Attribute-Controlled Translation with Preference Optimization

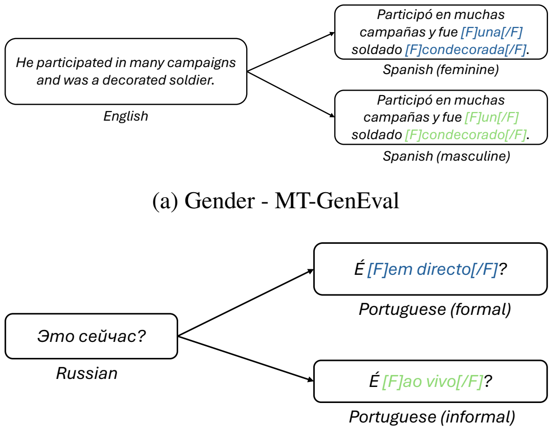
Inigo Jauregi Unanue[†], **Najmeh Sadoughi**[▽], **Vimal Bhat**[▽], **Zhu Liu**[▽], **Massimo Piccardi**[†]
[†]University of Technology Sydney, Australia
[▽]Amazon Prime Video, United States

Abstract

Attribute-controlled translation (ACT) seeks to produce translations that satisfy specific constraints on linguistic and stylistic attributes. While careful prompt engineering can enable large language models to perform strongly in this task, its effectiveness is mainly limited to models of very large size. For this reason, in this paper we set to improve the performance of language models of more contained size by leveraging the contrastive nature of ACT tasks with preference optimization, as well as exploiting knowledge distillation with synthetically-generated training samples from larger models. As a resource for this investigation, we also introduce PREF-FAME-MT, a large, contrastive, formality-controlled parallel corpus which has been generated by expanding the existing FAME-MT dataset with synthetic contrastive samples. Experiments conducted over three datasets for formality- and gender-controlled translation with 71 distinct language pairs have demonstrated the effectiveness of the proposed approach at simultaneously improving attribute matching and translation quality. We release all our code and datasets to allow reproduction and expansion of our work[1].

1 Introduction

Attribute-controlled translation (ACT) is a natural language processing task that targets two complementary objectives: 1) providing a high-quality translation from the source to the target language, and 2) generating text that conforms to predetermined linguistic and stylistic attributes such as formality, style, gender-specificity, and length. Control over these types of attributes is important as it enables translations that suit different contexts and applications. In other words, ACT can be framed as the integration of conventional machine translation and text style transfer (Sarti et al., 2023).


He participated in many campaigns and was a decorated soldier.
English
Participó en muchas campañas y fue [F]una[/F] soldado [F]condecorada[/F].
Spanish (feminine)
Participó en muchas campañas y fue [F]un[/F] soldado [F]condecorado[/F].
Spanish (masculine)
Это сейчас?
Russian
É [F]em directo[/F]?
Portuguese (formal)
É [F]ao vivo[/F]?
Portuguese (informal)

(a) Gender - MT-GenEval

(b) Formality - PREF-FAME-MT

Figure 1: Contrastive examples from ACT datasets. Lexical differences are annotated with [F] and [/F] tags.

Recent advances in ACT have been mainly driven by improvements in large language models (LLMs) and the use of prompt engineering (Lee et al., 2024b), in-context learning (Sarti et al., 2023), and fine-tuning (Schioppa et al., 2021). These advances have been complemented by the creation of dedicated benchmarks and evaluation metrics tailored to various ACT tasks (Costa-jussà et al., 2020; Nadejde et al., 2022; Currey et al., 2022; Rarrick et al., 2023; Wisniewski et al., 2024). Among them, contrastive datasets highlight the key lexical differences of translations with opposing attributes (Figure 1), and as such, have become valuable resources for both model training and contrastive evaluation.

However, the highest levels of performance in ACT seem to increasingly be restricted to language models of very large size (Robinson et al., 2023; Anthropic, 2024). While this issue may be mollified by aggressively fine-tuning smaller models, the scarcity of ACT training data presents a critical bottleneck, as most datasets only contain a few hundred annotated examples (Liu and Niehues, 2024). Therefore, more effective training approaches are

[1] GitHub link: `https://github.com/inigo-jauregi/act-po`

Findings of the Association for Computational Linguistics: EACL 2026, pages 4031–4057
March 24-29, 2026 ©2026 Association for Computational Linguistics

needed to improve the performance of smaller language models with such limited training data.

For this reason, in this paper we propose a novel fine-tuning approach based on contrastive annotation and preference optimization, jointly with sampling strategies to supplement the manually-annotated data with synthetic contrastive samples. Drawing inspiration from the success of reinforcement learning from human feedback (RLHF) in aligning LLMs with human preferences (Christiano et al., 2017; Ouyang et al., 2022), we remark that existing ACT datasets—that typically contain contrastive target pairs with opposite attributes—spontaneously lend themselves to preference-based learning (Schulman et al., 2017; Rafailov et al., 2023). For example, in the case of formality-controlled translation, an available formal reference naturally becomes the preferred candidate over an informal one. This approach can be systematically applied to any targeted attribute (e.g., formality, gender, verbosity), creating a rich preference signal that captures the bidirectional nature of style transfer tasks.

A distinct, additional challenge of ACT tasks is that they are inherently instances of an optimization of a dual nature, where the aims are to concurrently ensure compliance of the attribute and quality of the translation. We address this issue by crafting synthetic preference pairs that are carefully controlled along both these dimensions. Since the synthetic samples are generated by state-of-the-art LLMs, this approach can be seen as a form of knowledge distillation from larger to smaller language models.

Finally, we show that we can also leverage LLMs to convert existing translation resources into contrastive ACT datasets. The key idea is to prompt LLMs to perform style transfer over the reference sentences in the target language, thus generating synthetic contrastive examples that can be used for preference optimization. We apply this approach to an existing formality-controlled translation dataset named FAME-MT (Wisniewski et al., 2024) in 7 language pairs, creating a new, contrastive version that we aptly nickname PREF-FAME-MT, inclusive of a human-validated test set. The resulting dataset is more challenging and more lexically-diverse in formality compared to other popular formality-controlled datasets such as CoCoA-MT (Nadejde et al., 2022).

In summary, our paper makes the following contributions:

- **Preference optimization.** To the best of our knowledge, our paper is the first to propose the use of preference optimization for fine-tuning language models for ACT tasks.
- **Synthetic samples.** We propose an approach for generating synthetic preference pairs that contrast in both attribute accuracy and translation quality.
- **Contrastive dataset.** We present a new contrastive dataset for formality-controlled translation nicknamed PREF-FAME-MT which is simultaneously of large size and more lexically diverse in formality expressions.
- **Experimental results.** Experimental results in formality- and gender-controlled ACT tasks over three datasets and 71 language pairs demonstrate the effectiveness of the proposed fine-tuning approach.

2 Related Work

ACT Models. Controlling specific linguistic features in neural machine translation (NMT) models has emerged as an active research area in recent years. Early work has focused on providing control signals for various attributes such as formality (Sennrich et al., 2016; Niu et al., 2018), gender (Vanmassenhove et al., 2018; Saunders et al., 2020), length (Lakew et al., 2019; Takase and Okazaki, 2019) or style (Michel and Neubig, 2018; Wang et al., 2023). These approaches explore different solutions, but, in general, they all fine-tune NMT models using special input tokens or dedicated control vectors. More recently, researchers have proposed prompting multilingual LLMs to control translation outputs. For instance, Lee et al. (2024b) introduced an approach to include entity-level gender information in the prompt to guide the LLM to translate with correct gender inflections. At their turn, Sarti et al. (2023) proposed an in-context learning approach that uses retrieval-augmented generation (RAG) to retrieve relevant controlled translation examples from a document store. This line of work relies entirely on prompt engineering and dispenses with fine-tuning.

Preference Optimization. Aligning LLMs to human preferences has become a common approach to encourage models to generate high-quality, factual and safe responses (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). Training objectives such as proximal policy optimization (PPO) (Schulman et al., 2017) or direct preference optimization (DPO) (Rafailov et al., 2023) have been widely adopted in the research literature for this purpose. In machine translation specifically, several studies have shown that preference optimization can improve domain alignment and overall translation quality (Xu et al., 2024a; Uhlig et al., 2024; Vajda et al., 2025). However, no prior work that we are aware of has attempted to leverage preference optimization specifically for ACT.

Synthetic Data. Given the cost of collecting manually-annotated human preferences, several works have proposed approaches for generating synthetic preferences by sampling LLMs (Lee et al., 2024a; Cui et al., 2024). Among them, Geng et al. (2025) have shown that the performance of a given model can be improved by fine-tuning it with synthetic samples generated by lower-quality models, so long that there is a clear "learning delta" between the preferences. In machine translation, Xu et al. (2024a) have proposed using reference-free translation quality estimation models to turn the samples generated by multilingual LLMs into preference pairs. At their turn, Cui et al. (2025) have proposed a combination of quality scores and model confidence to select preferences for fine-tuning. In the same vein, our work proposes an approach for generating synthetic data for ACT tasks that enforces a learning delta in both translation quality and attribute compliance.

3 The Proposed Approach: Preference Optimization for ACT

In this section, we first describe the approach that we follow to train ACT models with preference optimization (Section 3.1), and then present our approach for generating synthetic contrastive samples (Section 3.2).

3.1 Preference Optimization

In ACT, a contrastive sample is a triplet (x, y_{a_1}, y_{a_2}) formed by a sentence, x, in the source language, and two translations, y_{a_1} and y_{a_2}, with opposing controllable attributes or styles i.e., formal/informal, masculine/feminine (see Figure 1). In the instruction-tuning (IT) framework, a prompt p_{a_t} acts as the control signal for the target attribute, $a_t \in \{a_1, a_2\}$. The prompt can be anything from a single token to an elaborate sentence. The model receives the prompt and the source sentence in input, and can be trained with a variety of training objectives. The most conventional is the negative log-likelihood (NLL) loss that maximizes the probability of the translation with the target attribute, y_{a_t}:

$$\begin{aligned}\mathcal{L}_{NLL}(\pi_\theta) = \\ -\mathbb{E}_{(x, y_{a_1}, y_{a_2}, a_t)\sim\mathcal{D}}[\log \pi_\theta(y_{a_t}|x, p_{a_t})]\end{aligned} \tag{1}$$

where π_θ is the probability assigned to the target translation by the model, and θ is its set of trainable parameters.

A popular, contemporary alternative is the preference optimization framework introduced by the seminal DPO paper of Rafailov et al. (2023), where both references are used in the training loss, taking the role of the preferred, y_{pref}, and the rejected, y_{rej}, candidate, respectively, depending on the target attribute value. In this work we adopt a computationally-lighter variant of DPO known as *contrastive preference optimization* (CPO) (Xu et al., 2024a), where the training objective is defined as:

$$\begin{aligned}\mathcal{L}_{CPO}(\pi_\theta) = -\mathbb{E}_{(x, y_{\text{pref}}, y_{\text{rej}}, a_t)\sim\mathcal{D}}\Bigg[\\ \lambda \log \sigma\Big(\beta \log \frac{\pi_\theta(y_{\text{pref}}|x, p_{a_t})}{\pi_\theta(y_{\text{rej}}|x, p_{a_t})}\Big) \\ + (1-\lambda) \log \pi_\theta(y_{\text{pref}}|x, p_{a_t})\Bigg],\end{aligned} \tag{2}$$

where the first term rewards the ratio between the probabilities of the preferred and rejected translations, while the second acts as an NLL regularizer; β and λ are their respective hyperparameters. The main advantage of CPO with respect to DPO is that it does not require storing an additional, reference model in memory, while performing on par in most reported cases (Xu et al., 2024b).

