

Fine-Tuning LLMs for Translation: General Forgetting Mitigation Does Not Preserve MT-Specific Instruction Following

Niklas Scholz¹ David Thulke^{1,2} Abdallah Nasir^{1,3}
Will Allred¹ Evgeny Matusov¹ Hermann Ney^{1,2}

¹AppTek GmbH, Aachen, Germany

²Machine Learning and Human Language Technology, RWTH Aachen University, Germany

³Applied Science Private University, Amman, Jordan

Abstract

Fine-tuning large language models on parallel data improves translation quality but can cause catastrophic forgetting. Mitigation methods are generally evaluated by retention on general benchmarks. We ask whether these findings transfer to machine translation (MT) fine-tuning and to MT-specific instruction following (MT-IF): instructions that modify a translation, such as formality, grammatical gender, and length control. We compare methods anchored to auxiliary data, to model outputs, and to the base model parameters, first in a screening study with Llama 3.2 1B Instruct, then on Llama 3.1 8B Instruct fine-tuned on bidirectional Arabic-English or Spanish-English data. Elastic Weight Consolidation preserves general capabilities best in both stages; on the 8B Spanish model the average score on general benchmarks drops 1.7 points versus 11.0 for standard fine-tuning, yet its scores for formality and grammatical gender control remain close to standard fine-tuning. Only data mixing with control-task examples preserves these controls, but its gains do not transfer to unseen prompts for the same task.

1 Introduction

Instruction-tuned large language models (LLMs) translate well between high-resource languages (Robinson et al., 2023) and can modify their translations in response to natural language instructions. Translation quality is lower for language pairs and domains that are poorly represented in the training data. Fine-tuning on parallel data can reduce this gap. In our experiments, it raises COMET from 45.6 to 71.5 for Amharic-to-English translation.

Fine-tuning on a narrow task can cause catastrophic forgetting, in which performance degrades on capabilities not represented in the fine-tuning data. One capability at risk during MT fine-tuning is following instructions that modify a translation.

Prompt: Translate the following sentence into Spanish while maintaining a formal tone.
Sentence: I am with you there, what is your favorite movie?

COMET: translation quality, General: avg. general capabilities, ↑ is better

Llama 3.1 8B Instruct COMET 86.0 General 66.8
Estoy de acuerdo **con usted**,
¿cuál es **su** película favorita? ✓ formal

Standard SFT COMET 87.0 General 55.8
Estoy de acuerdo **contigo**,
¿cuál es **tu** película favorita? ✗ informal

Elastic Weight Consolidation COMET 87.0 General 63.3
Estoy de acuerdo **contigo**,
¿cuál es **tu** película favorita? ✗ informal

Data Mixing COMET 87.0 General 60.6
Estoy de acuerdo **con usted**,
¿cuál es **su** película favorita? ✓ formal

Figure 1: Formality-controlled translation from CoCoA-MT (Nadejde et al., 2022) before and after fine-tuning Llama 3.1 8B Instruct on 1M Spanish↔English sentence pairs, with each model’s COMET (translation quality, FLORES es↔en) and average General capabilities scores. EWC recovers most of the general-capability loss but not the instructed formality; only Data Mixing with control-task examples retains it.

For example, a prompt can request a formal translation, select the masculine or feminine forms used for a person mentioned in the source, or impose a length constraint. Conventional neural machine translation (NMT) systems can support these controls, but typically require dedicated mechanisms and annotated training data. In applications such as subtitling, the controls can be part of the translation task (Matusov et al., 2019).

A number of methods to mitigate forgetting have been proposed. They differ in the reference signal they use to mitigate forgetting, which we call the anchor: auxiliary data that represents the capabilities to be kept, the output distribution of a model, or the parameters of the base model (Section 3). Ex-

isting evaluations fine-tune on data from domains such as mathematical reasoning and code, and measure retention with general benchmarks (Biderman et al., 2024; Yang et al., 2024; Diao et al., 2026). It remains unclear how well the mitigation methods transfer when MT is the fine-tuning task, and whether they protect MT-specific instruction following (MT-IF).

We address these questions for instruction-tuned models in two stages. Stage 1 is a screening study with Llama 3.2 1B Instruct, fine-tuned separately on Amharic-to-English, Arabic-to-English, and Spanish-to-English data. We compare the mitigation methods of Section 3 across hyperparameter settings, measuring translation quality and general capabilities (Section 4). Stage 2 is the main experiment. We apply selected configurations to Llama 3.1 8B Instruct and fine-tune it on bidirectional Arabic-English or Spanish-English data. In addition to translation quality and general capabilities, we evaluate MT-IF on formality, grammatical gender, and length control (Section 4.3).

On the 1B model, standard fine-tuning decreases the average score on the general benchmarks by 6.3 to 8.1 points across the three language pairs. Elastic Weight Consolidation (EWC) limits this decrease to 2.5 or less, and its configurations trace the Pareto frontier of learning against forgetting. The Stage 2 results show a different pattern for MT-IF. On the 8B model, EWC retains general capabilities, but its scores for formality and grammatical gender control remain close to those of standard fine-tuning (illustrated in Figure 1). Among the methods evaluated in Stage 2, only data mixing with examples of the corresponding control task preserves these controls. Ablations show no consistent transfer between grammatical gender and formality control. Our contributions are:

1. A comparison of forgetting mitigation methods anchored to auxiliary data, to model outputs, and to the base model parameters, for MT fine-tuning of instruction-tuned LLMs, across three language pairs and two model sizes.
2. Evidence from formality, grammatical gender, and length control that retention on general benchmarks can coexist with losses in MT-IF, so general benchmark scores alone can miss capabilities affected by MT fine-tuning.

2 Related Work

Our study draws on two lines of research, the mitigation of catastrophic forgetting during LLM fine-tuning and the control of translation attributes such as formality, grammatical gender, and length.

Forgetting when fine-tuning LLMs on MT data.

Stap et al. (2024) observe that fine-tuning pretrained LLMs on parallel data improves translation quality but degrades formality steering, few-shot translation, and document-level translation. They fine-tune pretrained-only base models, elicit the affected abilities with few-shot prompting, and evaluate a single mitigation method, mixing in monolingual data. We instead start from instruction-tuned models, compare eleven methods, and measure zero-shot instruction following, the interface through which these controls are used in practice. Liu and Niehues (2025) study the loss of translation ability for languages unseen during fine-tuning and find that the relative scale of model and data is a primary determinant, while parameter-efficient fine-tuning offers no clear advantage over full fine-tuning.

Forgetting mitigation for LLM fine-tuning. We draw the methods we compare from this literature (Kirkpatrick et al., 2017; Rolnick et al., 2019; Hu et al., 2022; Tian et al., 2024; Yang et al., 2026; Zheng et al., 2025; Wu et al., 2025; Diao et al., 2026) and describe each of them in Section 3. The evidence behind them was collected outside MT. Earlier approaches were validated on supervised and reinforcement learning benchmarks (Kirkpatrick et al., 2017; Rolnick et al., 2019) and on vision and language classification tasks (Tian et al., 2024), while recent work fine-tunes on mathematical reasoning, code, medical, or agentic data and measures retention with general benchmarks (Biderman et al., 2024; Wu et al., 2025; Diao et al., 2026).

Our experiments focus on mitigation during off-policy SFT, where the training responses are fixed rather than sampled from the current model. Recent comparisons, including one on MT, find that on-policy reinforcement learning often retains prior capabilities better than SFT (Shenfeld et al., 2026; Chen et al., 2026; Su et al., 2026), although Luo et al. (2026) report substantial forgetting during continual reinforcement learning on multimodal reasoning tasks. These studies do not evaluate retention of translation-specific instruction following.

Controlled translation in conventional NMT. Conventional NMT supports controlled translation of attributes such as politeness, length, grammatical gender, and formality through labeled data and dedicated mechanisms, including source- and word-level tags, modified positional encodings, and constrained rescoring or reranking (Sennrich et al., 2016; Lakew et al., 2019; Saunders et al., 2020; Wu et al., 2023). Raunak et al. (2024) fine-tune a single NMT model on parallel and task-specific instruction data, where each instruction is a short task description, such as “informal”, prepended to the source inside dedicated demarcation tokens. Instruction-tuned LLMs instead accept such requests as free-form natural language prompts. We study whether fine-tuning on parallel data degrades this ability and whether forgetting-mitigation methods preserve it.

MT instruction-following benchmarks. Attribute-specific MT benchmarks evaluate phenomena relevant to controlled translation. WinoMT (Stanovsky et al., 2019) and MuST-SHE (Bentivogli et al., 2020) target gender translation accuracy, CoCoA-MT (Nadejde et al., 2022) provides contrastive formal and informal references, MT-GenEval (Currey et al., 2022) adds gender-balanced counterfactual and contextual examples, and mGeNTE (Savoldi et al., 2025) instead evaluates gender-neutral translation. We adapt CoCoA-MT and MT-GenEval, together with FLORES, into targeted evaluations of formality, grammatical gender, and length control, as described in Section 4.3. Broader recent benchmarks assess translation and instruction adherence jointly. IF-MT tests two to four verifiable instructions for English-to-Chinese and English-to-Spanish translation and scores translation quality and adherence separately (Rei et al., 2026). Concurrent work introduces IFMTBench, which evaluates single and composed instructions across seven languages and several lexical, contextual, structural, and stylistic constraints (Sun et al., 2026).

3 Mitigation Methods

The compared methods differ in the reference signal they use to mitigate forgetting during fine-tuning. We call this signal the anchor and group the methods by it: auxiliary data that represents the capabilities to be kept (Section 3.1), the output distribution of a model on the fine-tuning data (Section 3.2), and the parameters of the base model

(Section 3.3). The anchor bounds what a method can preserve, which we return to in Section 7. Several of the methods are orthogonal and could be combined, but we evaluate each in isolation to compare their individual contributions.

3.1 Anchoring to Auxiliary Data

Data Mixing (Rolnick et al., 2019) replays a fraction of general-domain instruction data into the fine-tuning data instead of training on MT alone.

Model Merging (Yang et al., 2026) combines the parameters of two independently trained models. We merge a model fine-tuned on MT with a model trained on general-domain data by linear interpolation, that is, weight averaging.

Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) adds a quadratic penalty on the distance from the base parameters, scaled by a hyperparameter λ and weighted per parameter by the diagonal of the empirical Fisher information estimated on the auxiliary data. Its penalty is in parameter space, but the importance estimate that shapes it comes from the auxiliary data, which is why we group it here. Liu and Chang (2026) propose **EWC-DR**, which reverses the logits when computing Fisher information and thereby reduces the redundant protection of parameters that the standard EWC applies.

All methods in this group draw their anchor from the same subset of the Tülu 3 SFT mixture (Appendix A.1), so what they retain is bounded by the capabilities that this data covers.

3.2 Anchoring to Model Outputs

KL Divergence Regularization. We add the exact KL divergence between the output distributions of the model being trained and a frozen copy of the base model as a penalty scaled by λ on the SFT loss (Shenfeld et al., 2026).

Selective Token Masking (STM) removes tokens from the loss instead of penalizing divergence. Wu et al. (2025) observe that fine-tuning on LLM-generated responses forgets less than fine-tuning on ground-truth data, tracing this to the lower token-level perplexity of generated text, and therefore mask ground-truth tokens whose perplexity exceeds a fixed threshold. We instead set the threshold dynamically and mask the $X\%$ of tokens with the highest perplexity under a frozen copy of the base model.

Confident Conflict Masking (CC). Diao et al. (2026) identify confident conflicts, tokens with low probability and low entropy, as a cause of forgetting, because fitting them overwrites a confident prior. Masking the tokens in the bottom $X\%$ of both quantities mitigates forgetting, but discards useful training signal.

Entropy-Adaptive Fine-Tuning (EAFT) addresses this by weighting the loss of each token by its entropy normalized over the top- K distribution, instead of discarding the token (Diao et al., 2026).

CC and EAFT take the output from the model being trained rather than from a frozen copy. All methods in this group are computed on the MT fine-tuning data, so they bound how far the model moves on translation inputs and place no constraint on its behavior on other prompts.

