (Pan et al., 2021): employs a contrastive objective that pulls together representations of translation pairs while pushing apart representations of unrelated sentence pairs.



(a) M2M-124-0.2B



(b) M2M-124-0.6B

Figure 5: Cross-lingual alignment effects on translation-specific models on the LARGE dataset. Alignment methods bring gains in unseen language pairs, but suffer from persistent forgetting in unseen target languages.

The losses are applied on encoder for encoder-decoder models, and on the middle layers of decoder-only model (Liu and Niehues, 2025).

5.2 Translation-Specific Models

We first evaluate the three alignment methods on translation-specific models: the 0.2B and 0.6B variants of M2M-124. The results are shown in Figure 5, displaying change in translation quality for various language categories, comparing each alignment method against the plain fine-tuning baseline.

Gains in Unseen Language Pairs Among the three unseen categories, alignment methods primarily improve performance on unseen language pairs. These improvements are observed when plain fine-tuning causes forgetting (Figure 5a) and when it brings improvements (Figure 5b). For the 0.2B model, these methods reverse a -7.5 COMET loss by plain fine-tuning (60.5→53.0) into a gain of over 10 COMET. On the larger 0.6B model, the gains are more modest but consistent, ranging from +0.8 to +1.5 COMET over the plain fine-tuning baseline. Besides this category, alignment techniques do not benefit unseen source or target languages, as discussed next.

Persistent Forgetting in Unseen Target Languages The last column of Figure 5 shows that all three approaches still result in drastic, double-digit COMET degradation for this category. This suggests an inherent weakness of the token-based language control mechanism discussed in §4.3. Cross-lingual alignment, while beneficial for transfer, struggle to overcome this fundamental limitation.

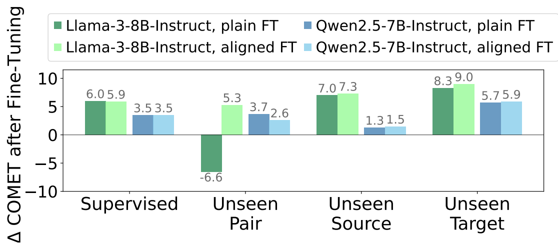


Figure 6: Cross-lingual alignment effects on instruction-following models on the SMALL dataset.

Similar Performance Patterns Across Alignment Methods The three evaluated alignment methods exhibit highly similar performance patterns. While the adversarial approach shows an advantage for unseen target languages (Figure 5), the improvement is insufficient to overcome the severe forgetting in this category. We argue this difference is of limited practical relevance, as the degradation results from an inherent limitation in the token-based language control that none of the methods fully resolve. Moreover, the instruction-based language control already demonstrates superior baseline performance in this setting (§4.2). Therefore, given their comparable overall effectiveness, we select a single representative alignment method for the subsequent analysis of instruction-following models.

5.3 Instruction-Following Models

We choose the contrastive approach for studying instruction-following models due to its generality, as the other two approaches require joint training with task-specific loss to avoid collapse. In Figure 6, results are shown for both Llama3-8B and Qwen2.5-7B with LoRA fine-tuning on both the SMALL and LARGE fine-tuning data configurations.

Impact on Unseen Source and Target Languages On unseen source languages, cross-lingual alignment generally leads to performance comparable to standard LoRA fine-tuning, in line with previous observations on task-specific models (§5.2). On unseen target languages, cross-lingual alignment provides a modest gain of 0.7 COMET (79.6→80.3) for Llama, whereas it offers no significant improvement for the Qwen model. These results suggest that the primary advantage of cross-lingual alignment is its ability to reverse forgetting on unseen language pairs. In conditions where standard fine-tuning already yields improvements, the additional gains from alignment are much milder. This forgetting pattern here, especially in the unseen target category, differs from those observed on translation-specific models in §5.2. The persistent strong forgetting observed previously is substantially reduced, with alignment occasionally surpassing the performance of standard fine-tuning. This suggests that as models move forward in their instruction-following capabilities, their potential for cross-lingual transfer is also enhanced.

Why Gains Concentrate on Unseen Pairs The most significant performance improvements are observed on unseen pairs, where both the source and target languages were included in the training data but never appearing together. As discussed in §3.2, this category is particularly challenging because fine-tuning can cause the model to overfit to spurious source-target associations, leading to outputs in an incorrect target language.

We interpret these results as evidence that cross-lingual alignment methods directly counteract this degradation. Encourage more language-invariant representations leads to disentangling semantic content from language-specific features. By breaking the spurious associations learned during training, alignment mitigates the effects of forgetting and restores the model's ability to generate the correct target language. Consequently, the performance gains are most substantial on these unseen pairs. Considering that the number of translation directions in a multilingual system scales quadratically, and that many languages may only have parallel data to English, breaking the spurious correlations that affect unseen pairs is of high practical importance for scalable translation models.

In contrast, for translation directions involving entirely unseen languages, the central challenge is a general lack of exposure rather than spurious correlations. Therefore, the impact of this alignment mechanism is much milder in those scenarios.

6 Conclusion

In this work, we aim to resolve ambiguities in the literature regarding when catastrophic forgetting occurs for multilingual fine-tuning for MT. Based on our findings, we provide the following practical recommendations: **1)** Consider the relative scale between model size and fine-tuning data. Larger datasets may require larger base models to prevent forgetting. **2)** Prioritize models with strong instruction-following abilities over specific architectural choices. **3)** Do not rely solely on parameter-efficient fine-tuning methods as a forgetting mitigation strategy. **4)** For models exhibiting forgetting, cross-lingual alignment is promising for unseen pairs where both source and target languages have been separately seen in fine-tuning. For instruction-following models, we recommend training with in-language instructions as an initial data-oriented approach before proceeding with cross-lingual alignment approaches.