3.2 Synthetic Samples for Dual Optimization

However, as we show later in Section 5, the straightforward use of attribute-contrastive samples from ACT datasets for preference optimization tends to significantly compromise translation quality. The reason is that the contrastive reference translations in these datasets typically only differ in the tokens that convey the attribute, but do not differ in overall translation quality. As such, they fail to create a "learning delta" that can act as a control signal for translation quality, leading to typical artifacts such as reduced fluency and drops in evaluation metrics.

For this reason, in this paper we address this issue by generating synthetic preference pairs that align with the dual nature of the task, simultaneously contrasting in attribute compliance and translation quality. The process for generating the synthetic preference pairs is fully described in Algorithm 1. For each input sentence in the training

Algorithm 1 Synthetic Sample Generation

```
Input: $\Pi_{LLM}$, $\mathcal{D}_{train}$
Output: $\mathcal{D}_{syn}$

$\mathcal{D}_{syn} \leftarrow \emptyset$
for each $(x, y_{\text{pref}}\ y_{\text{rej}}, a_t)$ in $\mathcal{D}_{train}$ do
    $(y_s^1, y_s^2, \ldots, y_s^k) \leftarrow$ getSamples($\pi_{LLM}$, $x$, $p_{a_t}$)
    KiwiScore$_{\text{rej}}$ $\leftarrow$ getKiwiScore($x$, $y_{\text{rej}}$)
    $candidates \leftarrow []$
    for each $y_s$ in $(y_s^1, y_s^2, \ldots, y_s^k)$ do
        if matchingAttribute($y_s$, $y_{\text{pref}}$) then
            KiwiScore$_s$ $\leftarrow$ getKiwiScore($x$, $y_s$)
            if KiwiScore$_s$ $>$ KiwiScore$_{\text{rej}}$ then
                $diff \leftarrow$ KiwiScore$_s$ - KiwiScore$_{\text{rej}}$
                $candidates.append((y_s, diff))$
            end if
        end if
    end for
    if $candidates$ is not empty then
        $best\_sample \leftarrow \arg\max_{(y_s,d) \in candidates} d$
        $y^s_{\text{pref}} \leftarrow best\_sample[0]$
        $y^s_{\text{rej}} \leftarrow y_{\text{rej}}$
        $\mathcal{D}_{syn} \leftarrow \mathcal{D}_{syn} \cup (x, y^s_{\text{pref}}, y^s_{\text{rej}})$
    end if
end for
return $\mathcal{D}_{syn}$
```

dataset, x, we first draw k translation samples from a chosen LLM [2]. Then, the samples are evaluated with two criteria: 1) matching the preferred reference, y_{pref}, in attribute; and 2) scoring a translation score higher than that of the rejected reference, y_{rej}. As in (Xu et al., 2024a), we use CometKiwi-XXL (Kiwi-XXL for brevity hereafter) (Rei et al., 2023) as our reference-free model to assess the translation quality of the samples. The sample with the highest translation score is selected as the new preferred candidate, y^s_{pref}, and the original rejected candidate is retained as y^s_{rej}. By forming synthetic data in this way, we ensure that the samples have a positive learning delta in translation quality, with the preferred candidates scoring higher than the rejected references, while preserving the preferred attribute. We collect all such synthetic preference pairs in a new dataset, $\mathcal{D}_{syn}$.

4 Datasets

In this section, we describe the datasets used for the experiments, which include two popular datasets for formality- and gender-controlled ACT (CoCoA-MT (Nadejde et al., 2022) and MT-GenEval (Currey et al., 2022), respectively), and a new dataset for formality-controlled ACT, nicknamed PREF-FAME-MT, that we introduce as part this work.

CoCoA-MT (Nadejde et al., 2022) is a formality-controlled, contrastive translation dataset in the conversation domain. The dataset covers 8 language pairs, {*en*} → {*de, es, fr, hi, it, ja, nl, pt*}, and for each source sentence, both a *formal* and an *informal* translations are provided. Interestingly, a large number of the English source sentences are shared across language pairs, which allows their potential use as pivots to align other language pairs. By using them in this way, we have managed to align all the target languages pairwise, obtaining 7 × 8 = 56 new language pairs. These new combinations have allowed us to carry out a deeper analysis of the proposed approach in lower-resource language pairs.

MT-GenEval (Currey et al., 2022) is a gender-controlled, contrastive translation dataset in the Wikipedia domain. For each source sentence, both a *feminine* and a *masculine* translations are provided, with all pronouns in each sentence having the same gender. The dataset covers 8 language pairs {*en*} → {*ar, de, es, fr, hi, it, pt, ru*}.

PREF-FAME-MT is a new, formality-controlled ACT dataset that we have generated by converting the existing FAME-MT (Wisniewski et al., 2024) into a contrastive dataset. FAME-MT consists of 11.2 million translations between various European languages that have been automatically classified as formal, informal or neutral. While remarkable in size, this dataset has two limitations: 1) it does not provide contrastive pairs, and therefore is not suitable for preference optimization; and 2) it has not been human-validated, challenging its use for model evaluation. To address the first issue, we have generated synthetic contrastive counterparts for the original examples. In detail, we have prompted an external LLM[3] to carry out attribute transfer with minimal edits on the original translation in FAME-MT, i.e., *formal* → *informal* and *informal* → *formal* (see full prompt in Appendix B.1). Note that the attribute transfer has been performed in a monolingual setting, i.e. the LLM was not provided the source sentence. In addition, we have automatically annotated the token spans that differ in each contrastive pair with the longest common subsequence (LCS) algorithm (Hirschberg, 1975), obtaining contrastive examples of the same format as those of CoCoA-MT and MT-GenEval. We have applied this process to a subset of 7 language pairs, mixing Germanic, Latin and Slavic languages: *da*→*es*, *de*→*fr*, *en*→*de*, *es*→*en*, *it*→*nl*,

[2] LLM details in Appendix A.2

[3] LLM details in Appendix A.1

	PREF-FAME-MT	CoCoA-MT
TTR (↑)	**0.700**	0.217
Gini Coeff. (↓)	**0.300**	0.730
Concentration (↓)		
Top-5	**0.085**	0.157
Top-10	**0.118**	0.246
Top-20	**0.156**	0.361

Table 1: Comparison of the lexical diversity of the formality-annotated token spans in the test sets of PREF-FAME-MT and CoCoA-MT.

pl→it and *ru→pt*. Following common practice, we have divided this dataset into training, validation, and test sets.

To address the second issue, we have asked professional translators to validate 200 examples from the test sets of each target language. Annotators were asked to ensure that the contrastive sentences were grammatically correct, were paraphrases with identical meaning, and had opposite formality style. For the test sets, we have only retained the samples that were deemed correct by the annotators (more details in Appendix A.1).

PREF-FAME-MT has been designed to address several limitations of CoCoA-MT. In the first place, it has a much larger, contrastive training set ($\sim$ 140K samples per language pair vs $\sim$ 400; although it is a silver-corpus, while the CoCoA-MT training set has been manually validated). In the second place, PREF-FAME-MT has a much higher lexical diversity in formality expressions compared to CoCoA-MT, which predominantly expresses formality only through the use of personal pronouns (Wisniewski et al., 2024). To corroborate this claim, we have collected all the formality-annotated token spans (i.e., the tokens between [F] and [/F] tags) in the test sets of the target languages shared by both datasets, and computed the type-token ratio (TTR), the Gini coefficient, and the concentration of the top-k annotated tokens. Table 1 shows that the formality expressions in PREF-FAME-MT are substantially more diverse than in CoCoA-MT. Appendix A provides more details of the three datasets.

5 Experiments

5.1 Setup

Models. We have carried out fine-tuning experiments with three language models of small-medium size and two transformer architectures. The first, **NLLB 600M**, is an encoder-decoder transformer that has been pretrained to translate between more than 200 languages (Costa-Jussà et al., 2022). During fine-tuning, we have updated all its parameters, and to control the targeted attribute we have simply prepended a special token to the source sentence, i.e. {`formal`, `informal`} for formality and {`feminine`, `masculine`} for gender. The other two are **Qwen3 8B** and **EuroLLM 9B**, both decoder-only models. Qwen3 8B is a distilled version of the larger, original Qwen3 model (Yang et al., 2025) that has been pretrained to cover 119 languages, while EuroLLM 9B (Martins et al., 2025) has been trained predominantly in languages spoken in Europe. For fine-tuning, we have used LoRA (Hu et al., 2022) for efficiency, and used the same prompts used by Sarti et al. (2023) for both formality and gender control for more direct comparability. Finally, we have added a large, closed-source LLM, **Claude Sonnet 3** (Anthropic, 2024), as an additional baseline. For this model, we have used both zero-shot (ZS) and few-shot in-context (IC) prompting, with similar prompts to those used for Qwen3 and EuroLLM.

Training details. We have followed the standard approach for preference fine-tuning. First, we have instruction-tuned (IT) the models on the training set using the loss defined in Equation 1, for a maximum of 30 epochs or until convergence on the validation set. A dedicated model has been trained for each language pair and each dataset. Then, we have applied CPO (Equation 2) to the best checkpoint using a lower learning rate, again until convergence on the validation set or a maximum number of epochs. We have experimented with both the original contrastive data and with the synthetically-generated samples (CPO-SD in Section 5.2).