3.3 Anchoring to the Base Model Parameters

Freeze keeps the bottom N layers at their base values and trains the remaining ones. Zheng et al. (2025) propose this on the grounds that task alignment is concentrated in the bottom layers and is disrupted during the first optimization steps.

Selective Projection Decay (SPD) (Tian et al., 2024) modifies the Adam optimizer to apply weight decay toward the pre-trained weights selectively. A gradient condition determines for each parameter tensor whether the update is heading toward a higher-loss region, and only those tensors receive the penalty, weighted by λ , while the remaining ones are updated without intervention.

Low-Rank Adaptation (LoRA) Hu et al. (2022) restricts fine-tuning updates to a low-rank subspace, which is believed to keep the model close to its base weights and thus limit forgetting (Sun et al., 2023). However, Biderman et al. (2024) show this holds only insofar as LoRA also learns less, with high-rank LoRA forgetting nearly as much as full fine-tuning in some settings.

4 Evaluation of Model Capabilities

We evaluate translation quality with COMET and BLEU, retention on general benchmarks, and three MT-specific controls: formality, grammatical gender, and length. These controlled-translation evaluations are targeted diagnostics of forgetting rather than a general benchmark of MT instruction following.

4.1 Translation Quality

We report reference-based COMET with Unbabel/wmt22-comet-da (Rei et al., 2022) and corpus BLEU with sacreBLEU (Post, 2018), using the default 13a tokenizer for our language pairs. Before scoring, hypotheses and references are punctuation- and whitespace-normalized.

4.2 General Capabilities

Our general-capability benchmarks are a subset of the Tülu 3 evaluation suite (Lambert et al., 2025), which covers knowledge recall, reasoning, math, coding, and instruction following. From its development split, we take GSM8K (Cobbe et al., 2021), DROP (Dua et al., 2019), TruthfulQA (Lin et al., 2022), Codex HumanEval (Chen et al., 2021), Codex HumanEval+ (Liu et al., 2023), and IFEval (Zhou et al., 2023), the last of which assesses general instruction following. From its unseen split, we take AGIEval English (Zhong et al., 2024) and GPQA (Rein et al., 2024), both zero-shot with chain-of-thought prompting. We run all of them with the task formulations and prompts of the Tülu 3 evaluation regime, using the OLMES evaluation standard (Gu et al., 2025). We report the mean of the eight task scores as the general average (General Avg).

4.3 Machine Translation-Specific IF

To evaluate MT-IF capabilities, we assess models on the following three tasks closely related to MT:

Formality Control. To assess formality, we prompt the model to produce a formal or informal translation on the CoCoA-MT dataset (Nadejde et al., 2022) for en→de and en→es, and score with its metric: a hypothesis is classified as formal if it contains at least one of the annotated formal marked phrases from the reference pair and no informal ones (symmetrically for informal), and is left unmatched otherwise. The reported accuracy is the proportion of matched hypotheses whose classified formality matches the requested tone.

Grammatical Gender Control. To assess control over grammatical gender, we use the MT-GenEval benchmark (Currey et al., 2022), which provides counterfactual sentence pairs that differ in the gender forms used for a person mentioned in the text. We instruct the model to produce a translation with the opposite gender forms of the source sentence (i.e., masculine → feminine, feminine

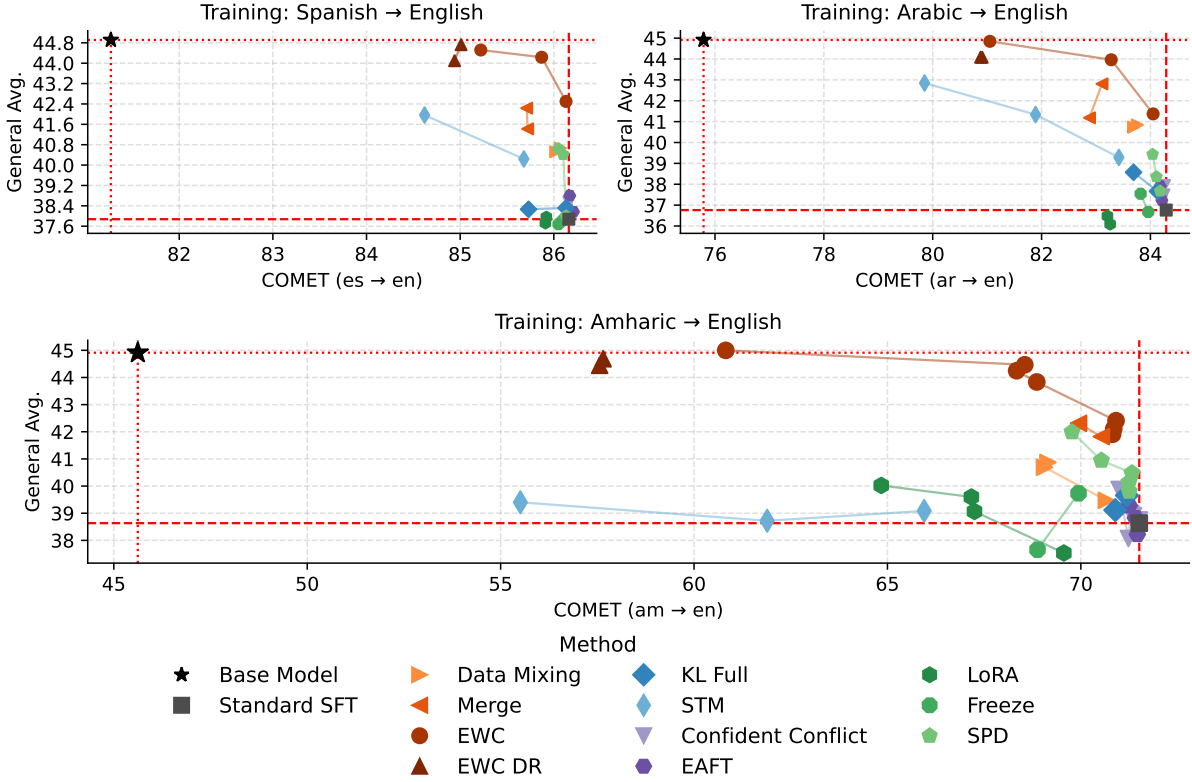


Figure 2: Stage 1: Translation quality (COMET) versus general capabilities (General Avg) for all evaluated configurations per training language on the Llama 3.2 1B Instruct model. Lines connect hyperparameter settings within each method; dashed lines mark the SFT baseline.

→ masculine). The instruction asks the model to override the forms suggested by the source, which separates instruction following from the model’s default choice of gender forms. We follow the same evaluation procedure from the benchmark by checking for the absence of unique opposite-gender words in the hypothesis.

Length Control. We adapt the FLORES-101 and FLORES-200 benchmarks (Goyal et al., 2022; Costa-jussà et al., 2024) for the bidirectional translation between English and the training language. In particular, we add additional instructions on producing shorter or longer translations than the character count of the source. As the produced sequence length is heavily influenced by fine-tuning, we compare whether the instruction affected the output by comparing the translation with longer/shorter instructions to the baseline translation (instead of measuring pure accuracy on IF). This shows the effect of the additional instruction on length control and whether the model actually considered the instruction. We additionally report the target/source (tgt/src) ratios to understand length trends better.

5 Stage 1: Transfer of Mitigation Techniques to Machine Translation Fine-tuning

5.1 Data

We fine-tune models on unidirectional translation data for three language pairs: Amharic → English, Arabic → English, and Spanish → English, each containing 1,000,068 samples. For Data Mixing, Model Merging, and for computing the Fisher matrix in EWC, we use subsets of the Tulu 3 SFT mixture (Lambert et al., 2025) of varying sizes, from which we remove synthetic instruction-following data to prevent overlap with the IFEval benchmark used for evaluation (see Appendix A.1 for more details on the distribution of our training data).

5.2 Experimental Setup

In the first stage of our experiments, we fine-tune the Llama 3.2 1B Instruct model with the mitigation methods previously introduced, comparing various hyperparameter settings. We compare the fine-tuned models against two baselines: the unmodified instruct model and standard SFT.

Anchoring	Method	Hyperparameters	am→en		de→en		General Avg
			COMET	BLEU	COMET	BLEU	
Llama 3.2 1B Instruct			45.6	0.2	82.9	27.6	44.9
Standard SFT			<u>71.5</u>	10.7	83.6	30.7	38.6
Auxiliary Data	Data Mixing	Tülu (full)	69.1	8.5	85.7	34.2	40.7
	Merge	Trans=0.75, Tülu=0.25	70.5	9.9	84.2	<u>32.1</u>	<u>41.8</u>
	EWC	$\lambda=1e2$, Tülu (75k)	70.9	9.9	83.9	31.4	42.4
	EWC-DR	$\lambda=1e3$, Tülu (10k)	71.3	10.6	83.7	30.7	38.6
Model Outputs	KL-Full	$\lambda=0.0025$	71.2	10.6	83.6	30.7	39.7
	STM	X=0.15	65.9	6.7	82.7	27.6	39.1
	CC	X=0.45	71.0	10.5	83.9	31.3	39.9
	EAFT	K=20, Weight=1.0	71.6	11.0	84.0	31.8	38.9
Base Model Parameters	Freeze	N=1	69.9	9.4	83.5	30.5	39.7
	SPD	$\lambda=2$	71.3	<u>10.9</u>	83.9	31.4	40.5
	LoRA	Rank=64, lr=1e-5	67.2	7.5	83.9	31.1	39.6

Table 1: Stage 1: Best-performing configuration per method for Amharic, selected by the highest sum of Amharic-language COMET and general average score. The cell shading encodes the absolute differences from the base instruct model, normalized relative to the largest observed change within each metric group (across languages).

Evaluation. We report BLEU and COMET scores for translation on the training language pair and on German-to-English, included as an out-of-distribution language pair (Section 4.1). We use FLORES-101 (Goyal et al., 2022) for Amharic and Spanish, and FLORES-200 (Costa-jussà et al., 2024) for Arabic and German. We additionally evaluate the forgetting of general capabilities, as described in Section 4.

5.3 Results

We present the MT scores and the general average of the best-performing configurations for Amharic in Table 1 (Arabic and Spanish in Appendix A.8). These configurations are chosen based on the highest sum of COMET and general average scores. Additionally, we visualize the results of all 40 Amharic, 27 Spanish, and 28 Arabic configuration training runs in Figure 2 showing the trade-off between general capabilities and MT performance.

Overall, we observe the SFT baselines of all three languages to maintain similar forgetting rates (Amharic: 6.3, Arabic: 8.1, Spanish: 7.0) despite the different improvements in learning the MT task (Amharic: 25.9, Arabic: 8.5, Spanish: 4.9). Even though German was not included as a language pair in training, German MT scores slightly improve after fine-tuning, indicating no forgetting on a similar translation task.

Most methods collapse toward the SFT baseline.

Across all three training languages, the majority of evaluated configurations are clustered around

the SFT baseline in Figure 2 (dashed lines). Only a small number of methods consistently separate from this cluster by forgetting less on general capabilities (General Avg) at comparable COMET scores. In particular, the EWC family (EWC, EWC-DR) and Merge appear in the upper-right region in all three plots, while Data Mixing and SPD show mediocre mitigation of forgetting.

Most methods are consistent across languages.

The only two techniques showing clear differences between languages are Selective Token Masking (STM) and LoRA, each diverging in opposite directions when comparing Amharic with Spanish and Arabic. LoRA performs comparatively better on Amharic by trading a smaller gain on MT performance (COMET: +21.6 vs. +25.9 in standard SFT) for a slightly smaller drop in the general average (−5.3), whereas STM performs better on Arabic and Spanish than for Amharic where it significantly hindered learning (COMET: 65.9).

EWC lies on the Pareto frontier. Individual EWC configurations trade translation quality against general capabilities, spreading along the upper-right boundary of Figure 2 between standard SFT and the base model rather than occupying a single point. Any method with a continuous coefficient produces such a curve between the two baselines, and the sweeps of Merge, Data Mixing, KL, and STM reach the region where COMET falls below standard SFT. Their curves lie below the EWC curve: for almost every configuration of another method, some EWC setting reaches both

a higher COMET and a higher general average, in all three languages. The remaining frontier points are configurations of standard SFT and a few other methods that gain up to a few tenths of COMET over the best EWC setting at the cost of several points on the general average.