Limitations

Our study has several limitations that should be considered when interpreting the results:

- Our translation experiments focus on English-centric language pairs, which reflects real-world data availability. Extension to non-English pivot scenarios would provide additional validation of our findings' generalizability.
- While we vary model and data scales systematically, computational constraints limit our exploration to larger size ranges. The dynamics of forgetting in even larger models remain to be investigated.
- We focus on machine translation as it provides a well-structured testbed for studying multilingual forgetting with clear evaluation metrics. Whether similar patterns emerge across other multilingual tasks remains an open question beyond the current scope.

Acknowledgement

We thank the reviewers for their helpful feedback. Part of this work was funded by the KiKIT (The Pilot Program for Core-Informatics at the KIT) of the Helmholtz Association. The authors gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (https://www.nhr-verein.de/en/our-partners). HoreKa is partly funded by the German Research Foundation (DFG).

References

AI @ Meta, Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, and 543 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *Preprint*, arXiv:1903.07091.

Alexandre Berard. 2021. Continual learning in multilingual NMT via language-specific embeddings. In *Proceedings of the Sixth Conference on Machine Translation*, pages 542–565, Online. Association for Computational Linguistics.

Dan Biderman, Jacob 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. *Transactions on Machine Learning Research*. Featured Certification.

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: Research Track*, pages 156–169, Nagoya Japan.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

Taisei Enomoto, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2025. A fair comparison without translationese: English vs. target-language instructions for multilingual LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 649–670, Albuquerque, New Mexico. Association for Computational Linguistics.

Johannes Eschbach-Dymanus, Frank Essenberger, Bianka Buschbeck, and Miriam Exel. 2024. Exploring the effectiveness of LLM domain adaptation for business IT machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 610–622, Sheffield, UK. European Association for Machine Translation (EAMT).

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35.

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. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Shuhao Gu and Yang Feng. 2020. Investigating catastrophic forgetting during continual training for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4315–4326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746, Miami, Florida, USA. Association for Computational Linguistics.

Damjan Kalajdzievski. 2024. Scaling laws for forgetting when fine-tuning large language models. *CoRR*, abs/2401.05605.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2023. Extending multilingual machine translation through imitation learning. *Preprint*, arXiv:2311.08538.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2022. Learning an artificial language for knowledge-sharing in multilingual translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 188–202, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15979–15996, Vienna, Austria. Association for Computational Linguistics.

Junpeng Liu, Kaiyu Huang, Hao Yu, Jiuyi Li, Jinsong Su, and Degen Huang. 2023. Continual learning for multilingual neural machine translation via dual importance-based model division. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12011–12027, Singapore. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond English: The impact of prompt translation strategies across languages and tasks in multilingual LLMs. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 81–104, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nat.*, 630(8018):841–846.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

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.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Qwen Team, 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, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. 2022. Effect of scale on catastrophic forgetting in neural networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa 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)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aquia Richburg and Marine Carpuat. 2024. How multilingual are large language models fine-tuned for translation? In *First Conference on Language Modeling*.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, Marzieh Fadaee, Sara Hooker, Antoine Bosselut, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, and 40 others. 2025. INCLUDE: evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Danielle Saunders and Steve DeNeefe. 2024. Domain adapted machine translation: What does catastrophic forgetting forget and why? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12660–12671, Miami, Florida, USA. Association for Computational Linguistics.

Simeng Sun, Maha Elbayad, Anna Sun, and James Cross. 2023. Efficiently upgrading multilingual machine translation models to support more languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1513–1527, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. 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, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D. McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn. 2018. Freezing subnetworks to analyze domain adaptation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 124–132, Brussels, Belgium. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Thuy-trang Vu, Shahram Khadivi, Xuanli He, Dinh Phung, and Gholamreza Haffari. 2022a. Can domains be transferred across languages in multi-domain multilingual neural machine translation? In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 381–396, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022b. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 89–99, Online. Association for Computational Linguistics.

Genta Winata, Lingjue Xie, Karthik Radhakrishnan, Shijie Wu, Xisen Jin, Pengxiang Cheng, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2023. Overcoming catastrophic forgetting in massively multilingual continual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 768–777, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Junhong Wu, Yuchen Liu, and Chengqing Zong. 2024. 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)*, pages 7180–7192, Mexico City, Mexico. Association for Computational Linguistics.

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.

Changtong Zan, Liang Ding, Li Shen, Yibing Zhen, Weifeng Liu, and Dacheng Tao. 2024. Building accurate translation-tailored llms with language aware instruction tuning. *Preprint*, arXiv:2403.14399.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

A Training and Inference Details

Implementation Frameworks: The M2M-124 experiments were conducted using FairSeq (Ott et al., 2019), while Qwen and Llama experiments utilized Hugging Face Transformers (Wolf et al., 2020).

Training For M2M-124, we used a batch size of 16,384 target tokens. For Qwen and Llama models, we used a batch size of 128 sentences. With M2M-124, we applied a warmup period of 2,500 steps with a learning rate of 1e-4. Training was limited to a maximum of 500K updates, with validation runs every 2,000 steps. Early stopping was triggered if validation loss does not improve for 10 consecutive runs. With Qwen and Llama, we used a warmup period of 200 steps with a default learning rate of 5e-4. For full fine-tuning of Qwen-7B and Llama-8B, the learning rate was reduced to 1e-4 to due to training instability with higher rates. Validation was conducted every 200 steps, with early stopping applied after 5 consecutive runs without improvement. Both model families employed an inverse square root learning rate schedule.

Decoding During inference, we used beam search with a beam size of 5 for M2M-124 experiments, while greedy search was applied for Qwen and Llama models, following Alves et al. (2024).