Evaluation. We have carried out a completely blind evaluation on the held-out test sets, prompting the models to predict two translations for each source sentence, i.e., a controlled translation for each attribute type. We have used two types of metrics: to evaluate the accuracy of the controlled attribute, we have used the established *Matched-Accuracy* ($\mathbf{M}_{\boldsymbol{Acc}}$) (Nadejde et al., 2022). With this metric, a prediction is labeled as of a certain attribute, say, a_1, if it contains at least one of the attribute-tagged spans annotated in the corresponding reference, y_{a_1}, and none of the spans annotated in y_{a_2}. The conditions are reversed when targeting a_2. However, according to this definition, some of the predictions may not fall into either category, which are simply omitted from the computation of the metric. For this reason, we add

Dataset	Model	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	BLEU	COMET	KIWI-XXL
CoCoA-MT	NLLB-600M-ZS (uncontrolled)	0.500	0.280	0.318	23.32	0.764	0.529
	NLLB-600M-IT	0.966	0.687	0.684	**40.74**	0.838	0.683
	NLLB-600M-IT + CPO	**0.992**†	**0.703**†	0.690	39.71	0.834	0.667
	NLLB-600M-IT + CPO-SD	0.979†	0.701†	**0.691**	40.28	**0.839**	**0.712**†
	Qwen3 8B-IT	0.964	0.670	0.653	37.73	0.835	0.726
	Qwen3 8B-IT + CPO	0.966†	0.671†	0.658†	37.98†	0.835	0.726
	Qwen3 8B-IT + CPO-SD	**0.969**†	**0.674**†	**0.659**†	**38.15**	**0.839**†	**0.735**†
	Claude 3 Sonnet-ZS	0.843	0.564	0.551	34.31	0.852	0.818
	Claude 3 Sonnet-IC (2-shot)	0.963	0.648	0.624	36.95	0.858	**0.820**
	Claude 3 Sonnet-IC (8-shot)	0.983	0.693	0.673	40.55	**0.864**	0.819
	Claude 3 Sonnet-IC (16-shot)	**0.987**	**0.703**	**0.684**	**41.17**	0.863	0.816
	RAMP (XGLM 7.5B)*	0.938	—	—	30.00	0.451	—
	RAMP (BLOOM 175B)*	0.973	—	—	41.90	0.711	—
MT-GenEval	NLLB-600M-ZS (uncontrolled)	0.500	0.335	0.352	34.32	0.814	**0.646**
	NLLB-600M-IT	0.940	0.673	0.591	37.52	0.825	0.578
	NLLB-600M-IT + CPO	**0.959**†	0.692†	0.603†	36.84	0.820	0.562
	NLLB-600M-IT + CPO-SD	0.953†	**0.695**†	**0.608**†	**37.96**†	**0.827**†	0.625†
	Qwen3 8B-IT	0.942	0.650	0.561	33.01	0.814	0.618
	Qwen3 8B-IT + CPO	0.941	0.651	0.562	33.06	0.815†	0.618
	Qwen3 8B-IT + CPO-SD	**0.949**†	**0.672**†	**0.579**†	**34.26**†	**0.825**†	**0.640**†
	Claude 3 Sonnet-ZS	0.936	0.612	0.512	29.21	0.750	0.500
	Claude 3 Sonnet-IC (2-shot)	0.933	0.673	0.602	**40.97**	**0.854**	**0.727**
	Claude 3 Sonnet-IC (8-shot)	**0.938**	**0.682**	**0.632**	39.07	0.843	0.719
	Claude 3 Sonnet-IC (16-shot)	0.933	0.669	0.625	36.89	0.826	0.702

Table 2: Average results on the CoCoA-MT and MT-GenEval test sets across all languages (en2all). The best result for each model type is in **bold**, and the best overall result for each dataset is underlined. (†) refers to statistically significant differences (p-value < 0.05) with respect to the respective IT baseline (more details in Appendix B.2). *From (Sarti et al., 2023).

two variants that account for all the predictions and can offer further insights: 1) a *Matched-Accuracy-Strict* ($\mathbf{M}_{Acc\text{-}Strict}$) metric, which, as the name suggests, is a stricter version of $\mathbf{M}_{Acc}$ that also marks the "neutral" predictions, i.e., sentences that do not contain annotated phrases from either of the references or contain mixed attributes, as incorrect; 2) a *Token-recall* ($\mathbf{T}_{recall}$) metric, which computes the percentage of all annotated spans from the target references that are present in the prediction.

In addition, we have used a set of metrics to evaluate the translation quality of the predictions: BLEU (Papineni et al., 2002), the most standard reference-based n-gram matching metric; COMET (Rei et al., 2020), a popular reference-based neural learned metric; and KIWI-XXL (Rei et al., 2023), a contemporary neural learned metric that we have used in a reference-free style to assess the adequacy of the translation irrespective of the reference.

More details of the hyperparameter settings and evaluation metrics are provided in Appendix B.

5.2 Results

CoCoA-MT and MT-GenEval

Table 2 reports the results for the original CoCoA-MT and MT-GenEval language pairs (en→x). For reference, the base NLLB zero-shot (ZS) model (which is tested in an uncontrolled translation setting, i.e. no control prompt is provided) has achieved significantly better translation quality on the MT-GenEval dataset (34.32 BLEU, 0.814 COMET) than on CoCoA-MT (23.32 BLEU, 0.764 COMET), potentially highlighting the different linguistic properties of these datasets or different extents of alignment with its pretraining. In any case, the instruction tuning of the base model (IT) has neatly improved the performance in every metric for both datasets. When applying CPO fine-tuning with the original contrastive data (IT+CPO), the model has achieved an additional improvement in all the attribute matching accuracy metrics, showing that preference fine-tuning has been effective at improving the control of the attribute. However, the results also show that this improvement has come at a cost in translation quality, with a clear drop in BLEU, COMET and KIWI-XXL com-

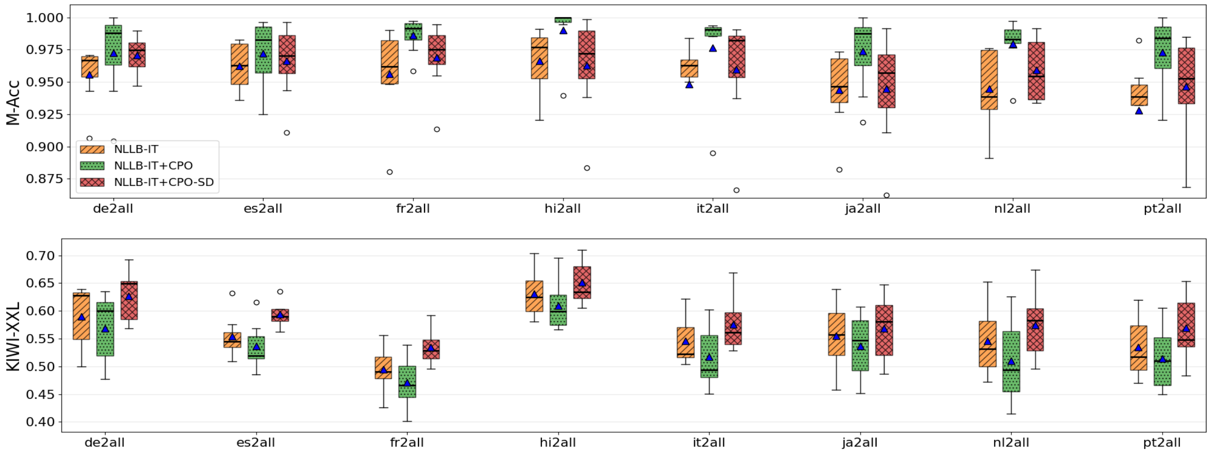


Figure 2: Comparative results for the non-English language pairs of CoCoA-MT.

pared to the IT baseline for both datasets. Conversely, when applying CPO to the synthetic data (IT+CPO-SD), the model has been able to achieve a much more even improvement across the various metrics compared to the instruction-tuned baseline, namely: for CoCoA-MT: +1.3 pp M_{Acc}, +1.4 pp $M_{Acc\text{-}Strict}$, +0.7 pp T_{recall}, −0.46 pp BLEU, +0.1 pp COMET, +2.9 pp KIWI-XXL; and for MT-GenEval: +1.3 pp M_{Acc}, +2.2 pp $M_{Acc\text{-}Strict}$, +1.7 pp T_{recall}, +0.44 pp BLEU, +0.2 pp COMET, +4.7 pp KIWI-XXL.

We have observed similar trends with the Qwen3 8B model. When applying CPO with the original data, the improvement in attribute accuracy does not translate into a corresponding improvement in translation quality, but with the synthetic data the improvements have extended to all metrics. While the differences are smaller than with the NLLB model (possibly due to the use of LORA with this model), the trend is still noticeable and confirms the potential of the approach for larger models, if sufficient computational resources for their fine-tuning are available. It is also worth noting that Qwen3 has performed worse than NLLB in most metrics, despite being a larger model, possibly due to its lack of machine translation pretraining.

We have also included Claude 3 Sonnet in the evaluation, using few-shot prompting in place of fine-tuning since it is not possible to locally fine-tune a closed-source model such as this. Despite this model being anecdotally much larger than NLLB (the actual number of parameters is undisclosed), it has been outperformed by NLLB in multiple metrics, e.g., −0.7 pp T_{recall} between their best configurations with CoCoA-MT, and −1.3 pp $M_{Acc\text{-}Strict}$ with MT-GenEval. We can therefore remark that our proposed approach has been able to lift the performance of smaller models such as NLLB and Qwen3 8B to levels comparable to those of a contemporary LLM. This can be important for a number of applications that cannot rely on LLMs due to cost or privacy reasons.

As a final comment, the results from NLLB on CoCoA-MT have also outperformed those reported by RAMP (Sarti et al., 2023), a state-of-the-art ACT approach. We note that a much smaller model such as NLLB, when fine tuned with preference optimization, has been able to surpass RAMP with BLOOM 175B in M_{Acc}.

CoCoA-MT Non-English Language Pairs

Figure 2 shows the M_{Acc} and KIWI-XXL scores of the NLLB models for the CoCoA-MT non-English language pairs (the results are averaged by source language). Overall, we have observed similar trends to those of the English pairs. The model fine-tuned with CPO on the original contrastive data has consistently outperformed the other models in terms of attribute accuracy in all language pairs, but has consistently underperformed in terms of KIWI-XXL. Conversely, the CPO model fine-tuned with the synthetic data has reported more conservative improvements in M_{Acc}, but has clearly outperformed all the others in terms of translation quality. This confirms its ability to strike a very desirable trade-off between attribute compliance and translation quality. The only notable exception to this trend has been *hi2all*, where the IT baseline has obtained a higher average M_{Acc} compared to the CPO-SD model. As a potential explanation for this, we have noted that the model that we have used to generate the synthetic data (Claude 3 Sonnet) has much lower translation metrics in

Model	M$_{Acc}$	M$_{Acc\text{-}Strict}$	T$_{recall}$	BLEU	COMET	KIWI-XXL
NLLB-600M-ZS	0.500	0.223	0.225	18.46	0.717	0.509
NLLB-600M-IT (orig. FAME-MT)	0.786	0.440	0.392	32.80	**0.837**	**0.789**
NLLB-600M-IT	0.841	0.493	0.446	**34.17**	**0.837**	0.785
NLLB-600M-IT + CPO	**0.902**†	**0.512**†	0.448	31.32	0.822	0.748
NLLB-600M-IT + CPO-SD	0.850†	0.501†	**0.449**	33.66	**0.837**	0.785
Qwen3 8B-IT	0.861	0.514	0.474	35.69	**0.850**	0.829
Qwen3 8B-IT + CPO	**0.862**	0.514	0.474	**35.74**	**0.850**	0.829
Qwen3 8B-IT + CPO-SD	**0.862**	**0.517**	**0.476**†	35.73	**0.850**	**0.830**
EuroLLM 9B-IT	0.826	0.468	0.433	35.16	0.853	0.848
EuroLLM 9B-IT + CPO	**0.840**†	**0.478**†	**0.439**†	35.13	0.855†	0.851†
EuroLLM 9B-IT + CPO-SD	0.823	0.469	0.434	**35.25**	**0.856**†	**0.855**†

Table 3: Average results over the test sets of the 7 language pairs of PREF-FAME-MT. The best result for each model type is in **bold**, and the best overall result for each dataset is underlined. (†) refers to statistically significant differences (p-value < 0.05) with respect to the respective IT baseline (more details in Appendix B.2).