Implications for Stage 2 Experiments. The language-dependent differences of LoRA and STM noted above suggest that the extent of learning required, rather than the language itself, drives the differences in forgetting mitigation trends. As we are not aware of resources to evaluate MT-IF on Amharic, we reduce the Stage 2 experiments to Spanish and Arabic. Additionally, as EWC performed better than EWC-DR, we drop the latter.

6 Stage 2: Comparison to Machine Translation Instruction Following Tasks

6.1 Data

We fine-tune models with bidirectional, open data from two languages to improve reproducibility: Spanish $\leftrightarrow$ English and Arabic $\leftrightarrow$ English. Each dataset has a total size of 1,000,000 samples, consisting of 500,000 pairs per direction (see [Appendix A.1](#) for detailed information on the training data distribution).

Alongside the Tülu 3 subsets used in Stage 1, we mix in data closely related to our MT-IF tasks: the training split of CoCoA-MT (Spanish, N=400; German N=400) ([Nadejde et al., 2022](#)) and the dev split of MT-GenEval (Arabic, Spanish, and German evenly distributed; N=800) ([Currey et al., 2022](#)). We primarily use the same prompts as during evaluation, but additionally run variants with a different prompt to assess whether the observed effects generalize to unseen prompts (see [Appendices A.5](#) and [A.6](#) for prompts used).

6.2 Experimental Setup

We move to Llama 3.1 8B Instruct, whose base performance on the MT-IF tasks is high enough to measure their forgetting, and to open, bidirectional training data. Model size, translation directions, and data source therefore change together between the two stages. We take the best performing configurations per method identified in the previous stage as a starting point and only slightly re-tune them for the larger model trained on open, bidirectional data, in contrast to re-screening all possible configurations.

Evaluation. We reuse the Stage 1 evaluation setup (COMET, BLEU, and the general average; [Section 5.2](#)), but evaluate translation in both directions. Beyond these scores, we evaluate our models on the three MT-IF control tasks, as defined in [Section 4.3](#): Formality Control (Spanish, German; only for Spanish training runs), Grammatical Gender Control (Training Language, German), and Length Control (Training Language). We evaluate MT-IF only in Stage 2, as CoCoA-MT and MT-GenEval only cover translations from English to other languages and the MT-IF capabilities of the 1B model are too low to measure forgetting. We summarize the controls as the MT-IF average (MT-IF Avg): the mean of the training-language scores of the three tasks, where formality and grammatical gender contribute their average accuracy and length control contributes the mean compliance rate under the shorter and longer instructions. For Arabic, the average covers only grammatical gender and length control, as CoCoA-MT provides no Arabic formality data.

6.3 Results

6.3.1 Forgetting of General Capabilities

We present the MT scores and the general average of our bidirectional training runs on Spanish in [Table 2](#) and on Arabic in [Table 20](#) (Appendix). Standard SFT forgets more on general capabilities at 8B than at 1B in Stage 1, with the general average (base 66.8) dropping 11.0 points for Spanish and 8.3 points for Arabic, despite much smaller MT learning gains (COMET +1.0 vs. +4.9 at 1B for Spanish). Since model size, translation directions, and data source change together between the stages, we cannot attribute this difference to one of them. The ordering of the methods is largely stable across stages: EWC retains the most in both, Merge, SPD, and Data Mixing retain moderately, and the methods anchored to model outputs stay within two points of standard SFT. Changes in ordering occur where Stage 1 differences were already small. CC and Freeze, within 1.3 points of standard SFT at 1B, now fall 0.8 to 1.6 points below it. The exception is STM, which retained 4.1 to 4.5 points more than standard SFT for Spanish and Arabic in Stage 1 at the cost of COMET, and at 8B loses 2.5 to 2.8 points more than standard SFT (53.0 for Spanish, 56.0 for Arabic) while still learning less.

EWC retains general capabilities; other methods only partially. Consistent with Stage 1 of our

Anchoring	Method	Hyperparameters	es $\leftrightarrow$ en		de $\leftrightarrow$ en		General Avg	es MT-IF Avg
			COMET	BLEU	COMET	BLEU		
Llama 3.1 8B Instruct			86.0	26.7	87.6	37.3	66.8	71.1
Standard SFT			87.0	30.4	87.4	37.2	55.8	32.3
Auxiliary Data	Data Mixing	Tülu (full)	86.9	30.1	87.9	38.2	60.0	34.5
		MT-IF (1600)	87.0	30.5	87.9	38.4	56.5	65.8
		Tülu (800) + MT-IF (1600)	87.0	30.3	87.8	37.8	60.6	67.0
		+ diff MT-IF prompt	87.0	30.4	87.9	38.0	60.9	33.7
		Tülu (800) + Gender (800)	87.0	30.4	87.8	37.7	61.7	51.2
		Tülu (800) + Formality (800)	87.0	30.3	87.4	37.1	61.6	45.5
	Merge	Trans=0.5, Tülu+MT-IF=0.5	87.0	30.3	87.8	37.9	61.5	36.8
	EWC	$\lambda=1e3$, Tülu (10k)	87.0	30.0	87.6	37.7	65.1	35.3
		$\lambda=1e3$, MT-IF (1600)	87.0	30.4	87.7	38.0	58.1	34.8
		$\lambda=1e3$, Tülu (800) + MT-IF (1600)	87.0	30.2	87.7	37.5	63.3	35.0
Model Outputs	KL-Full	$\lambda=0.05$	86.9	30.1	87.6	37.5	57.8	32.1
	STM	$X=0.25$	86.5	28.2	87.1	35.3	53.0	32.5
	CC	$X=0.25$	87.0	30.4	87.4	36.8	55.0	31.5
	EAFT	$K=20$, Weight=1.0	87.0	30.2	87.6	37.4	57.8	31.9
Base Model Parameters	Freeze	$N=4$	87.0	30.3	87.5	37.2	54.3	32.5
	SPD	$\lambda=5$	86.8	29.9	87.8	37.7	60.2	53.3
	LoRA	Rank=32, Alpha=64	86.9	30.1	87.6	37.8	61.3	33.2

Table 2: Stage 2: MT, General Avg, and Avg MT-IF scores for bidirectional Spanish training on the Llama 3.1 8B Instruct model. Color shade of cells encodes the absolute difference to the base model, normalized within each metric group (across languages).

experiments, EWC with Tülu for parameter importance estimation keeps the drop in the general average close to zero. With only the Tülu subset, EWC limits the drop to -1.7 for Spanish and -0.7 for Arabic, far smaller than the -3.6 to -13.8 drops of other methods. This mirrors the Stage 1 pattern for EWC, now under bidirectional translation with a stronger base model. LoRA, Merge, and SPD all show moderate retention of the general average.

6.3.2 Forgetting of MT-IF Capabilities

We observe that while general capabilities have been preserved strongly or moderately by specific methods, such as EWC, this is not the case for MT-IF (i.e., Formality Control, Grammatical Gender Control, Length Control) (see Figure 3). Standard SFT drops the average MT-IF score by -38.8 for Spanish ($71.1 \rightarrow 32.3$) and -37.8 for Arabic ($66.8 \rightarrow 28.9$). The only methods that separate from the large cluster around this baseline are SPD, with drops of -17.8 for Spanish and -27.1 for Arabic, and Data Mixing with MT-IF data (using the same prompts as evaluation), ranging from -4.1 to -25.6 for Spanish and from -11.7 to -15.0 for Arabic. The Arabic mix containing only formality data, a control task that is not part of the Arabic MT-IF average, even underperforms SFT by 4.6 points.

EWC does not preserve MT-IF. Although EWC retains general capabilities best, it falls into the cluster around the SFT baseline on MT-IF (see Figure 3). EWC with the 10k Tülu subset, the configuration with the smallest drop in the general average for both languages, reaches MT-IF averages of 35.3 for Spanish and 32.7 for Arabic, close to standard SFT (32.3 and 28.9). Estimating parameter importance on additional MT-IF data does not change this: the general average drops further while the MT-IF averages stay near the SFT baseline.

Data Mixing preserves only the target MT-IF capability. We observe that Data Mixing with MT-IF data is the only method that retains most of the MT-IF capabilities (see Figure 3). However, this mitigation is limited to the specific MT-IF task that is included in the training data. For example, Data Mixing with Tülu (800) + Gender (800) is only actively improving on grammatical gender control capabilities (to 77.9% accuracy vs. standard SFT’s 17.8%), while showing more forgetting than standard SFT on Formality Control (53.7% accuracy vs. 56.1%) and Length Control (Table 19). This shows that the retention of MT-IF capabilities is limited to the specific task that is included in the training data, and does not transfer to other MT-IF tasks. On the mixed-in task itself, scores surpass even the base model in all cases except German formality control.

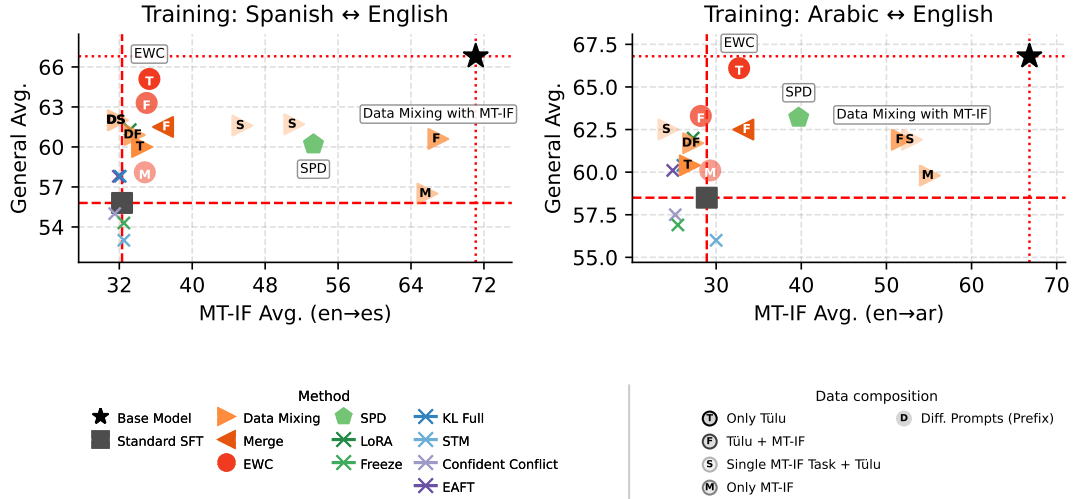


Figure 3: Stage 2: General capabilities (General Avg) versus MT-IF capabilities per training language for fine-tuned, bidirectional Llama 3.1 8B Instruct models (MT-IF averaged over Formality, Grammatical Gender, and Length Control for Spanish; Length and Grammatical Gender Control for Arabic). Dotted/dashed red lines mark the base model / standard SFT reference points. Additional text labels within markers indicate the data composition.

Same-prompt gains do not transfer to unseen prompts on the same task. However, we observe clear differences between using the same prompt as during evaluation and different prompts in Spanish training runs (see Figure 3 for the dashed-edge markers that distinguish these variants). Evidently, for Data Mixing with different Formality Control Prompts, Formality scores reduce to an average accuracy of 53.8, which is even lower than the score of standard SFT (56.1). The recovery of MT-IF capabilities with Data Mixing is thus partially because of prompt memorization, rather than general MT-IF capability retention.

Grammatical Gender and Formality Control. Per-method scores are in Tables 17, 18 and 22 (Appendix). Standard SFT reduces grammatical gender control scores in both training runs and formality control scores in the Spanish run by more than 30 points. The majority of methods (CC, EAFT, Freezer, KL, LoRA, STM) stay close to these scores on both tasks. SPD and Merge stay above these scores on both tasks. On the training language pair, SPD retains more in the Spanish run (38.5 versus 23.4 for Merge on gender control, 72.0 versus 61.0 on formality control), while Merge retains more gender control in the Arabic run (34.3 versus 29.7).