MT-GenEval (en-es)	PREF-FAME-MT (es-en)
source They were married in Washington, D.C., on 16 February 1922 when she was 17 years old; Converse was about a decade older and had been married before.	**source** La mayoría de los costes se pueden atribuir al consumo de energía.
reference (masculine) Se casaron en Washington, D.C., el 16 de febrero de 1922, cuando **él** tenía 17 años; Converse era una década mayor y había estado **casado** antes.	**reference** (informal) Most of the costs **come from** energy consumption.
NLLB-IT (KIWI-XXL: 0.060) Se casaron en Washington, D. C., el 16 de febrero de 1922, cuando **él** tenía 17 años; Converse tenía unos diez años de edad y se había **casado** antes.	**NLLB-IT** (KIWI-XXL: 0.999) Most of the costs can be attributed to energy consumption.
NLLB-IT + CPO (KIWI-XXL: 0.060) Se casaron en Washington, D. C., el 16 de febrero de 1922, cuando **él** tenía 17 años; Converse tenía unos diez años de edad y se había **casado** antes.	**NLLB-IT + CPO** (KIWI-XXL: 0.988) Most of the costs **come from** power consumption.
NLLB-IT + CPO-SD (KIWI-XXL: 0.813) Se casaron en Washington, D. C., el 16 de febrero de 1922, cuando **él** tenía 17 años; Converse tenía unos diez años más y se había **casado** antes.	

Table 4: Prediction examples for MT-GenEval en-es (left) and PREF-FAME-MT es-en (right).

language pairs that include Hindi (see Appendix A.2 for details). As a result, far fewer synthetic samples have met the validity criteria of Algorithm 1, and can be argued to be of overall lower quality. This is likely to have affected the performance of the proposed CPO-SD approach in these language pairs, and stresses the importance of high-quality synthetic samples for its effectiveness.

PREF-FAME-MT Results

Table 3 shows the results obtained by the NLLB model on the newly-released PREF-FAME-MT dataset. Given the substantially lower scores obtained by the ZS model, we can argue that this dataset is more challenging compared to CoCoA-MT. For this dataset, we have carried out the instruction-tuning experiments in two alternative ways: a) using the original training data of FAME-MT, and b) using instead the synthetic contrastive training data generated as described in Section 4. It is interesting to note that the instruction-tuning with the synthetic data has achieved much higher attribute matching scores (e.g., $+5.5$ pp M$_{Acc}$), and a marked increase in BLEU ($+1.37$ pp). While the test set has been generated in the same way and the results may suggest circularity, all its samples have been carefully manually validated. In turn, the model fine-tuned with CPO has displayed a similar trend to the previous experiments: higher attribute matching scores compared to the instruction-tuned model ($+6.1$ pp in M$_{Acc}$), and lower scores in translation quality metrics (e.g., -3.7 pp in KIWI-XXL). Finally, similarly to the other datasets, we have observed that the model trained with synthetic data has achieved more balanced improvements, recov-

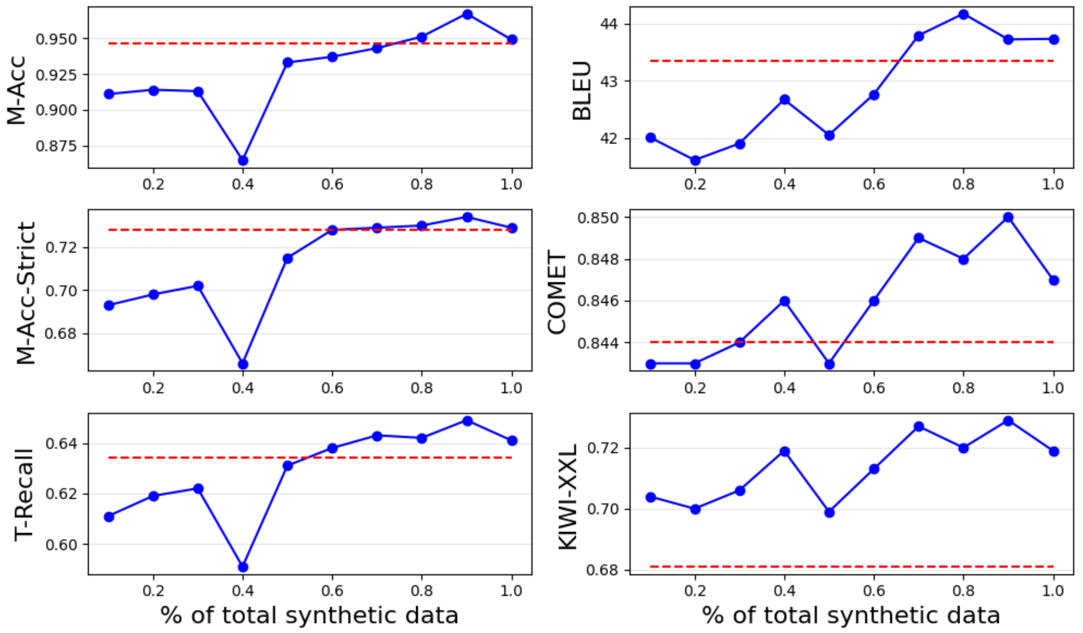


Figure 3: Sensitivity analysis of the different metrics over the CoCoA-MT en-es test set. The red dashed line shows the performance of the respective IT baseline.

ering the COMET and KIWI-XXL scores of the IT baseline, while mildly improving its attribute control performance in terms of M_{Acc}, $\mathrm{M}_{Acc\text{-}Strict}$ and T_{recall}.

In turn, the decoder-only models, Qwen3 8B and EuroLLM 9B, have achieved mild improvements in the controlled attribute when applying preference optimization. In contrast to NLLB, these models have retained translation quality when fine-tuned with CPO on top of the model instruction-tuned with the synthetic contrastive training data (IT). A plausible explanation is that their larger capacity, compared to NLLB, makes them more robust to generalization. Between the two models, Qwen3 seems to have performed better in terms of target attribute control, while EuroLLM has proved slightly better in translation quality.

Qualitative Analysis

Table 4 shows two prediction examples from the MT-GenEval and PREF-FAME-MT test sets, respectively. For the former, all models have correctly mapped the gender pronouns to masculine, but only the model trained with the synthetic data has accurately translated the phrase "a decade older" as "ten years older" (the others have rendered it as "ten years old", and have been heavily penalized by KIWI-XXL). For the latter, the model trained with CPO has generated a translation with a clear informal tone, as requested, using the phrase "come from" instead of "can be attributed".

Synthetic Data Sensitivity Analysis

Finally, we have investigated the sensitivity of the results to the use of different amounts of synthetic data for fine-tuning over the CoCoA-MT en-es test set. Figure 3 shows a clear trend: using more synthetic data tends to improve the results in all the metrics. Further results over the CoCoA-MT en-de test set are presented in Appendix E.

6 Conclusion

This paper has explored the use of preference optimization over attribute-contrastive ACT datasets for improving the performance of small-medium size language models in ACT tasks. Our experimental results show that the proposed preference optimization with attribute-contrastive training data has been able to improve attribute-matching metrics in most cases. In addition, to counter a corresponding drop in translation quality, we have proposed generating synthetic contrastive data where both attribute compliance and translation quality are controlled. Extensive experiments over three datasets, including the newly-released PREF-FAME-MT, have shown that the proposed approach has been able to lift the performance of small-medium size language models such as NLLB and Qwen3 8B to levels comparable to those of a state-of-the-art LLM in many cases.

Limitations

We acknowledge a number of limitations in the proposed approach. The first is that performing knowledge distillation by sampling synthetic training data may inherit the biases of the LLMs used to generate such data, potentially propagating problematic patterns from larger models to smaller ones. Additionally, the performance of LLMs varies greatly across languages, and for many low-resource languages there may not exist any model capable of generating synthetic data of adequate quality.

In the second place, the gender-controlled translation dataset that we have employed in our work, MT-GenEval, only considers masculine and feminine genders, excluding non-binary or gender-diverse identities. As a result, our gender-controlled translation models may reinforce binary gender assumptions, even in gender-neutral source expressions. More work is needed to build datasets that amend these limitations.

Finally, we have not been able to carry out a very exhaustive search of the various hyperparameters involved in the training and inference of the models, due to time limitations. A deeper search of these hyperparameters could further improve the performance of the models, and, potentially, change their relative rankings.

Acknowledgments

This research project was supported by an Amazon Research Award (ARA). We also thank the anonymous reviewers for all their precious comments and suggestions.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Model Card.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2020. GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, France. European Language Resources Association.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Guofeng Cui, Pichao Wang, Yang Liu, Zemian Ke, Zhu Liu, and Vimal Bhat. 2025. CRPO: Confidence-reward driven preference optimization for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 560–574, Vienna, Austria. Association for Computational Linguistics.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, 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.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, and Pang Wei Koh. 2025. The delta learning hypothesis: Preference tuning on weak data can yield strong gains. *arXiv preprint arXiv:2507.06187*.

D. S. Hirschberg. 1975. A linear space algorithm for computing maximal common subsequences. *Commun. ACM*, 18(6):341–343.

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*.

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.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024a. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Minwoo Lee, Hyukhun Koh, Minsung Kim, and Kyomin Jung. 2024b. Fine-grained Gender Control in Machine Translation with Large Language Models. 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 5416–5430, Mexico City, Mexico. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2024. How transferable are attribute controllers on pretrained multilingual translation models? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 334–348, St. Julian's, Malta. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and 1 others. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. 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.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 845–854, New York, NY, USA. Association for Computing Machinery.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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.

Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. 2023. RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1476–1490, Toronto, Canada. Association for Computational Linguistics.

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.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. 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 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaden Uhlig, Joern Wuebker, Raphael Reinauer, and John DeNero. 2024. Cross-lingual human-preference alignment for neural machine translation with direct quality optimization. *arXiv preprint arXiv:2409.17673*.

Dario Vajda, Domen Vreš, and Marko Robnik-Šikonja. 2025. Improving llms for machine translation using synthetic preference data. *arXiv preprint arXiv:2508.14951*.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2023. Controlling styles in neural machine translation with activation prompt. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2606–2620, Toronto, Canada. Association for Computational Linguistics.

Dawid Wisniewski, Zofia Rostek, and Artur Nowakowski. 2024. FAME-MT dataset: Formality awareness made easy for machine translation purposes. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 164–180, Sheffield, UK. European Association for Machine Translation (EAMT).