Length Control. Per-method length control scores are in Tables 19 and 21 (Appendix). The base model follows shorter and longer instructions relatively well. After standard SFT, compliance

collapses and the tgt/src ratios of the three control settings converge; almost every other method stays within this range. SPD is the only method that keeps compliance with both instructions clearly above the SFT level. While EWC (Tulu 10k) also partially follows the longer-translation instruction, its COMET scores under this instruction drop to 74.3 for Spanish and 71.1 for Arabic, well below other methods. This suggests that EWC pushes the model toward producing longer sequences without ensuring correctness of translations.

7 Analysis and Discussion

Why does EWC protect general capabilities but not MT-IF? Figure 4 speaks against the simplest cause, a Fisher estimated on Tulu that misses the parameters behind MT-IF: the normalized layer-wise Fisher importance of the three anchor datasets used in Stage 2 has nearly the same profile, with most importance in layers 0–5 and the final layer (parameter-level differences remain possible). Consistent with this, a Fisher estimated on MT-IF data alone did not preserve MT-IF in the Spanish run (average 34.8 versus 32.3 for standard SFT) and lowered the general average (58.1 versus 65.1 with the Tulu Fisher). We therefore attribute the asymmetry to the MT gradients rather than to the importance estimate. EWC with the Tulu Fisher reaches the same COMET as standard SFT (87.0) while keeping the general average at 65.1 versus 55.8, which is only possible if MT fine-tuning updates the pa-

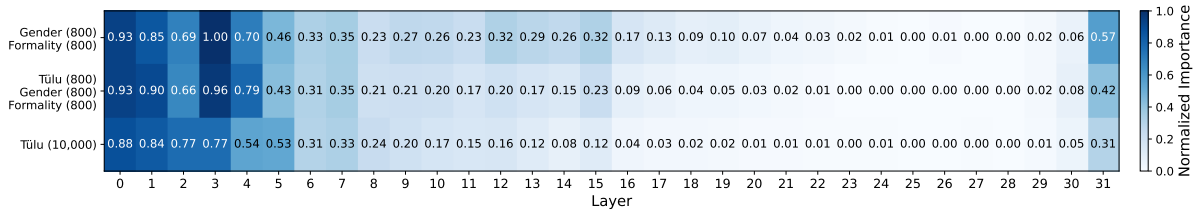


Figure 4: Normalized layer-wise Fisher importance on Llama 3.1 8B Instruct, estimated on Tulu, MT-IF data, and their combination. The importance profiles largely coincide.

rameters behind the general benchmarks weakly. The MT-IF controls, in contrast, are exercised on translation inputs, so the parameters that carry them receive the MT gradient throughout fine-tuning, which a fixed quadratic penalty does not hold back. We did not measure per-parameter drift, so this explanation remains a hypothesis.

Why do the methods anchored to model outputs fail? The methods in this group (KL, STM, CC, EAFT) compute their constraint on the MT training batches (Section 3.2), so they bound the drift only on inputs on which the model must change and leave the evaluated behavior unconstrained. KL regularization, for instance, penalizes divergence from the base model on the translation inputs themselves and places no constraint on the prompts used by the general benchmarks or the MT-IF evaluations. CC additionally masks almost nothing on MT data. In the Spanish 8B run ($X=0.25$), each condition alone is met by roughly a quarter of the tokens in a batch (on average 2773 with low entropy and 2619 with low probability), but fewer than one token per batch (0.36) meets both. Tokens that are improbable under the base model are thus uncertain rather than confidently wrong, and with almost no tokens masked, CC reduces to standard SFT, in line with its scores in Table 2.

SPD is a partial exception. Among the methods without auxiliary data, only SPD separates from the SFT cluster on MT-IF (average drops of 17.8 for Spanish and 27.1 for Arabic, against 38.8 and 37.8 for standard SFT), plausibly because its penalty applies per tensor and step, when the update heads toward a higher-loss region, rather than following a fixed importance estimate. Our results do not isolate this mechanism, and the retained scores remain well below those of data mixing.

Implications. Methods anchored to the base model parameters assume that the fine-tuning task does not need to update the weights that carry the

capability to be preserved. Under MT fine-tuning this holds for general capabilities but not for MT-IF, which is itself translation behavior. Data mixing escapes this limit by continuing to train the control rather than constraining the model, but as the Spanish formality run shows, the retention is tied to the training prompt.

8 Conclusion

We compared forgetting mitigation methods anchored to auxiliary data, to model outputs, and to the base model parameters, for fine-tuning Llama 3.2 1B Instruct and Llama 3.1 8B Instruct on parallel data for three language pairs. EWC retains general capabilities best on both models. On the 8B model, it limits the drop in the general average to 1.7 points for Spanish and 0.7 for Arabic, against 11.0 and 8.3 for standard fine-tuning. Its scores for formality and grammatical gender control nevertheless remain close to those of standard fine-tuning, even when the data used to estimate parameter importance includes control-task examples. Of the methods compared, only mixing control-task examples into the training data preserves the corresponding control, and the effect does not extend to other control tasks or to unseen prompts for the same task. Among the methods without control-task data, SPD retains the most MT-IF, at scores well below data mixing.

We conclude that retention on general benchmarks does not establish that MT-specific instruction following is preserved. Evaluations of MT fine-tuning should measure the controls the application requires, and preserving them currently requires matching data in the training mix. Open questions are whether broader control-task data, for example synthetic examples covering more tasks and prompt variants, can preserve MT-IF without overfitting to the seen prompts, and whether the findings extend to further controls such as glossary constraints and gender-neutral translation.

Limitations

Our comparison is restricted to SFT-based adaptation and does not include on-policy reinforcement learning or preference optimization. As discussed in [Section 2](#), recent prior work showed that on-policy methods retain prior capabilities better than SFT. Including them would have required model-generated trajectories and task-specific rewards or preference pairs, a substantial addition to our framework that would complicate a fair comparison.

We adapt the evaluation for MT-IF from related work and inherit some limitations. For control of formality and grammatical gender, we follow the evaluation approaches proposed in CoCoA-MT and MT-GenEval. These rely on predefined marked phrases and words and some translation variants may be missed. MT-GenEval’s masculine/feminine annotation, which our prompts follow, treats gender as binary and excludes non-binary referents. We use it because its scoring is deterministic and cheap to run, whereas gender-neutral translation currently requires a trained classifier or an LLM judge for evaluation ([Savoldi et al., 2025](#)). Our results therefore cover the selection between masculine and feminine forms and do not show whether MT fine-tuning degrades gender-neutral translation. The MT-IF evaluation also covers only translation from English into Spanish, German, and Arabic, where CoCoA-MT and MT-GenEval provide data, and formality control only for Spanish and German. The length-control score compares each instructed translation with the model’s own uninstructed translation, so it measures whether the instruction changed the output length and not whether a length target was met.

Model size, translation directions, and data source change together between Stage 1 and Stage 2, and the configurations selected on the 1B model are only slightly re-tuned for the 8B model. The design therefore does not show which of these factors causes the larger forgetting of standard SFT at 8B or the reversal for STM, and the Stage 2 configurations may not be optimal for the larger model. We only use data from our internal standard data preparation pipeline which may introduce some biases. Experiments are further limited to two model sizes of the Llama 3 family and to unidirectional and bidirectional translation of English and three target languages (two high and one low-resource). While we assume that our findings generalize to other languages and models, this is still untested.

Acknowledgments

This work was partially supported by NeuroSys, which as part of the initiative “Clusters4Future” is funded by the Federal Ministry of Education and Research BMBF (funding ID 03ZU2106DD).

Generative AI assistants (OpenAI Codex and Claude Code) were used to proofread and help rephrasing parts of the manuscript, to help with LaTeX formatting, and to support with the implementation of the experiments. All model generated suggestions were reviewed and verified by the authors, who take full responsibility for the content of this work.

References

- Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. [Gender in danger? evaluating speech translation technology on the MuST-SHE corpus](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, 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.
- Howard Chen, Noam Razin, Karthik R Narasimhan, and Danqi Chen. 2026. [Retaining by doing: The role of on-policy data in mitigating forgetting](#). In *Forty-third International Conference on Machine Learning*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. [Evaluating large language models trained on code](#). *Preprint*, arxiv:2107.03374 [cs.LG].
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arxiv:2110.14168 [cs.LG].
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Mailhard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John

- Hoffman, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Anna Currey, Maria Nadejde, Raghavendra Reddy Papagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. [MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Muxi Diao, Lele Yang, Wuxuan Gong, Yutong Zhang, Zhonghao Yan, Yufei Han, Kongming Liang, Weiran Xu, and Zhanyu Ma. 2026. [Entropy-adaptive fine-tuning: Resolving confident conflicts to mitigate forgetting](#). *Preprint*, arxiv:2601.02151 [cs.LG].
- Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. [DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs](#). 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 2368–2378, Minneapolis, Minnesota. 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](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. 2025. [OLMES: A standard for language model evaluations](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5020–5048, Albuquerque, New Mexico. Association for 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 *International Conference on Learning Representations*.
- James Kirkpatrick, Razvan Pascanu, Neil 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. 2017. [Overcoming catastrophic forgetting in neural networks](#). *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. [Controlling the output length of neural machine translation](#). In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxin Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, and 4 others. 2025. [TÜLU 3: Pushing Frontiers in Open Language Model Post-Training](#). In *Second Conference on Language Modeling*.
- 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)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Danni Liu and Jan Niehues. 2025. [Conditions for catastrophic forgetting in multilingual translation](#). In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 347–359, Suzhou, China. Association for Computational Linguistics.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. 2023. [Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 21558–21572. Curran Associates, Inc.
- Xuan Liu and Xiaobin Chang. 2026. Elastic weight consolidation done right for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3930–3940.
- Mao-Lin Luo, Zhe-Xu Wang, Zi-Hao Zhou, Bo Ye, Jian Zhao, Min-Ling Zhang, and Tong Wei. 2026. [RL forgets! towards continual policy optimization](#). *Preprint*, arXiv:2607.04364.
- Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. [Customizing neural machine translation for subtitling](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.
- Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. [CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. 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.

- Vikas Raunak, Roman Grundkiewicz, and Marcin Junczys-Dowmunt. 2024. [On instruction-finetuning neural machine translation models](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1155–1166, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Amin Farajian, Pedro Teixeira, and Andre Martins. 2026. [TOWER+: Bridging generality and translation specialization in multilingual LLMs](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29614–29635, San Diego, California, United States. Association for Computational Linguistics.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. [GPQA: A graduate-level google-proof q&a benchmark](#). In *First Conference on Language Modeling*.
- Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. [ChatGPT MT: Competitive for high- \(but not low-\) resource languages](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.
- David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. [Experience replay for continual learning](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. [Neural machine translation doesn’t translate gender coreference right unless you make it](#). In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.
- Beatrice Savoldi, Giuseppe Attanasio, Eleonora Cupin, Eleni Gkovedarou, Janiça Hackenbuchner, Anne Lauscher, Matteo Negri, Andrea Piergentili, Manjinder Thind, and Luisa Bentivogli. 2025. [Mind the inclusivity gap: Multilingual gender-neutral translation evaluation with mGenTE](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13698–13720, Suzhou, China. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Controlling politeness in neural machine translation via side constraints](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.
- Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. 2026. [RL’s razor: Why online reinforcement learning for-gets less](#). In *The Fourteenth International Conference on Learning Representations*.
- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. [Evaluating gender bias in machine translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.
- 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)*, pages 6189–6206, Bangkok, Thailand. Association for Computational Linguistics.
- Zeli Su, Ziyin Zhang, Zhou Liu, Xuexian Song, Zhankai Xu, Longfei Zheng, Xiaolu Zhang, Rong Fu, Guixian Xu, and Wentao Zhang. 2026. [Reinforcement learning with semantic rewards enables low-resource language expansion without alignment tax](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 17772–17786, San Diego, California, United States. Association for Computational Linguistics.
- Mingrui Sun, Mao Zheng, Zheng Li, and Mingyang Song. 2026. [Ifmtbench: A comprehensive benchmark for multilingual translation instruction following](#). *Preprint*, arXiv:2605.28218.
- Simeng Sun, Dhawal Gupta, and Mohit Iyyer. 2023. [Exploring the impact of low-rank adaptation on the performance, efficiency, and regularization of rlhf](#). *Preprint*, arXiv:2309.09055.
- Junjiao Tian, Chengyue Huang, and Zsolt Kira. 2024. [Rethinking Weight Decay for Robust Fine-Tuning of Foundation Models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 22418–22440. Curran Associates, Inc.
- Chao-Chung Wu, Zhi Rui Tam, Chieh-Yen Lin, Yun-Nung (Vivian) Chen, Shao-Hua Sun, and Hung-yi Lee. 2025. [Mitigating forgetting in llm fine-tuning via low-perplexity token learning](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 1708–1744. Curran Associates, Inc.
- Zhanglin Wu, Zongyao Li, Daimeng Wei, Hengchao Shang, Jiabin Guo, Xiaoyu Chen, Zhiqiang Rao, Zhengzhe Yu, Jinlong Yang, Shaojun Li, Yuhao Xie, Bin Wei, Jiawei Zheng, Ming Zhu, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. [Improving neural machine translation formality control with domain adaptation and reranking-based transductive learning](#). In

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 180–186, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2026. [Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications, and opportunities](#). *ACM Comput. Surv.*, 58(8).