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024a. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024b. Is dpo superior to ppo for llm alignment? a comprehensive study. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Details of the Datasets

To provide further details of the CoCoA-MT dataset[4], Table 5 reports the number of training samples for each language pair. For the training set, we have included all the contrastive samples available from both subdomains, i.e., *Topical Chat* and *Telephony*. For validation, we have held out 10% of samples randomly sampled from the training data. We have used the official test sets released by the authors including the three available subdomains, i.e., *Topical Chat*, *Telephony* and *Call Center*. Non-English language pairs have been aligned using the shared English sentences. As such, they include a subset of sentences from the original dataset. However, while the English source sentences are only neutral, in the non-English pairs we include both the formal and informal variants as separate source sentences. Therefore, we have doubled the amount of available samples.

In turn, Table 6 shows the number of training samples in the MT-GenEval dataset[5]. We have used the sentence-level dev and test splits released by the authors, not including any disambiguating context sentences in the source. From the dev set, we use 1,000 samples for training and 200 for validation. But we use both the masculine and feminine version of every English sentence as separate inputs, therefore doubling the dataset size, obtaining 2,000, 400 and 600 train, dev, and test samples, respectively.

src	tgt	train	val	test	src	tgt	train	val	test
				Original dataset					
en	de	360	40	600	en	it	360	40	600
	es	360	40	600		ja	900	100	594
	fr	360	40	600		nl	360	40	597
	hi	360	40	600		pt	360	40	599
				Non-English aligned pairs					
de	es	691	77	1,144	it	de	311	35	736
	fr	547	61	990		es	313	35	732
	hi	671	75	1,124		fr	329	37	764
	it	311	35	736		hi	309	35	736
	ja	235	27	194		ja	239	27	168
	nl	185	21	170		nl	160	18	288
	pt	183	21	190		pt	178	20	304
es	de	691	77	1,144	ja	de	235	27	194
	fr	550	62	978		es	248	28	200
	hi	680	76	1,134		fr	255	29	188
	it	313	35	732		hi	237	27	200
	ja	248	28	200		it	239	27	168
	nl	185	21	174		nl	237	27	162
	pt	183	21	204		pt	235	27	128
fr	de	547	61	990	nl	de	185	21	170
	es	550	62	978		es	185	21	174
	hi	536	60	960		fr	185	21	210
	it	329	37	764		hi	178	20	174
	ja	255	29	188		it	160	18	288
	nl	185	21	210		ja	237	27	162
	pt	174	20	244		pt	318	36	450
hi	de	671	75	1,124	pt	de	183	21	190
	es	680	76	1,134		es	183	21	204
	fr	536	60	960		fr	174	20	244
	it	309	35	736		hi	180	20	216
	ja	237	27	200		it	178	20	304
	nl	178	20	174		ja	235	27	128
	pt	180	20	216		nl	318	36	450

Table 5: Number of contrastive samples in each language pair in the CoCoA-MT dataset.

A.1 PRE-FAME-MT Dataset Construction

Table 7 shows the number of samples per language pair in the new PREF-FAME-MT dataset. As described in Section 4, the dataset has been created by extending the existing FAME-MT dataset (Wisniewski et al., 2024)[6] with a synthetic contrastive sample for each original sample. The original dataset contains 50K formal, 50K informal and 50K neutral translations for each language pair (150K total). For the 7 selected language pairs, we have divided the data into 145K/1.5K/3K for training, validation and testing, respectively, maintaining the balance of each formality type. Then, we have used a style-transfer prompt (Figure 4) with a Claude 4 Sonnet model[7] to generate the synthetic contrastive samples. We have only generated synthetic samples for translation labeled as formal or informal, while for neutral samples we have simply retained the original translation.

In addition, we have carried out a human-evaluation of the generated test set in order to remove inexact contrastive samples and obtain reliable test data. We have randomly sampled 300 samples from the 3K reserved for testing in each language pair, 100 from each class (formal, informal, neutral). Then, we have asked 7 professional translators (one for each target language) to validate the correctness of the contrastive samples generated for the formal and informal original translations. Figure 5 shows the instructions provided to the annotators, which have been used to evaluate whether the

[4] CoCoA-MT data source (CDLA-Sharing-1.0 License): https://github.com/amazon-science/contrastive-controlled-mt

[5] MT-GenEval data source (CC-BY-SA-3.0 license): https://github.com/amazon-science/machine-translation-gender-eval

[6] FAME-MT data source: https://github.com/laniqo-public/fame-mt

[7] AWS Bedrock model ID: apac.anthropic.claude-sonnet-4-20250514-v1:0

src	tgt	train	val	test
en	ar	2,000	400	600
	de	2,000	400	600
	es	2,000	400	600
	fr	2,000	400	600
	hi	2,000	400	600
	it	2,000	400	600
	pt	2,000	400	600
	ru	2,000	400	600

Table 6: Number of contrastive samples in each language pair in the MT-GenEval dataset.

src	tgt	train	val	test
da	es	145.5K	1,500	208
de	fr	145.5K	1,500	203
en	de	145.5K	1,500	225
es	en	145.5K	1,500	249
it	nl	145.5K	1,500	246
pl	it	145.5K	1,500	248
ru	pt	145.5K	1,500	217

Table 7: Number of contrastive samples in each language pair in the PREF-FAME-MT dataset.

contrastive pair was CORRECT or INCORRECT. Table 8 shows the results of the human-evaluation. Correct contrastive examples range within 50-75% of the generated samples, depending on the language pair. Finally, we have only kept the samples that annotators have labeled as CORRECT and concatenate them with the 100 neutral samples, obtaining the test sets reported in Table 7.

A.2 Synthetic Data

In this section we provide further details on the generation of the synthetic contrastive data described in Section 3.2, contrasting both attribute compliance and translation quality in the sample pairs with the aim to improve ACT performance. For CoCoA-MT and MT-GenEval, as input data, we have used the source sentences of their respective training sets for every language pair. We have generated the translation samples with Claude 3 Sonnet[8] with 16-shot IC learning, and the prompt described in Figure 8. For PREF-FAME-MT, we have sampled 20,000 sentences from the training set of each language pair and formality target, and have used Claude 4.5 Sonnet[9] for sampling instead. For each source sentence and target attribute combination in the training data, we have sampled 32 translations (i.e., $k = 32$ in Algorithm 1), varying the temperature parameter to obtain diverse predictions ($temp = \{0.0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$). We have relied on the attribute labeling criteria used in the M_{Acc} metric to check whether the attribute of the sample matched the attribute of the preferred reference (y_{ref}). To obtain the KIWI score, we have used the same model as in evaluation, KIWI-XXL (Rei et al., 2023)[10]. The sampling costs of the Claude models with Bedrock's API was $0.00014 USD on average per sample.

```
Here is a sentence written in <FORMALITY> style:
<TRANSLATION>; Please provide a version of the sentence
written in <CONTRASTIVE_FORMALITY> style between curly
brackets, making minimal changes and without changing the
meaning of the sentence.
```

Figure 4: Style-transfer prompt for PREF-FAME-MT.

```
Review the sentences below and determine if the
contrastive sentence correctly transforms the language
style while maintaining the same meaning, and is
grammatically correct.

Label as correct only if:
1. Both sentences are grammatically correct
2. Both sentences are semantically equivalent (same
meaning)
3. The original sentence and all it's highlighted token
spans suggest a <FORMALITY> style
4. The contrastive sentence and all it's highlighted
token spans suggest a <CONTRASTIVE_FORMALITY> style

Original Sentence: <ORIGINAL_SENTENCE>
Contrastive Sentence: <CONTRASTIVE_SENTENCE>
```

Figure 5: Instructions for the human annotators.

Table 9 and 10 show the statistics of the generated synthetic data for the CoCoA-MT and MT-GenEval datasets, respectively. These are samples that match the selection criteria described in Algorithm 1. In CoCoA-MT, for most language pairs we have been able to generate samples that represent between 60-90% of the original training dataset, except for language pairs that have Japanese ($\sim 40\%$) and Hindi ($\sim 20\%$) as target languages. In MT-GenEval, we have been able to generate samples that represent $\sim 50\%$ of the original data, except for Hindi (9.2%). This shows that the ability to generate synthetic data depends on the quality of the LLM in the specific language pair, particularly the target language. In this case, Claude 3 Sonnet has shown a lower performance in Hindi. In future work, we aim to experiment with better LLMs

[8] AWS Bedrock model ID: anthropic.claude-3-sonnet-20240229-v1:0

[9] AWS Bedrock model ID: au.anthropic.claude-sonnet-4-5-20250929-v1:0

[10] Hugginface: Unbabel/wmt23-cometKiwi-da-xxl

Language Pair	# Evaluated	# Correct	%
da-es	200	108	54.0
de-fr	200	103	51.5
en-de	200	125	62.5
es-en	200	149	74.5
it-nl	200	146	73.0
pl-it	200	148	74.5
ru-pt	200	117	58.5

Table 8: Human-evaluation results.

for each language. Table 11 shows the synthetic data generated for the PREF-FAME-MT dataset. In general, the percentage of successful contrastive data is lower than for the other datasets. This could be due to various reasons: 1) the noisy nature of the FAME-MT dataset, 2) noisy contrastive translations, and 3) the more challenging and diverse nature of this dataset. We will explore this in future work.

B Experimental Setup Details

In this section, we provide further details on the models, prompts, their hyperparameter settings, and evaluation metrics.

B.1 Models

As described in Section 5.1, in our experiments we have used three open-source language models of small-medium size downloaded from Huggingface. The first is NLLB-600[11], an encoder-decoder transformer based on the BART architecture with 600M parameters, pretrained to translate between more than 200 languages. Figure 6 shows the prompt employed as control signal for this model for each case, i.e., formality and gender, where we have simply prepended the target attribute label to the input sentence in the source language (<INPUT_SRC>). The second model is Qwen3 8B[12], a decoder-only language model with 8B parameters. The third is EuroLLM 8B[13], another decoder-only model trained mainly in European languages. Figure 7 shows the respective prompts used as control signals for these two models, where we replace <TGT_LANG> with the language code of the target sentence.

For transparency and reproducibility of the results, Table 12 describes the hyperparameters used with both models. Note that as validation criteria for early stopping both during IT and CPO training, we have used the average between BLEU and M_{Acc}, thus aiming to select checkpoints that balance attribute accuracy and translation quality.

src	tgt	# samples	%	src	tgt	# samples	%
Original dataset							
en	de	562	78.0	en	it	471	65.4
	es	470	65.2		ja	823	45.7
	fr	504	70.0		nl	590	81.9
	hi	143	19.8		pt	585	81.2
Non-English aligned pairs							
de	es	984	71.2	it	de	457	73.4
	fr	784	71.6		es	438	69.9
	hi	309	23.0		fr	462	70.2
	it	437	70.2		hi	116	18.7
	ja	182	38.7		ja	204	42.6
	nl	322	87.0		nl	287	89.6
	pt	301	82.2		pt	298	83.7
es	de	1,053	76.1	ja	de	381	81.0
	fr	751	68.2		es	360	72.5
	hi	285	20.9		fr	378	74.1
	it	398	63.5		hi	84	17.7
	ja	210	42.3		it	331	69.2
	nl	337	91.0		nl	413	87.1
	pt	286	78.1		pt	369	78.5
fr	de	802	1,094	nl	de	313	84.5
	es	730	66.3		es	285	77.0
	hi	235	21.9		fr	295	79.7
	it	404	61.3		hi	118	33.1
	ja	206	40.3		it	255	79.6
	nl	322	87.0		ja	198	41.7
	pt	279	80.1		pt	518	81.4
hi	de	1,017	75.7	pt	de	284	77.5
	es	980	72.0		es	254	69.3
	fr	794	74.0		fr	264	75.8
	it	421	68.1		hi	94	26.1
	ja	210	44.3		it	265	74.4
	nl	320	89.8		ja	169	35.9
	pt	303	84.1		nl	557	87.57

Table 9: Number of synthetic contrastive samples in each language pair in the CoCoA-MT dataset, and the percentage it represents with respect to the number of inputs in the training data (unique input sentences and target attribute combinations).