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. [Self-distillation bridges distribution gap in language model fine-tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025. [Spurious forgetting in continual learning of language models](#). In *The Thirteenth International Conference on Learning Representations*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. [AGIEval: A human-centric benchmark for evaluating foundation models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arxiv:2311.07911 [cs.CL].

Corpus	Base	CONCAT
OPUS-NLLB	872,787	–
OPUS-CCAligned	47,563	–
OPUS-Tanzil	39,617	–
OPUS-bible-uedin	23,635	–
OPUS-GoURMET	6,279	–
OPUS-MultiCCAligned	4,093	–
OPUS-XLEnt	3,803	–
OPUS-OpenSubtitles	1,289	–
OPUS-GlobalVoices	316	21
OPUS-GNOME	296	4
OPUS-Ubuntu	172	39
OPUS-TED2020	93	4
OPUS-wikimedia	41	–
Proprietary data	16	–
Total	1,000,000	68

Table 3: Number of Amharic sentence pairs contributed by each corpus for the Stage 1 SFT training dataset. Non-OPUS sources are aggregated as proprietary data.

A Appendix

A.1 Information on Training Data

The MT training data consists of two groups. *Base* comprises sentence-level parallel segments drawn from OPUS and other sources. *CONCAT* is built from *Base* segments for which document-level metadata is available: consecutive sentences from the same document are concatenated into a single training example.

Stage 1. For unidirectional training runs on Llama 3.2 1B Instruct, we use Amharic → English, Arabic → English, and Spanish → English data (Tables 3 to 5). Each language pair comprises 1,000,000 *Base* sentence pairs plus 68 additional *CONCAT* samples (1,000,068 total). Non-OPUS sources are aggregated as proprietary data; Amharic is drawn almost entirely from OPUS (primarily NLLB).

Stage 2. For bidirectional training runs on Llama 3.1 8B Instruct, we use only public data from OPUS (Table 6 for Spanish, Table 7 for Arabic). During sampling, we explicitly minimized contributions from larger corpora, since these tend to consist of noisier, crawled data. For both *Base* and *CONCAT*, each language’s 500,000 sentence pairs are split evenly across both translation directions to enable bidirectional training.

A.2 Training & Evaluation Settings

A.2.1 Training

- **Epochs:** 1.0

Corpus	Base	CONCAT
Proprietary data	762,628	22
OPUS-TED2020	61,929	2
OPUS-NeuLab-TedTalks	56,368	3
OPUS-MultiUN	35,934	19
OPUS-UNPC	28,406	18
OPUS-GlobalVoices	13,369	–
OPUS-OpenSubtitles	13,150	4
OPUS-Tatoeba	9,211	–
OPUS-News-Commentary	4,659	–
OPUS-TED2013	3,906	–
OPUS-ELRC	3,846	–
OPUS-KDE4	2,624	–
OPUS-Mozilla-I10n	2,114	–
OPUS-tico-19	946	–
OPUS-Ubuntu	711	–
OPUS-UN	124	–
OPUS-EUbookshop	75	–
Total	1,000,000	68

Table 4: Number of Arabic sentence pairs contributed by each corpus for the Stage 1 SFT training dataset. Non-OPUS sources are aggregated as proprietary data.

- **Learning Rate (Non-PEFT):** 1e-6
- **Learning Rate (PEFT):** 1e-5
- **Optimizer:** AdamW (default weight decay; weight decay and AdamW only adapted for SPD regularization)
- **Warmup Ratio:** 0.003
- **Effective Batch Size:** 64
- **Max Sequence Length:** 1024 tokens
- **Chat Template:** Llama-3 Instruct format, with loss masked to assistant-turn content only.

A.2.2 Evaluation Sampling Arguments

General Capabilities. We use the OLMES setup (Gu et al., 2025) for evaluating general capabilities, inheriting the task-specific sampling arguments of their implementations to ensure reproducibility.

MT(-IF). For MT-IF Evaluation, we use greedy decoding (temperature = 0) with a maximum output length of 512 tokens.

A.3 Computational Resources

For the unidirectional training runs on Llama 3.2 1B Instruct, we used either one A100 with 80GB or one GPU with 48GB. We report the average computation time across languages of the training techniques used for the bidirectional Llama 3.1 8B Instruct training in Table 9.

Corpus	Base	CONCAT
Proprietary data	203,757	15
OPUS-ELRC	90,449	–
OPUS-DGT	88,367	3
OPUS-Europarl	78,739	2
OPUS-EUbookshop	78,002	8
OPUS-SciELO	48,193	–
OPUS-GlobalVoices	47,387	1
OPUS-ELITR-ECA	40,974	4
OPUS-MultiUN	39,451	3
OPUS-JRC-Acquis	38,903	4
OPUS-Tatoeba	35,037	–
es_LA.OPUS-OpenSubtitles2018	34,928	12
OPUS-UNPC	34,514	4
OPUS-TED2020	31,017	6
OPUS-NeuLab-TedTalks	25,279	3
OPUS-EMEA	20,498	–
OPUS-EuroPat	16,707	–
OPUS-OpenSubtitles	13,306	1
OPUS-KDE4	9,155	1
OPUS-News-Commentary	7,048	1
OPUS-ECB	6,475	–
OPUS-Mozilla-I10n	3,733	–
OPUS-Books	3,424	–
OPUS-OpenOffice	1,435	–
OPUS-MDN_Web_Docs	620	–
OPUS-Ubuntu	539	–
OPUS-TED2013	412	–
OPUS-tico-19	406	–
OPUS-ECDC	292	–
OPUS-UN	290	–
OPUS-tldr-pages	289	–
OPUS-EUconst	234	–
OPUS-GNOME	140	–
Total	1,000,000	68

Table 5: Number of Spanish sentence pairs contributed by each corpus for the Stage 1 SFT training dataset. Non-OPUS sources are aggregated as proprietary data.

A.4 Prompt for MT Training & Evaluation

System: You are a translation assistant who always responds with an {target language} translation. You should write the translation only without any formatting.

User: Translate the given sentence into {target language}.
Sentence:
{source sentence}

A.5 Prompts for MT-IF Formality Control

A.5.1 Training & Evaluation

System: You are a translation assistant who always responds in an {target language} translation. You should only write the translation without any

Corpus	Base	CONCAT
OPUS-SciELO	48,193	–
OPUS-GlobalVoices	47,387	18,945
OPUS-ELITR-ECA	40,974	63,626
OPUS-JRC-Acquis	38,903	25,130
OPUS-Tatoeba	35,037	–
OPUS-ELRC	33,584	–
OPUS-DGT	32,866	38,634
OPUS-TED2020	31,017	58,439
OPUS-Europarl	29,490	31,800
OPUS-EUbookshop	29,230	86,181
OPUS-NeuLab-TedTalks	25,279	46,687
OPUS-EMEA	20,498	–
OPUS-MultiUN	15,022	16,802
es_LA.OPUS-OpenSubtitles2018	13,318	51,304
OPUS-UNPC	13,160	29,092
OPUS-KDE4	9,155	9,325
OPUS-News-Commentary	7,048	15,760
OPUS-ECB	6,475	–
OPUS-EuroPat	6,421	–
OPUS-OpenSubtitles	5,129	6,615
OPUS-Mozilla-I10n	3,733	–
OPUS-Books	3,424	–
OPUS-OpenOffice	1,435	–
OPUS-MDN_Web_Docs	620	1,020
OPUS-Ubuntu	539	503
OPUS-TED2013	412	–
OPUS-tico-19	406	–
OPUS-ECDC	292	–
OPUS-UN	290	–
OPUS-tldr-pages	289	–
OPUS-EUconst	234	–
OPUS-GNOME	140	137
Total	500,000	500,000

Table 6: Number of Spanish sentence pairs contributed by each OPUS corpus for the Stage 2 SFT training dataset.

formatting and without any reasoning.

User: Translate the following sentence into {target language} while maintaining [a|an] [formal|informal] tone.
Sentence:
{source sentence}

A.5.2 Different Training Prompts

System: Provide a {target language} translation only. Do not include explanations, formatting, or any text besides the translation itself.

User: Translate the following text into {target language}. Use the [informal|formal] form of address.

Corpus	Base	CONCAT
OPUS-MultiUN	157,691	144,588
OPUS-UNPC	124,691	202,562
OPUS-TED2020	61,929	44,055
OPUS-OpenSubtitles	57,736	64,417
OPUS-NeuLab-TedTalks	56,368	40,398
OPUS-GlobalVoices	13,369	2,396
OPUS-Tatoeba	9,211	–
OPUS-News-Commentary	4,659	463
OPUS-TED2013	3,906	–
OPUS-ELRC	3,846	–
OPUS-KDE4	2,624	744
OPUS-Mozilla-I10n	2,114	–
OPUS-tico-19	946	–
OPUS-Ubuntu	711	371
OPUS-UN	124	–
OPUS-EUbookshop	75	6
Total	500,000	500,000

Table 7: Number of Arabic sentence pairs contributed by each OPUS corpus for the Stage 2 SFT training dataset.

Category	# Prompts
General	116,872
Knowledge Recall	104,982
Math Reasoning	334,252
Coding	142,275
Safety & Non-Compliance	110,983
Multilingual	100,000
Instruction Following (removed)	29,980

Table 8: Composition of the Tülu 3 SFT-Mixture by category. We remove the synthetic instruction-following category to avoid overlap with the instructions evaluated in IFEval.

Text to translate:
{source sentence}

A.6 Prompts for MT-IF Grammatical Gender Control

The prompts use the male/female labels of the MT-GenEval annotation and are reproduced as used in our experiments.

A.6.1 Training & Evaluation

System: You are a translation assistant who always responds with a {target language} translation. You should write the translation only without any formatting.

User: Translate the given sentence into {target language}. In your translation the person referred to should be [male|female]. Translate all

Training Technique	# GPUs	Avg. Training Time (h)
Standard SFT	4x A100	16.89
CC	6x A100	19.26
EAFT	6x A100	17.25
Freeze	4x A100	14.15
KL-Full	4x H200	6.31
LoRA	4x A100	22.77
SPD	6x A100	18.04
STM	6x A100	21.23
Data Mixing (Tülu full)	4x A100	53.50
Data Mixing (MT-IF 1600)	4x A100	15.11
Data Mixing (Tülu + MT-IF; 2400)	4x A100	15.54
Data Mixing (Single MT-IF + Tülu; 1600)	4x A100	16.84
EWC (10k Tülu)	2x H200	14.79
EWC (MT-IF 1600)	2x H200	14.69
EWC (Tülu + MT-IF; 2400)	2x H200	14.99

Table 9: GPUs used and average training time (in hours) for each technique on the reported bidirectional Llama 3.1 8B Instruct training runs.

gendered words in the sentence accordingly, even if it uses opposite-gender pronouns.