As a baseline comparison, we have used the Claude 3 Sonnet model, i.e., the same model used to generate the synthetic data (Appendix A.2). Figure 8 shows the prompts used with this model, which are similar to those used for Qwen3 8B. The main difference is that when we do few-shot IC learning, we randomly sample examples from the training dataset (<EXAMPLE_SRC>, <EXAMPLE_TGT>) and their annotated tokens (<FORMALITY_TOKENS> or <GENDER_TOKENS>). We apply ZS, 2-shot, 8-shot and 16-shot IC testing. When retrieving examples from the training data, we have evenly distributed them across both attribute types. We have disabled the

[11] Huggingface: `facebook/nllb-200-distilled-600M`

[12] Huggingface: `Qwen/Qwen3-8B`

[13] Huggingface: `utter-project/EuroLLM-9B-Instruct`

src	tgt	# samples	%
en	ar	2,092	52.3
	de	1,970	49.2
	es	1,770	44.2
	fr	2,006	50.1
	hi	369	9.2
	it	1,821	45.5
	pt	2,048	51.2
	ru	1,868	46.7

Table 10: Number of synthetic contrastive samples in each language pair in the MT-GenEval dataset, and the percentage it represents with respect to the number of inputs in the training data (unique input sentences and target attribute combinations).

src	tgt	# samples	%
da	es	7,127	17.8
de	fr	8,148	20.3
en	de	4,745	11.8
es	en	3,607	9.0
it	nl	8,759	21.8
pl	it	4,693	11.7
ru	pt	4,237	10.5

Table 11: Number of synthetic contrastive samples in each language pair in the PREF-FAME-MT dataset, and the percentage it represents with respect to the number of inputs in the training data (unique input sentences and target attribute combinations).

temperature parameter to limit the diversity of the generated samples during testing.

B.2 Evaluation Metrics

As discussed in Section 5.1, we have relied on the corpus-level M$_{Acc}$ metric proposed by Nadejde et al. (Nadejde et al., 2022). In this metric, the estimated attribute of a prediction is correct if:

1. It contains at least one lexical match with the list of annotated phrases (i.e., words between [F] ... [/F]) made in the target reference translation.

2. It contains not a single lexical match with the list of annotated phrases in the contrastive reference translation.

Then, considering that n_c is the number of correct attribute predictions in the test set and n_i is the number of incorrect attribute predictions, M$_{Acc}$ is simply computed as:

$$M_{Acc} = \frac{n_c}{n_c + n_i} \qquad (3)$$

Formality-control prompt:

```
<FORMALITY> <INPUT_SRC>
```

Gender-control prompt:

```
<GENDER> <INPUT_SRC>
```

Figure 6: Control prompt for the NLLB model.

Formality-control prompt:

```
Here is a sentence {<INPUT_SRC>}; Here is its
<TGT_LANG> translation written in <FORMALITY> style {
```

Gender-control prompt:

```
Here is a sentence {<INPUT_SRC>}; Here is its
<TGT_LANG> translation in which every mentioned person's
gender is <GENDER> {
```

Figure 7: Control prompt for the Qwen3 8B and EuroLLM 9B models.

However, note that a prediction may not satisfy neither of the aforementioned conditions. On the one hand, the generated sentence may not contain any lexical match with the target nor the contrastive reference. On the other, the sentence may contain mixed matches from both references. M$_{Acc}$ ignores both those cases.

For this reason, we have proposed two additional, complementary metrics. The first is a stricter version of M$_{Acc}$:

$$M_{Acc\text{-}Strict} = \frac{n_c}{n_c + n_i + n_n + n_m} \qquad (4)$$

where n_n is the number of neutral predictions, i.e., sentences that do not match token-level markers in either of the references, and n_m are the number of mixed predictions that contain matches from both references. In M$_{Acc\text{-}Strict}$ both cases are considered incorrect predictions.

The second metric is the T$_{Recall}$ metric. Note from Figure 1a that references may contain more than one attribute marker in a reference. T$_{Recall}$ measures the percentage of matched annotations in all target references in the test set:

$$T_{Recall} = \frac{t_c}{t_c + t_m} \qquad (5)$$

where t_c is the total number of matched annotations and t_m is the number of missed annotations.

Meanwhile, we have carried out statistical significance tests of the chosen metrics comparing the fine-tuned baselines (IT) against both the proposed model variants (IT+CPO and IT+CPO-SD), and reported the results in Tables 2 and 3. In specific, we have computed the sample-level pairwise

	NLLB-600M	Qwen3 8B / EuroLLM 9B
batch size	4	1
max sequence length	512	512
epochs	30	30
early stopping	true	true
patience	10	10
val. criteria	avg(BLEU,M_{Acc})	avg(BLEU,M_{Acc})
val. check interval (IT)	1 epoch	1 epoch
val. check interval (CPO)	30 steps	30 steps
learning rate (IT)	5e-5	5e-5
learning rate (CPO)	5e-6	5e-7
β (CPO)	1.0	1.0
λ (CPO)	0.25	0.25
padding side	right	left
fixed seed	42	42
LoRA	false	true
LoRA rank	-	16
LoRA α.	-	32
LoRA dropout	-	0.1
LoRA bias	-	None

Table 12: Hyperparameters for the NLLB-600M and Qwen3 8B models.

Formality-control prompt:

```
Here is a sentence {<EXAMPLE_SRC>}; Please provide the
<TGT_LANG> translation written in <FORMALITY> style
between curly brackets: {<EXAMPLE_TGT>}; The translated
sentence conveys a <FORMALITY> style by using words
such as <FORMALITY_TOKENS>.
Here is a sentence {<INPUT_SRC>}; Please provide the
<TGT_LANG> translation written in <FORMALITY> style
between curly brackets: {
```

Gender-control prompt:

```
Here is a sentence {<EXAMPLE_SRC>}; Please provide
the <TGT_LANG> translation in which every mentioned
person's gender is <GENDER> between curly brackets:
{<EXAMPLE_TGT>}; In the translation, the <GENDER>
gender of the person is made explicit by words such as
<GENDER_TOKENS>.
Here is a sentence {<INPUT_SRC>}; Please provide the
<TGT_LANG> translation in which every mentioned person's
gender is <GENDER> between curly brackets: {
```

Figure 8: Control prompt for Claude 3 Sonnet.

source (CoCoA-MT en-es) I read them sometimes, I mostly rely on Rotten Tomatoes, and you?
reference (informal) Los leo de vez en cuando, pero me baso sobre todo en Rotten Tomatoes, ¿y **tú**?
NLLB-IT (Kiwi-XXL: 0.098) Las leo a veces, dependo sobre todo de Los tomates podridos, ¿y **tú**?
NLLB-IT + CPO (Kiwi-XXL: 0.271) Las leo a veces, me fio principalmente a Tomates Rochosos, ¿y **tú**?
NLLB-IT + CPO-SD (Kiwi-XXL: 0.813) Las leo a veces, me fio principalmente a Rotten Tomatoes, ¿y **tú**?

Table 13: Prediction example from CoCoA-MT en-es test set.

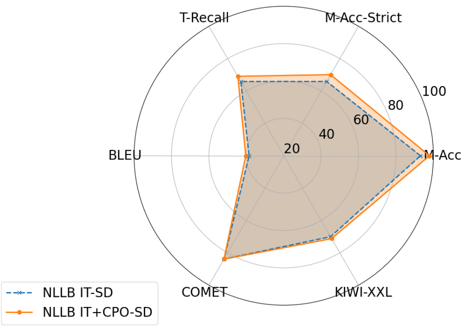


Figure 9: Instruction tuning with the synthetic data (IT-SD) vs preference optimization with the same data (CPO-SD) (CoCoA-MT test sets, en2all average).

bootstrap tests for the M-Acc, M-Acc-Strict, T-Recall, COMET, and KIWI-XXL metrics, following the recommendation from Dror et al. (2018), and corpus-level paired t-tests for the BLEU metric.

C Qualitative Analysis Example

Table 13 shows another prediction example for qualitative analysis. Only the model trained with the synthetic data has been able to retain the term "Rotten Tomatoes" correctly untranslated since it is a proper noun, and as such has achieved a much higher Kiwi-XXL score.

D Instruction-Tuning Ablation With the Synthetic Data

For further insight on the synthetic data, we have carried out an experiment where we have used the synthetic preferred samples with a conventional NLL objective (Equation 1) after fine-tuning with the regular training set. Figure 9 shows a radar plot of the six metrics for this configuration (noted as IT-SD) and preference optimization with the same data (CPO-SD). The plot shows that the performance in translation quality has been roughly comparable, but the performance of preference optimization has been noticeably higher in all the attribute matching metrics. This confirms the general effectiveness of the preference optimization framework at improving the control of the targeted attribute.

E Synthetic Data Sensitivity Analysis

In Figure 10 we show the sensitivity analysis carried out in the en-de CoCoA-MT dataset. These results confirm the trend, more synthetic data contributes to better results.

F Results per Language Pair

For transparency and reproducibility of the experiments, Tables 14-26 show the results of the experiments for each language pair.

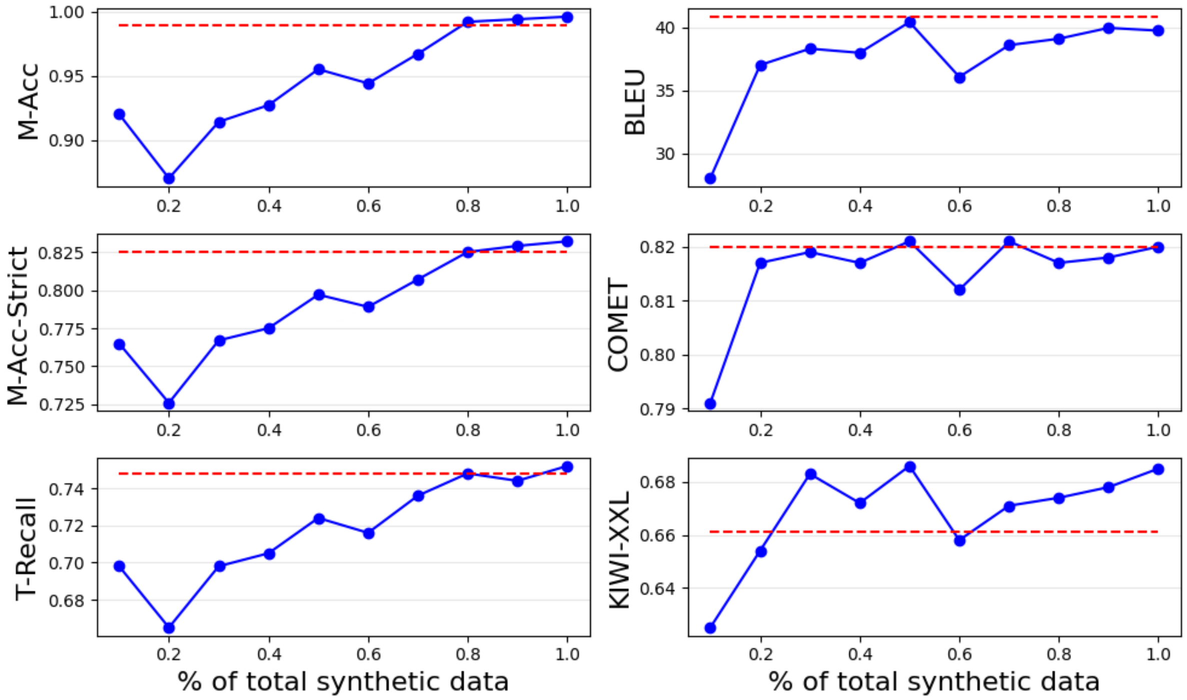


Figure 10: Sensitivity analysis of the different metrics over the CoCoA-MT en-de test set. The red dashed line shows the performance of the respective IT baseline.

Lang. Pair	Model	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	BLEU	COMET	KIWI-XXL
en-de	NLLB-600M-ZS (uncontrolled)	0.500	0.310	0.272	21.94	0.746	0.509
	NLLB-600M-IT	0.989	0.825	0.748	40.87	0.820	0.661
	NLLB-600M-IT + CPO	0.998	0.835	0.747	39.27	0.813	0.638
	NLLB-600M-IT + CPO-SD	0.996	0.832	0.752	39.76	0.820	0.685
	Qwen3 8B-IT	0.989	0.791	0.706	37.99	0.818	0.698
	Qwen3 8B-IT + CPO	0.989	0.790	0.705	37.98	0.818	0.698
	Qwen3 8B-IT + CPO-SD	0.989	0.794	0.708	37.95	0.818	0.700
	Claude 3 Sonnet-ZS	0.988	0.757	0.655	35.81	0.847	0.809
	Claude 3 Sonnet-IC (2-shot)	1.000	0.791	0.702	39.85	0.856	0.819
	Claude 3 Sonnet-IC (8-shot)	0.998	0.827	0.742	43.51	0.860	0.814
	Claude 3 Sonnet-IC (16-shot)	1.000	0.828	0.747	43.97	0.861	0.812
en-es	NLLB-600M-ZS (uncontrolled)	0.500	0.330	0.376	32.41	0.807	0.613
	NLLB-600M-IT	0.946	0.728	0.634	43.35	0.844	0.681
	NLLB-600M-IT + CPO	0.996	0.750	0.642	40.70	0.838	0.666
	NLLB-600M-IT + CPO-SD	0.949	0.729	0.641	43.73	0.847	0.719
	Qwen3 8B-IT	0.982	0.754	0.655	43.73	0.849	0.736
	Qwen3 8B-IT + CPO	0.982	0.755	0.656	43.73	0.849	0.736
	Qwen3 8B-IT + CPO-SD	0.989	0.763	0.663	44.58	0.854	0.751
	Claude 3 Sonnet-ZS	0.951	0.714	0.644	42.10	0.864	0.822
	Claude 3 Sonnet-IC (2-shot)	0.985	0.749	0.658	43.95	0.866	0.818
	Claude 3 Sonnet-IC (8-shot)	0.989	0.778	0.690	46.80	0.871	0.813
	Claude 3 Sonnet-IC (16-shot)	0.993	0.786	0.696	47.44	0.872	0.813
en-fr	NLLB-600M-ZS (uncontrolled)	0.500	0.309	0.299	24.87	0.742	0.481
	NLLB-600M-IT	0.974	0.790	0.695	39.41	0.819	0.640
	NLLB-600M-IT + CPO	0.999	0.805	0.701	38.90	0.816	0.628
	NLLB-600M-IT + CPO-SD	0.988	0.823	0.726	41.02	0.823	0.690
	Qwen3 8B-IT	0.980	0.798	0.686	39.94	0.835	0.767
	Qwen3 8B-IT + CPO	0.980	0.796	0.687	40.30	0.834	0.762
	Qwen3 8B-IT + CPO-SD	0.980	0.798	0.688	40.04	0.835	0.767
	Claude 3 Sonnet-ZS	0.971	0.752	0.659	36.88	0.839	0.802
	Claude 3 Sonnet-IC (2-shot)	0.986	0.775	0.683	39.60	0.846	0.809
	Claude 3 Sonnet-IC (8-shot)	0.994	0.800	0.711	42.89	0.853	0.811
	Claude 3 Sonnet-IC (16-shot)	0.996	0.818	0.729	43.52	0.851	0.809
en-hi	NLLB-600M-ZS (uncontrolled)	0.500	0.148	0.329	22.30	0.725	0.550
	NLLB-600M-IT	0.985	0.304	0.791	34.74	0.810	0.691
	NLLB-600M-IT + CPO	0.993	0.308	0.785	33.58	0.806	0.684
	NLLB-600M-IT + CPO-SD	0.960	0.290	0.761	34.52	0.806	0.716
	Qwen3 8B-IT	0.907	0.252	0.675	25.99	0.757	0.611
	Qwen3 8B-IT + CPO	0.923	0.268	0.708	26.64	0.763	0.621
	Qwen3 8B-IT + CPO-SD	0.926	0.261	0.698	27.47	0.775	0.646
	Claude 3 Sonnet-ZS	0.698	0.197	0.487	24.23	0.806	0.752
	Claude 3 Sonnet-IC (2-shot)	0.958	0.285	0.622	26.93	0.812	0.763
	Claude 3 Sonnet-IC (8-shot)	0.981	0.309	0.687	29.84	0.818	0.768
	Claude 3 Sonnet-IC (16-shot)	0.991	0.321	0.713	30.50	0.820	0.768

Table 14: CoCoA-MT en2all results per language pair.

Lang. Pair	Model	M_{Acc}	$\mathrm{M}_{Acc\text{-}Strict}$	T_{recall}	BLEU	COMET	KIWI-XXL
en-it	NLLB-600M-ZS (uncontrolled)	0.500	0.278	0.286	29.58	0.794	0.536
	NLLB-600M-IT	0.969	0.743	0.648	45.41	0.850	0.627
	NLLB-600M-IT + CPO	0.983	0.746	0.658	44.35	0.847	0.620
	NLLB-600M-IT + CPO-SD	0.971	0.744	0.659	45.05	0.854	0.676
	Qwen3 8B-IT	0.989	0.722	0.620	42.85	0.856	0.720
	Qwen3 8B-IT + CPO	0.988	0.720	0.625	43.14	0.857	0.720
	Qwen3 8B-IT + CPO-SD	0.996	0.729	0.625	42.92	0.860	0.733
	Claude 3 Sonnet-ZS	0.784	0.505	0.450	41.90	0.866	0.798
	Claude 3 Sonnet-IC (2-shot)	0.979	0.615	0.529	43.57	0.872	0.797
	Claude 3 Sonnet-IC (8-shot)	0.997	0.687	0.585	47.88	0.877	0.788
	Claude 3 Sonnet-IC (16-shot)	0.997	0.710	0.606	48.54	0.876	0.783
en-ja	NLLB-600M-ZS (uncontrolled)	0.500	0.195	0.149	8.05	0.739	0.457
	NLLB-600M-IT	0.912	0.460	0.390	24.44	0.840	0.717
	NLLB-600M-IT + CPO	0.969	0.492	0.400	24.21	0.836	0.698
	NLLB-600M-IT + CPO-SD	0.972	0.501	0.403	22.84	0.837	0.698
	Qwen3 8B-IT	0.927	0.462	0.394	25.00	0.848	0.759
	Qwen3 8B-IT + CPO	0.930	0.459	0.393	25.01	0.847	0.759
	Qwen3 8B-IT + CPO-SD	0.934	0.462	0.398	25.13	0.851	0.767
	Claude 3 Sonnet-ZS	0.825	0.394	0.335	22.65	0.881	0.862
	Claude 3 Sonnet-IC (2-shot)	0.891	0.452	0.362	23.04	0.880	0.864
	Claude 3 Sonnet-IC (8-shot)	0.914	0.484	0.405	25.61	0.887	0.874
	Claude 3 Sonnet-IC (16-shot)	0.925	0.499	0.415	26.21	0.888	0.870
en-nl	NLLB-600M-ZS (uncontrolled)	0.500	0.346	0.481	19.08	0.769	0.534
	NLLB-600M-IT	0.972	0.839	0.816	45.13	0.856	0.742
	NLLB-600M-IT + CPO	1.000	0.865	0.834	44.36	0.851	0.723
	NLLB-600M-IT + CPO-SD	0.999	0.867	0.835	42.30	0.850	0.762
	Qwen3 8B-IT	0.946	0.781	0.774	35.86	0.849	0.776
	Qwen3 8B-IT + CPO	0.949	0.789	0.780	36.44	0.849	0.779
	Qwen3 8B-IT + CPO-SD	0.952	0.792	0.782	36.45	0.850	0.780
	Claude 3 Sonnet-ZS	0.960	0.795	0.799	36.90	0.871	0.859
	Claude 3 Sonnet-IC (2-shot)	0.994	0.831	0.816	37.80	0.874	0.861
	Claude 3 Sonnet-IC (8-shot)	0.999	0.853	0.830	39.99	0.876	0.866
	Claude 3 Sonnet-IC (16-shot)	0.997	0.845	0.825	40.49	0.874	0.860
en-pt	NLLB-600M-ZS (uncontrolled)	0.500	0.323	0.350	28.34	0.793	0.553
	NLLB-600M-IT	0.976	0.807	0.748	52.53	0.869	0.704
	NLLB-600M-IT + CPO	0.995	0.816	0.748	52.31	0.865	0.682
	NLLB-600M-IT + CPO-SD	0.994	0.820	0.745	53.00	0.872	0.748
	Qwen3 8B-IT	0.986	0.793	0.709	50.45	0.866	0.737
	Qwen3 8B-IT + CPO	0.985	0.789	0.707	50.56	0.866	0.735
	Qwen3 8B-IT + CPO-SD	0.984	0.792	0.709	50.63	0.867	0.739
	Claude 3 Sonnet-ZS	0.561	0.397	0.372	33.98	0.856	0.842
	Claude 3 Sonnet-IC (2-shot)	0.912	0.681	0.614	40.84	0.867	0.833
	Claude 3 Sonnet-IC (8-shot)	0.992	0.804	0.731	47.86	0.874	0.818
	Claude 3 Sonnet-IC (16-shot)	0.994	0.816	0.740	48.69	0.875	0.813

Table 15: CoCoA-MT en2all results per language pair.