Sentence:
{source sentence}

A.6.2 Different Training Prompts

System: Provide a {target language} translation only. Do not include explanations, formatting, or any text besides the translation itself.

User: Translate the following text into {target language}. Ensure the translation represents the subject as [male|female]. Apply [male|female] gender forms to all applicable words, regardless of the original pronouns.

Text to translate:
{source sentence}

A.7 Prompts for MT-IF Length Control

A.7.1 Baseline

System: You are a translation assistant who always responds in an {target language} translation. You should only write the translation without any formatting and without any reasoning.

User: Translate the following sentence into {target language}.

Sentence:

{source sentence}

A.7.2 Shorter / Longer

System: You are a translation assistant who always responds in an {target language} translation. You should only write the translation without any formatting and without any reasoning.

User v1: Translate the following sentence into {target language} while ensuring it has [more|fewer] characters than the original sentence.

Sentence:

{source sentence}

User v2: Translate the given sentence into {target language}. Additional instruction to follow while translating: Produce a translated sentence with [more|fewer] than {src chars} characters.

Sentence:

{source sentence}

User v3: Translate the following sentence into {target language} ensuring it has [more|fewer] than {src chars} characters.

Sentence:

{source sentence}

A.8 Stage 1: Best-Performing Configurations of Spanish & Arabic Training

Anchoring	Method	Hyperparameters	ar→en		de→en		General Avg
			COMET	BLEU	COMET	BLEU	
Llama 3.2 1B Instruct			75.8	18.1	82.9	27.6	44.9
Standard SFT			84.3	31.9	85.9	34.9	36.8
Auxiliary Data	Data Mixing	Tülu (500k)	83.8	31.2	86.2	35.3	40.8
	Merge	Trans=0.5, Tülu=0.5	83.1	29.7	86.4	36.0	42.8
	EWC	λ=1e3, Tülu (10k)	83.3	30.1	86.1	35.0	44.0
	EWC-DR	λ=1e4, Tülu (10k)	80.9	24.8	86.1	33.8	44.1
Model Outputs	KL-Full	λ=0.03	83.7	30.2	85.8	33.7	38.6
	STM	X=0.25	81.9	28.2	85.8	34.6	41.3
	CC	X=0.2	84.3	31.8	85.9	34.2	37.9
	EAFT	K=20, Weight=1.5	<u>84.2</u>	31.7	85.9	35.1	37.9
Base Model Parameters	Freeze	N=2	83.8	31.8	86.0	35.2	37.5
	SPD	λ=5	84.0	31.7	<u>86.2</u>	35.2	39.4
	LoRA	Rank=128, Alpha=64, lr=1e-5	83.2	29.7	85.4	34.0	36.5

Table 10: Best-performing configuration per method for Arabic, selected by the highest sum of Arabic COMET and general average score. Cell shading encodes the absolute difference from the base instruct model, normalized relative to the largest observed change within each metric group (across languages).

Anchoring	Method	Hyperparameters	es→en		de→en		General Avg
			COMET	BLEU	COMET	BLEU	
Llama 3.2 1B Instruct			81.3	20.3	82.9	27.6	44.9
Standard SFT			86.2	<u>28.4</u>	84.3	32.3	37.9
Auxiliary Data	Data Mixing	Tülu (500k)	<u>86.1</u>	28.1	85.6	33.8	40.7
	Merge	Trans=0.5, Tülu=0.5	85.7	27.2	86.5	35.8	42.2
	EWC	$\lambda=1e3$, Tülu (10k)	85.9	27.5	85.4	34.0	<u>44.2</u>
	EWC-DR	$\lambda=1e4$, Tülu (10k)	85.0	25.6	<u>85.9</u>	34.3	44.7
Model Outputs	KL-Full	$\lambda=0.003$	<u>86.1</u>	28.2	84.6	32.5	38.3
	STM	X=0.25	84.6	24.7	85.6	34.2	42.0
	CC	X=0.15	86.2	28.3	84.2	32.4	38.5
	EAFT	K=20, Weight=1.0	86.2	28.2	84.2	31.9	38.8
Base Model Parameters	Freeze	N=1	<u>86.1</u>	28.2	85.7	<u>34.5</u>	37.8
	SPD	$\lambda=5$	<u>86.1</u>	28.0	85.4	34.1	40.6
	LoRA	Rank=64, Alpha=64, lr=1e-5	85.9	27.8	82.6	30.1	38.0

Table 11: Best-performing configuration per method for Spanish, selected by the highest sum of Spanish-language COMET and general average score. Cell shading encodes the absolute difference from the base instruct model, normalized relative to the largest observed change within each metric group (across languages).

Anchoring	Method	Hyperparameters	AGI-Eval	HumanEval	HumanEval+	DROP	GPQA	GSM8K	IFEval	TruthfulQA	General Avg
Llama 3.2 1B Instruct			38.0	65.9	59.3	32.2	25.9	44.7	54.3	39.0	44.9
Standard SFT			32.6	54.5	49.4	30.0	26.8	33.9	39.9	41.9	38.6
Auxiliary Data	Data Mixing	Tülu (full)	34.4	56.7	53.8	33.0	23.7	42.8	40.3	41.0	40.7
	Merge	Trans=0.75, Tülu=0.25	34.2	59.1	55.7	30.9	24.6	40.7	47.9	41.6	41.8
	EWC	$\lambda=1e2$, Tülu (75k)	33.7	61.0	55.1	30.1	28.3	37.4	49.4	44.3	42.4
	EWC-DR	$\lambda=1e3$, Tülu (10k)	32.6	54.0	49.2	29.8	27.0	35.0	39.0	42.1	38.6
Model Outputs	KL	$\lambda=0.003$	32.5	54.1	48.7	29.6	27.0	34.3	36.8	41.9	38.1
	KL-Distill	$\lambda=0.05, \tau = 5$	32.8	59.6	52.1	29.9	24.1	33.3	39.9	<u>44.2</u>	39.5
	KL-Full	$\lambda=0.0025$	33.5	57.4	51.9	29.4	27.2	33.7	42.0	42.2	39.7
	STM	$X=0.15$	31.9	56.1	53.1	30.0	<u>28.1</u>	33.2	37.7	42.5	39.1
	CC	$X=0.45$	32.5	57.3	53.5	30.0	26.8	35.2	42.5	41.3	39.9
	EAFT	$K=20$, Weight=1.0	32.3	55.2	50.3	30.2	<u>27.5</u>	35.1	39.2	41.7	38.9
Base Model Parameters	Freeze	$N=1$	33.4	58.0	50.8	28.9	27.7	33.9	43.3	42.0	39.7
	SPD	$\lambda=2$	33.1	58.2	54.6	30.5	26.6	35.6	42.1	43.1	40.5
	LoRA	Rank=64, lr=1e-5	32.4	59.8	53.1	27.3	28.3	31.6	39.6	44.7	39.6

Table 12: Per-task scores on the general benchmarks for the best-performing configuration of each method for Amharic (same models as in Table 1).

A.9 Per-Task General Benchmark Scores for All Training Runs

Anchoring	Method	Hyperparameters	AGI-Eval	HumanEval	HumanEval+	DROP	GPQA	GSM8K	IFEval	TruthfulQA	General Avg
Llama 3.2 1B Instruct			38.0	65.9	59.3	32.2	25.9	44.7	54.3	39.0	44.9
Standard SFT			30.5	53.1	48.4	28.3	22.3	30.9	39.0	41.6	36.8
Auxiliary Data	Data Mixing	Tülu (500k)	35.4	59.3	51.3	33.1	23.2	42.7	41.4	40.2	40.8
	Merge	Trans=0.5, Tülu=0.5	37.0	62.5	56.3	32.2	24.6	41.7	46.6	41.6	42.8
	EWC	$\lambda=1e3$, Tülu (10k)	36.5	64.1	58.9	31.6	21.9	39.4	55.1	44.2	44.0
	EWC-DR	$\lambda=1e4$, Tülu (10k)	37.7	62.0	58.2	32.4	22.1	41.8	52.5	45.8	44.1
Model Outputs	KL-Full	$\lambda=0.03$	32.1	54.3	51.5	28.5	25.2	32.5	41.6	42.9	38.6
	STM	$X=0.25$	35.2	61.4	54.1	30.5	20.8	38.6	47.9	42.2	41.3
	CC	$X=0.2$	31.4	55.7	51.9	28.5	23.7	31.3	39.4	41.6	37.9
	EAFT	$K=20$, Weight=1.5	31.7	54.4	50.5	29.0	23.0	32.8	39.9	41.9	37.9
Base Model Parameters	Freeze	N=2	31.3	54.3	50.3	27.0	22.1	34.3	38.4	42.5	37.5
	SPD	$\lambda=5$	31.4	59.4	53.7	29.5	22.3	34.7	40.7	43.8	39.4
	LoRA	Rank=128, Alpha=64, lr=1e-5	25.5	56.8	52.9	26.8	19.4	31.4	36.2	42.7	36.5

Table 13: Per-task scores on the general benchmarks for the best-performing configuration of each method for Arabic (same models as in Table 10).

Anchoring	Method	Hyperparameters	AGI-Eval	HumanEval	HumanEval+	DROP	GPQA	GSM8K	IFEval	TruthfulQA	General Avg
Llama 3.2 1B Instruct			38.0	65.9	59.3	32.2	25.9	44.7	54.3	39.0	44.9
Standard SFT			32.1	56.4	48.6	29.4	24.1	31.7	37.5	43.2	37.9
Auxiliary Data	Merge	Tülu (500k)	34.6	57.4	50.8	33.3	25.4	42.1	41.6	40.1	40.7
		Trans=0.5, Tülu=0.5	36.3	59.7	54.0	32.2	24.6	41.2	47.7	42.2	42.2
		$\lambda=1e3$, Tülu (10k)	38.0	61.9	58.6	31.4	25.0	40.8	53.8	44.4	44.2
		$\lambda=1e4$, Tülu (10k)	38.3	63.5	58.3	32.5	26.1	41.7	52.9	44.5	44.7
Model Outputs	KL-Full	$\lambda=0.003$	32.5	54.6	50.6	29.6	25.7	32.0	39.0	42.6	38.3
	STM	$X=0.25$	37.1	62.4	55.1	31.7	21.0	39.9	49.0	39.6	42.0
	CC	$X=0.15$	31.9	58.4	50.8	29.7	24.1	32.8	37.3	43.3	38.5
	EAFT	$K=20$, Weight=1.0	32.5	55.8	53.7	29.8	24.6	32.1	39.2	42.8	38.8
Base Model Parameters	Freeze	N=1	32.4	56.9	47.4	29.3	24.8	32.3	36.2	43.3	37.8
	SPD	$\lambda=5$	33.8	58.4	54.9	30.4	24.6	35.6	42.5	44.9	40.6
	LoRA	Rank=64, Alpha=64, lr=1e-5	30.7	56.9	48.1	28.1	24.1	31.2	37.7	46.8	38.0

Table 14: Per-task scores on the general benchmarks (as percentages) for the best-performing configuration of each method for Spanish (same model selection as in Table 11).

Anchoring	Method	Hyperparameters	AGI-Eval	HumanEval	HumanEval+	DROP	GPQA	GSM8K	IFEval	TruthfulQA	General Avg
Llama 3.1 8B Instruct			64.5	85.3	79.8	61.5	26.8	83.2	78.7	55.0	66.8
Standard SFT			52.5	82.1	76.4	36.4	25.9	75.3	53.0	44.7	55.8
Auxiliary Data	Data Mixing	Tülu (full)	57.9	87.1	78.9	45.7	27.7	76.6	60.3	46.2	60.0
		Tülu (800) + MT-IF (1600)	58.3	82.1	76.2	61.1	27.0	76.3	55.1	48.5	60.6
		+ diff MT-IF prompt	59.1	80.8	75.3	62.5	27.5	77.3	56.9	47.6	60.9
		MT-IF (1600)	52.7	83.5	77.0	34.7	28.3	74.3	56.4	45.5	56.5
		Tülu (800) + Gender (800)	59.5	81.6	77.2	62.0	28.3	77.3	60.3	47.9	61.7
		+ diff Gender prompt	58.7	83.7	76.1	63.1	27.0	78.0	61.6	47.6	62.0
		Tülu (800) + Formality (800)	59.2	83.8	74.6	62.3	27.7	77.9	59.9	47.2	61.6
		+ diff Formality prompt	59.1	83.8	76.8	63.0	27.5	78.0	61.2	46.7	62.0
	Merge	Trans=0.5, Tülu+MT-IF=0.5	58.1	86.9	79.7	47.1	26.3	79.5	68.2	46.2	61.5
	EWC	$\lambda=1e3$, Tülu (10k)	60.3	85.3	80.2	62.1	28.3	82.5	72.1	50.0	65.1
		$\lambda=1e3$, Tülu (800) + MT-IF (1600)	57.3	85.4	78.5	64.2	25.2	77.7	69.7	46.5	63.1
		$\lambda=1e3$, MT-IF (1600)	56.8	84.9	78.7	34.0	25.7	76.9	60.6	47.3	58.1
Model Outputs	KL-Full	$\lambda=0.05$	53.5	83.6	76.6	46.8	25.0	74.4	56.0	46.4	57.8
	STM	$X=0.25$	50.3	82.1	73.7	25.1	24.3	71.2	52.9	44.1	53.0
	CC	$X=0.25$	49.7	83.2	75.9	30.2	26.8	73.6	55.1	45.4	55.0
	EAFT	$K=20$, Weight=1.0	52.3	84.1	75.4	46.5	25.0	76.0	58.0	45.5	57.8
Base Model Parameters	Freeze	N=4	49.4	82.9	74.0	31.2	26.8	75.4	49.0	45.5	54.3
	SPD	$\lambda=5$	61.4	85.2	79.4	27.1	24.6	81.7	69.1	53.2	60.2
	LoRA	Rank=32, Alpha=64	48.3	81.4	78.5	65.8	24.6	78.5	65.2	47.9	61.3

Table 15: Per-task scores on the general benchmarks for bidirectional Spanish training on the Llama 3.1 8B Instruct model.

Anchoring	Method	Hyperparameters	AGI-Eval	HumanEval	HumanEval+	DROP	GPQA	GSM8K	IFEval	TruthfulQA	General Avg
Llama 3.1 8B Instruct			64.5	85.3	79.8	61.5	26.8	83.2	78.7	55.0	66.8
Standard SFT			53.4	85.2	77.4	48.7	25.0	75.4	56.9	45.6	58.5
Auxiliary Data	Data Mixing	Tülu (full)	57.0	84.0	80.1	46.0	29.9	77.5	61.7	46.7	60.4
		Tülu (800) + MT-IF (1600)	58.9	84.7	77.9	65.6	25.4	76.4	59.3	46.8	61.9
		+ diff MT-IF prompt	58.8	83.2	76.1	65.6	28.1	76.7	58.0	47.0	61.7
		MT-IF (1600)	54.0	84.9	78.5	45.3	30.6	76.7	61.4	46.7	59.8
		Tülu (800) + Gender (800)	58.2	83.8	76.2	65.1	26.3	78.2	59.5	48.0	61.9
		Tülu (800) + Formality (800)	58.0	83.0	77.6	65.2	28.3	78.1	61.4	48.3	62.5
	Merge	Trans=0.5, Tülu+MT-IF=0.5	56.6	88.4	78.8	52.8	28.1	78.9	68.8	47.2	62.5
	EWC	$\lambda=1e3$, Tülu (10k)	60.7	87.7	78.4	64.0	28.6	82.5	76.7	50.4	66.1
		$\lambda=1e3$, Tülu (800) + MT-IF (1600)	57.7	84.6	76.0	45.4	27.2	77.7	66.5	47.6	60.1
		$\lambda=1e3$, MT-IF (1600)	55.9	85.0	77.0	66.0	24.3	78.9	69.9	47.9	63.3
Model Outputs	KL-Full	$\lambda=0.05$	53.8	85.2	77.3	57.6	27.9	74.5	59.2	47.9	60.4
	STM	$X=0.25$	51.2	85.0	76.4	35.3	24.6	75.1	54.2	46.1	56.0
	CC	$X=0.25$	52.2	85.9	78.2	41.4	25.7	75.0	55.1	46.8	57.5
	EAFT	$K=20$, Weight=1.0	52.3	84.1	80.8	55.0	26.1	75.9	59.5	46.7	60.1
Base Model Parameters	Freeze	$N=4$	47.9	84.7	75.1	45.9	25.7	77.0	50.5	48.3	56.9
	SPD	$\lambda=5$	59.9	83.9	80.2	46.4	28.8	82.7	71.0	52.6	63.2
	LoRA	Rank=32, Alpha=64	51.4	84.9	80.1	67.4	24.1	77.8	64.3	45.7	62.0

Table 16: Per-task scores on the general benchmarks for bidirectional Arabic training on the Llama 3.1 8B Instruct model.

A.10 Stage 2: Detailed Formality, Grammatical Gender and Length Control Scores for Spanish Training Runs

Anchoring	Method	Hyperparameters	en→es (N=600)					en→de (N=600)				
			COMET	BLEU	Formal M-Acc.	Informal M-Acc.	Avg.	COMET	BLEU	Formal M-Acc.	Informal M-Acc.	Avg.
Llama 3.1 8B Instruct			84.9	38.8	79.1 (M=359)	94.0 (M=252)	86.6	83.3	35.3	96.1 (M=434)	91.8 (M=437)	94.0
Standard SFT			85.7	40.5	50.0 (M=302)	62.2 (M=286)	56.1	81.8	31.8	58.4 (M=416)	59.1 (M=406)	58.8
Auxiliary Data	Data Mixing	Tülu (full)	85.6	40.9	52.5 (M=301)	66.4 (M=283)	59.5	83.0	33.4	67.4 (M=426)	57.3 (M=419)	62.4
		Tülu (800) + MT-IF (1600)	86.5	47.7	99.1 (M=425)	91.0 (M=255)	95.1	83.1	36.6	98.5 (M=401)	87.7 (M=447)	93.1
		+ diff MT-IF prompt	86.3	44.8	92.8 (M=391)	89.6 (M=259)	91.2	84.1	41.5	99.0 (M=480)	94.6 (M=482)	96.8
		Tülu (800) + Gender (800)	85.6	40.5	46.0 (M=287)	61.3 (M=284)	53.7	82.3	31.6	55.1 (M=423)	60.0 (M=425)	55.6
		+ diff Gender prompt	85.9	42.0	46.2 (M=292)	67.0 (M=276)	56.6	81.7	31.7	52.9 (M=435)	64.5 (M=431)	58.7
		Tülu (800) + Formality (800)	86.4	47.2	98.6 (M=427)	92.3 (M=248)	95.5	82.5	35.9	98.7 (M=379)	88.0 (M=458)	93.4
		+ diff Formality prompt	85.5	40.3	45.9 (M=290)	61.8 (M=280)	53.8	81.5	31.0	54.9 (M=430)	60.5 (M=420)	57.7
	Merge	Trans=0.5, Tülu+MT-IF=0.5	85.7	41.6	47.2 (M=299)	74.7 (M=273)	61.0	82.9	34.9	67.8 (M=451)	61.9 (M=449)	64.9
	EWC	$\lambda=1e3$, Tülu (10k)	85.7	40.8	52.3 (M=304)	58.6 (M=290)	55.5	82.3	32.9	65.3 (M=429)	55.6 (M=430)	60.5
		$\lambda=1e3$, Tülu (800) + MT-IF (1600)	85.8	40.9	52.1 (M=303)	64.2 (M=285)	58.2	82.4	32.8	69.2 (M=419)	54.7 (M=422)	61.2
		$\lambda=1e3$, MT-IF (1600)	85.7	40.9	53.6 (M=308)	62.6 (M=286)	58.1	82.4	32.8	70.0 (M=427)	53.4 (M=423)	61.7
Model Outputs	KL-Full	$\lambda=0.05$	85.7	41.0	47.5 (M=295)	64.4 (M=281)	56.0	82.1	32.0	54.6 (M=425)	63.6 (M=418)	59.1
	STM	$X=0.25$	84.9	36.7	54.4 (M=298)	58.0 (M=283)	56.2	80.8	29.2	60.6 (M=376)	54.5 (M=374)	57.6
	CC	$X=0.25$	83.9	34.3	11.2 (M=259)	93.7 (M=255)	52.3	81.6	31.9	25.8 (M=414)	86.2 (M=413)	56.0
	EAFT	$K=20$, Weight=1.0	85.7	40.7	48.1 (M=291)	62.7 (M=279)	55.4	82.0	32.1	52.5 (M=434)	62.0 (M=434)	57.3
Base Model Parameters	Freeze	$N=4$	85.7	40.7	49.2 (M=299)	64.5 (M=282)	56.9	82.3	31.9	58.0 (M=419)	59.2 (M=412)	58.6
	SPD	$\lambda=5$	85.5	41.0	57.8 (M=325)	86.1 (M=267)	72.0	82.7	33.1	72.8 (M=419)	78.4 (M=426)	75.6
	LoRA	Rank=32, Alpha=64	85.5	40.2	51.3 (M=302)	66.9 (M=287)	59.1	81.9	32.3	71.9 (M=420)	54.2 (M=421)	63.1

Table 17: Formality Control scores (including number of matches used for accuracy scores) for bidirectional Spanish training on the Llama 3.1 8B Instruct model.

Anchoring	Method	Hyperparameters	en→es (N=300)					en→de (N=300)				
			COMET	BLEU	Fem. Acc.	Masc. Acc.	Avg.	COMET	BLEU	Fem. Acc.	Masc. Acc.	Avg.
Llama 3.1 8B Instruct			83.1	44.5	42.3	60.7	51.5	81.2	32.8	63.3	50.7	57.0
Standard SFT			83.0	43.1	18.3	17.3	17.8	81.4	29.2	8.7	9.7	9.2
Auxiliary Data	Data Mixing	Tülu (full)	83.0	43.6	20.0	18.3	19.2	82.5	31.9	8.0	4.0	6.0
		Tülu (800) + MT-IF (1600)	86.2	52.8	81.7	75.0	78.4	85.3	40.8	84.3	82.7	83.5
		+ diff MT-IF prompt	83.5	40.9	22.0	19.7	20.8	82.9	32.5	19.7	15.3	17.5
		MT-IF (1600)	86.2	52.5	85.3	78.7	82.0	85.4	41.4	85.7	84.3	85.0
		Tülu (800) + Gender (800)	86.2	52.9	80.0	75.7	77.9	85.4	40.5	83.0	84.0	83.5
		+ diff Gender prompt	82.8	43.7	18.0	17.0	17.5	81.5	29.6	6.3	7.7	7.0
		Tülu (800) + Formality (800)	82.6	43.6	18.3	17.7	18.0	81.3	29.6	8.3	6.7	7.5
		+ diff Formality prompt	83.3	44.6	20.3	19.0	19.7	81.7	29.9	9.0	15.7	12.3
	Merge	Trans=0.5, Tülu+MT-IF=0.5	82.9	32.4	22.7	24.0	23.4	82.9	32.4	22.7	24.0	23.4
	EWC	$\lambda=1e3$, Tülu (10k)	82.8	43.2	19.0	17.3	18.2	82.0	30.3	8.3	8.7	8.5
		$\lambda=1e3$, Tülu (800) + MT-IF (1600)	82.8	43.1	18.3	17.0	17.7	81.9	30.4	8.7	7.7	8.2
		$\lambda=1e3$, MT-IF (1600)	82.7	43.0	19.0	17.3	18.2	81.8	30.3	8.3	7.7	8.0
Model Outputs	KL-Full	$\lambda=0.05$	83.3	43.8	18.7	17.0	17.9	81.7	29.6	7.3	8.0	7.7
	STM	$X=0.25$	82.3	37.8	18.0	17.0	17.5	81.2	27.8	8.7	9.3	9.0
	CC	$X=0.25$	81.6	39.8	20.3	19.0	19.7	79.8	28.2	8.3	10.3	9.3
	EAFT	$K=20$, Weight=1.0	83.2	43.8	18.0	17.3	17.7	81.8	29.4	8.0	8.7	8.4
Base Model Parameters	Freeze	$N=4$	83.2	42.9	17.7	17.7	17.7	82.1	29.3	7.0	8.7	7.9
	SPD	$\lambda=5$	84.2	46.0	47.0	30.0	38.5	82.9	32.2	39.3	38.3	38.8
	LoRA	Rank=32, Alpha=64	82.8	42.6	17.0	15.7	16.4	82.0	30.2	6.7	8.3	7.5

Table 18: Grammatical Gender Control scores for bidirectional Spanish training on the Llama 3.1 8B Instruct model. Fem. Acc. and Masc. Acc. are the accuracies of producing the requested feminine and masculine forms.

			en↔es							
			Baseline		Shorter			Longer		
Anchoring	Method	Hyperparameters	Avg. COMET	Tgt/Src	Avg. COMET	% Shorter	Tgt/Src	Avg. COMET	% Longer	Tgt/Src
Llama 3.1 8B Instruct			86.0	1.040	86.0	80.35	0.941	84.0	70.20	1.349
Standard SFT			87.0	0.998	86.6	21.40	0.994	83.7	24.45	1.010
Auxiliary Data	Data Mixing	Tülu (full)	86.9	0.997	86.8	26.90	0.990	86.0	22.95	1.004
		Tülu (800) + MT-IF (1600)	87.0	0.997	86.9	27.25	0.987	86.0	27.70	1.006
		+ diff MT-IF prompt	87.0	0.997	86.9	22.58	0.981	85.8	22.97	0.999
		MT-IF (1600)	87.0	0.998	86.7	23.30	0.991	84.1	24.90	0.999
		Tülu (800) + Gender (800)	87.0	0.998	86.9	21.10	0.992	86.9	23.00	1.001
		+ diff Gender prompt	87.0	0.998	86.9	19.55	0.995	86.3	23.45	1.000
		Tülu (800) + Formality (800)	87.0	0.998	86.9	21.65	0.994	86.4	24.50	1.002
		+ diff Formality prompt	87.0	0.997	86.9	20.70	0.992	86.6	22.80	1.001
	Merge	Trans=0.5, Tülu+MT-IF=0.5	87.0	0.996	86.9	27.25	0.988	85.7	24.65	1.001
	EWC	λ=1e3, Tülu (10k)	87.0	1.000	84.2	27.10	1.027	74.3	37.20	1.103
λ=1e3, Tülu (800) + MT-IF (1600)		87.0	0.999	85.8	27.10	1.006	77.4	31.30	1.040	
λ=1e3, MT-IF (1600)		87.0	0.998	86.0	26.60	0.992	79.0	29.35	1.028	
Model Outputs	KL-Full	λ=0.05	86.9	1.002	86.8	20.80	1.000	85.5	24.00	1.008
	STM	X=0.25	86.6	0.993	86.2	24.00	0.989	83.5	23.60	0.998
	CC	X=0.25	87.0	0.997	86.3	20.45	0.992	83.1	24.50	0.996
	EAFT	K=20, Weight=1.0	87.0	0.998	86.6	19.75	0.998	83.2	25.30	1.014
Base Model Parameters	Freeze	N=4	87.0	0.997	86.9	23.25	0.990	86.3	22.85	1.002
	SPD	λ=5	86.9	0.998	86.7	58.35	0.951	86.7	40.60	1.015
	LoRA	Rank=32, Alpha=64	86.9	0.997	86.8	21.15	1.000	85.1	26.90	1.044

Table 19: Effect of length instruction (none/shorter/longer) on length control scores for en↔es translation with Llama 3.1 8B Instruct, evaluated on FLORES (bidirectional, 2x N=1012).

A.11 Stage 2: Arabic Scores

Anchoring	Method	Hyperparameters	ar↔en		de↔en		General Avg	ar
			COMET	BLEU	COMET	BLEU		MT-IF Avg
Llama 3.1 8B Instruct			84.4	26.4	87.6	37.3	66.8	66.8
Standard SFT			86.7	32.9	87.2	37.2	58.5	28.9
Auxiliary Data	Data Mixing	Tülu (full)	86.7	32.5	87.7	38.0	60.4	26.9
		Tülu (800) + MT-IF (1600)	86.8	32.7	87.8	38.4	61.9	51.8
		+ diff MT-IF prompt	86.8	32.6	87.8	38.2	61.7	27.3
		MT-IF (1600)	86.7	32.8	87.4	37.6	59.8	55.1
		Tülu (800) + Gender (800)	86.7	32.6	87.4	37.6	61.9	53.0
		Tülu (800) + Formality (800)	86.7	32.7	87.4	37.8	62.5	24.4
	Merge	Trans=0.5, Tülu+MT-IF=0.5	86.4	31.8	87.9	38.5	62.5	33.2
	EWC	$\lambda=1e3$, Tülu (10k)	86.6	32.3	87.4	37.7	66.1	32.7
		$\lambda=1e3$, Tülu (800) + MT-IF (1600)	86.6	32.7	87.3	37.4	63.3	28.2
		$\lambda=1e3$, MT-IF (1600)	86.7	32.7	87.5	37.4	60.1	29.3
Model Outputs	KL-Full	$\lambda=0.05$	86.8	33.1	87.4	37.7	60.4	26.1
	STM	$X=0.25$	86.2	31.0	87.1	36.6	56.0	30.0
	CC	$X=0.25$	86.8	33.0	87.0	36.9	57.5	25.2
	EAFT	$K=20$, Weight=1.0	86.8	32.9	87.3	37.4	60.1	24.9
Base Model Parameters	Freeze	$N=4$	86.6	32.6	87.3	37.1	56.9	25.5
	SPD	$\lambda=5$	86.6	32.0	87.5	37.5	63.2	39.7
	LoRA	Rank=32, Alpha=64	86.2	31.1	87.4	37.2	62.0	27.3

Table 20: MT, General Avg, and Avg MT-IF scores for bidirectional Arabic training on the Llama 3.1 8B Instruct model. Color shade of cells encodes the absolute difference to the base model, normalized within each metric group (across languages).

			ar↔en							
			Baseline		Shorter			Longer		
Anchoring	Method	Hyperparameters	Avg. COMET	Tgt/Src	Avg. COMET	% Shorter	Tgt/Src	Avg. COMET	% Longer	Tgt/Src
Llama 3.1 8B Instruct			84.4	1.069	84.0	78.4	0.942	83.2	66.8	1.329
Standard SFT			86.7	1.006	84.5	34.3	1.003	78.3	33.2	1.061
Auxiliary Data	Data Mixing	Tülu (full)	86.6	0.998	86.3	37.7	0.978	82.7	30.0	0.997
		Tülu (800) + MT-IF (1600)	86.8	1.005	86.3	36.3	0.982	80.0	32.1	1.021
		+ diff MT-IF prompt	86.8	1.004	85.1	37.5	0.981	77.5	31.2	1.018
		MT-IF (1600)	86.7	1.005	86.4	36.5	0.991	82.4	36.5	1.060
		Tülu (800) + Gender (800)	86.6	1.005	86.4	34.3	0.996	82.5	30.1	1.012
		Tülu (800) + Formality (800)	86.7	1.004	86.2	32.4	1.002	81.1	30.6	1.012
	Merge	Trans=0.5, Tülu+MT-IF=0.5	86.3	1.005	86.2	35.2	0.987	84.5	29.0	1.002
	EWC	λ=1e3, Tülu (10k)	86.6	1.007	80.5	43.0	0.989	71.1	43.0	1.218
		λ=1e3, Tülu (800) + MT-IF (1600)	86.6	1.009	83.5	35.2	1.021	77.5	38.5	1.100
		λ=1e3, MT-IF (1600)	86.7	1.006	83.3	37.0	1.016	77.1	39.0	1.093
Model Outputs	KL-Full	λ=0.05	86.8	1.012	85.4	32.7	1.009	79.1	36.1	1.099
	STM	X=0.25	86.2	1.003	84.1	35.5	0.982	76.1	27.6	0.999
	CC	X=0.25	86.8	1.011	85.4	33.1	1.006	80.0	31.1	1.057
	EAFT	K=20, Weight=1.0	86.7	1.008	85.7	32.4	0.993	79.5	28.1	1.006
Base Model Parameters	Freeze	N=4	86.6	1.004	86.5	32.9	0.990	80.1	32.3	1.045
	SPD	λ=5	86.5	1.009	83.2	49.8	0.988	75.2	49.4	1.183
	LoRA	Rank=32, Alpha=64	86.1	1.011	81.2	41.1	0.977	73.9	28.4	0.999

Table 21: Effect of length instructions (none/shorter/longer) on length control scores for Arabic ↔ English translation with Llama 3.1 8B Instruct, evaluated on FLORES (bidirectional, 2x N=1012).

Anchoring	Method	Hyperparameters	en→ar (N=300)					en→de (N=300)				
			COMET	BLEU	Fem. Acc.	Masc. Acc.	Avg.	COMET	BLEU	Fem. Acc.	Masc. Acc.	Avg.
Llama 3.1 8B Instruct			76.6	17.3	70.3	51.7	61.0	81.2	32.8	63.3	50.7	57.0
Standard SFT			76.2	16.0	30.0	18.0	24.0	80.3	28.6	8.3	8.7	8.5
Auxiliary Data	Data Mixing	Tülu (full)	78.9	17.6	23.0	17.0	20.0	82.2	31.7	10.0	12.3	11.2
		Tülu (800) + MT-IF (1600)	80.9	21.2	73.7	65.0	69.4	85.2	41.2	79.7	80.0	79.9
		+ diff MT-IF prompt	79.5	17.9	23.0	17.3	20.2	83.3	35.2	36.7	38.0	37.3
		MT-IF (1600)	80.8	22.5	73.3	74.0	73.7	83.8	39.1	80.7	83.0	81.8
		Tülu (800) + Gender (800)	80.9	22.7	72.7	75.0	73.8	83.9	39.1	81.3	82.7	82.0
		Tülu (800) + Formality (800)	79.1	17.7	19.0	15.7	17.3	80.7	29.9	7.3	9.7	8.5
	Merge	Trans=0.5, Tülu+MT-IF=0.5	79.7	18.4	42.3	26.3	34.3	83.0	34.9	36.0	35.3	35.7
	EWC	$\lambda=1e3$, Tülu (10k)	76.8	16.0	23.3	21.7	22.5	81.1	29.9	13.7	11.3	12.5
		$\lambda=1e3$, Tülu (800) + MT-IF (1600)	78.2	17.1	20.3	18.7	19.5	81.3	29.8	11.0	9.0	10.0
		$\lambda=1e3$, MT-IF (1600)	78.1	17.3	21.7	19.3	20.5	81.1	29.6	11.0	8.7	9.9
Model Outputs	KL Full	$\lambda=0.05$	78.7	17.2	19.3	16.3	17.8	80.9	29.5	7.3	8.3	7.8
	STM	$X=0.25$	74.9	14.0	28.7	28.3	28.5	80.5	28.0	8.0	8.0	8.0
	CC	$X=0.25$	78.3	16.7	20.0	16.7	18.3	79.8	28.2	8.3	10.3	9.3
	EAFT	K=20, Weight=1.0	78.4	17.0	19.7	19.3	19.5	80.8	29.6	6.7	7.3	7.0
Base Model Parameters	Freeze	N=4	79.2	18.1	20.0	17.0	18.5	80.5	28.5	10.0	10.0	10.0
	SPD	$\lambda=5$	79.6	18.3	35.0	24.3	29.7	82.0	32.1	31.3	24.7	28.0
	LoRA	Rank=32, Alpha=64	78.4	16.7	21.7	18.0	19.8	80.8	28.9	6.0	8.0	7.0

Table 22: Grammatical Gender Control scores for bidirectional Arabic training on the Llama 3.1 8B Instruct model. Fem. Acc. and Masc. Acc. are the accuracies of producing the requested feminine and masculine forms.