Lang. Pair	Model	M_{Acc}	$\mathrm{M}_{Acc\text{-}Strict}$	T_{recall}	BLEU	COMET	KIWI-XXL
de-es	NLLB-600M-ZS (uncontrolled)	0.500	0.197	0.244	10.16	0.583	0.226
	NLLB-600M-IT	0.970	0.716	0.602	33.23	0.802	0.629
	NLLB-600M-IT + CPO	0.995	0.727	0.609	32.52	0.798	0.618
	NLLB-600M-IT + CPO-SD	0.974	0.721	0.611	32.88	0.803	0.649
de-fr	NLLB-600M-ZS (uncontrolled)	0.500	0.235	0.216	10.12	0.548	0.174
	NLLB-600M-IT	0.970	0.759	0.648	30.89	0.769	0.560
	NLLB-600M-IT + CPO	0.992	0.763	0.647	29.43	0.760	0.527
	NLLB-600M-IT + CPO-SD	0.972	0.761	0.656	30.25	0.772	0.588
de-hi	NLLB-600M-ZS (uncontrolled)	0.500	0.111	0.217	9.02	0.515	0.255
	NLLB-600M-IT	0.969	0.294	0.771	26.63	0.735	0.628
	NLLB-600M-IT + CPO	0.983	0.297	0.766	25.38	0.730	0.612
	NLLB-600M-IT + CPO-SD	0.976	0.298	0.777	26.77	0.735	0.654
de-it	NLLB-600M-ZS (uncontrolled)	0.500	0.129	0.131	7.64	0.580	0.173
	NLLB-600M-IT	0.966	0.650	0.554	27.38	0.801	0.500
	NLLB-600M-IT + CPO	0.988	0.667	0.563	26.76	0.791	0.477
	NLLB-600M-IT + CPO-SD	0.983	0.666	0.569	26.10	0.803	0.568
de-ja	NLLB-600M-ZS (uncontrolled)	0.500	0.156	0.103	3.60	0.570	0.172
	NLLB-600M-IT	0.943	0.458	0.408	18.88	0.793	0.639
	NLLB-600M-IT + CPO	0.943	0.463	0.401	18.86	0.788	0.634
	NLLB-600M-IT + CPO-SD	0.951	0.466	0.401	19.19	0.792	0.652
de-nl	NLLB-600M-ZS (uncontrolled)	0.500	0.252	0.335	10.13	0.590	0.267
	NLLB-600M-IT	0.906	0.750	0.710	27.29	0.800	0.635
	NLLB-600M-IT + CPO	0.904	0.758	0.695	25.26	0.790	0.599
	NLLB-600M-IT + CPO-SD	0.946	0.788	0.734	27.54	0.808	0.692
de-pt	NLLB-600M-ZS (uncontrolled)	0.500	0.194	0.205	6.86	0.585	0.217
	NLLB-600M-IT	0.965	0.742	0.639	25.59	0.780	0.537
	NLLB-600M-IT + CPO	1.000	0.776	0.637	25.42	0.772	0.510
	NLLB-600M-IT + CPO-SD	0.989	0.776	0.665	27.22	0.787	0.582

Table 16: CoCoA-MT de2all results per language pair.

Lang. Pair	Model	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	BLEU	COMET	Kiwi-XXL
es-de	NLLB-600M-ZS (uncontrolled)	0.500	0.201	0.160	12.32	0.637	0.334
	NLLB-600M-IT	0.981	0.745	0.650	30.18	0.769	0.535
	NLLB-600M-IT + CPO	0.996	0.759	0.662	29.60	0.765	0.511
	NLLB-600M-IT + CPO-SD	0.996	0.759	0.663	31.21	0.773	0.575
es-fr	NLLB-600M-ZS (uncontrolled)	0.500	0.232	0.221	16.46	0.660	0.330
	NLLB-600M-IT	0.982	0.776	0.649	32.76	0.788	0.533
	NLLB-600M-IT + CPO	0.995	0.782	0.646	31.78	0.784	0.517
	NLLB-600M-IT + CPO-SD	0.988	0.788	0.658	31.84	0.787	0.562
es-hi	NLLB-600M-ZS (uncontrolled)	0.500	0.114	0.222	12.63	0.606	0.345
	NLLB-600M-IT	0.962	0.280	0.733	26.79	0.745	0.576
	NLLB-600M-IT + CPO	0.969	0.294	0.740	26.23	0.741	0.568
	NLLB-600M-IT + CPO-SD	0.970	0.284	0.736	26.96	0.746	0.590
es-it	NLLB-600M-ZS (uncontrolled)	0.500	0.204	0.198	17.48	0.726	0.430
	NLLB-600M-IT	0.957	0.637	0.542	31.80	0.822	0.545
	NLLB-600M-IT + CPO	0.982	0.657	0.547	30.67	0.814	0.519
	NLLB-600M-IT + CPO-SD	0.970	0.656	0.553	31.87	0.824	0.603
es-ja	NLLB-600M-ZS (uncontrolled)	0.500	0.169	0.149	7.35	0.679	0.360
	NLLB-600M-IT	0.938	0.457	0.414	18.27	0.805	0.631
	NLLB-600M-IT + CPO	0.944	0.462	0.397	17.89	0.794	0.615
	NLLB-600M-IT + CPO-SD	0.943	0.462	0.406	17.56	0.805	0.634
es-nl	NLLB-600M-ZS (uncontrolled)	0.500	0.298	0.417	11.32	0.664	0.361
	NLLB-600M-IT	0.936	0.761	0.713	21.43	0.771	0.547
	NLLB-600M-IT + CPO	0.924	0.741	0.711	21.43	0.771	0.538
	NLLB-600M-IT + CPO-SD	0.910	0.744	0.713	22.10	0.775	0.588
es-pt	NLLB-600M-ZS (uncontrolled)	0.500	0.257	0.237	16.06	0.728	0.388
	NLLB-600M-IT	0.977	0.759	0.630	31.32	0.828	0.509
	NLLB-600M-IT + CPO	0.990	0.759	0.630	30.53	0.824	0.485
	NLLB-600M-IT + CPO-SD	0.984	0.784	0.637	32.33	0.831	0.602

Table 17: CoCoA-MT es2all results per language pair.

Lang. Pair	Model	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	BLEU	COMET	Kiwi-XXL
fr-de	NLLB-600M-ZS (uncontrolled)	0.500	0.255	0.212	12.69	0.665	0.398
	NLLB-600M-IT	0.990	0.756	0.641	26.24	0.764	0.475
	NLLB-600M-IT + CPO	0.996	0.746	0.634	25.11	0.756	0.447
	NLLB-600M-IT + CPO-SD	0.994	0.754	0.639	26.41	0.767	0.518
fr-es	NLLB-600M-ZS (uncontrolled)	0.500	0.277	0.297	18.63	0.748	0.501
	NLLB-600M-IT	0.984	0.705	0.573	28.84	0.809	0.555
	NLLB-600M-IT + CPO	0.994	0.719	0.52	28.28	0.807	0.539
	NLLB-600M-IT + CPO-SD	0.986	0.720	0.590	29.97	0.815	0.592
fr-hi	NLLB-600M-ZS (uncontrolled)	0.500	0.144	0.312	13.29	0.636	0.386
	NLLB-600M-IT	0.980	0.299	0.766	22.75	0.730	0.502
	NLLB-600M-IT + CPO	0.990	0.305	0.766	22.08	0.727	0.489
	NLLB-600M-IT + CPO-SD	0.986	0.300	0.773	23.19	0.732	0.510
fr-it	NLLB-600M-ZS (uncontrolled)	0.500	0.189	0.192	15.92	0.744	0.410
	NLLB-600M-IT	0.949	0.656	0.540	27.25	0.820	0.425
	NLLB-600M-IT + CPO	0.991	0.669	.539	26.58	0.808	0.401
	NLLB-600M-IT + CPO-SD	0.954	0.664	0.554	23.20	0.817	0.495
fr-ja	NLLB-600M-ZS (uncontrolled)	0.500	0.156	0.127	6.53	0.714	0.251
	NLLB-600M-IT	0.880	0.394	0.358	17.20	0.811	0.532
	NLLB-600M-IT + CPO	0.958	0.459	0.387	17.26	0.803	0.512
	NLLB-600M-IT + CPO-SD	0.913	0.432	0.371	17.83	0.814	0.559
fr-nl	NLLB-600M-ZS (uncontrolled)	0.500	0.297	0.440	11.28	0.696	0.328
	NLLB-600M-IT	0.948	0.754	0.740	20.62	0.793	0.491
	NLLB-600M-IT + CPO	0.974	0.764	0.754	21.43	0.788	0.465
	NLLB-600M-IT + CPO-SD	0.972	0.773	0.766	21.71	0.795	0.536
fr-pt	NLLB-600M-ZS (uncontrolled)	0.500	0.264	0.252	15.53	0.725	0.384
	NLLB-600M-IT	0.962	0.696	0.581	23.32	0.804	0.480
	NLLB-600M-IT + CPO	0.997	0.723	0.603	22.58	0.789	0.441
	NLLB-600M-IT + CPO-SD	0.975	0.721	0.602	22.10	0.803	0.528

Table 18: CoCoA-MT fr2all results per language pair.

Lang. Pair	Model	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	BLEU	COMET	KIWI-XXL
hi-de	NLLB-600M-ZS (uncontrolled)	0.500	0.298	0.252	15.28	0.744	0.581
	NLLB-600M-IT	0.987	0.749	0.625	26.97	0.794	0.671
	NLLB-600M-IT + CPO	1.000	0.740	0.619	26.09	0.786	0.657
	NLLB-600M-IT + CPO-SD	0.998	0.742	0.622	25.08	0.788	0.678
hi-es	NLLB-600M-ZS (uncontrolled)	0.500	0.302	0.331	20.93	0.801	0.645
	NLLB-600M-IT	0.991	0.687	0.575	28.50	0.823	0.703
	NLLB-600M-IT + CPO	0.998	0.694	0.573	28.00	0.819	0.695
	NLLB-600M-IT + CPO-SD	0.988	0.690	0.583	26.45	0.823	0.710
hi-fr	NLLB-600M-ZS (uncontrolled)	0.500	0.328	0.309	18.40	0.746	0.568
	NLLB-600M-IT	0.981	0.760	0.630	27.65	0.795	0.625
	NLLB-600M-IT + CPO	1.000	0.771	0.629	26.24	789	0.599
	NLLB-600M-IT + CPO-SD	0.990	0.770	0.638	27.32	0.796	0.632
hi-it	NLLB-600M-ZS (uncontrolled)	0.500	0.233	0.218	16.46	0.778	0.555
	NLLB-600M-IT	0.976	0.632	0.496	23.07	0.810	0.592
	NLLB-600M-IT + CPO	0.994	0.633	0.493	23.00	0.801	0.570
	NLLB-600M-IT + CPO-SD	0.938	0.591	0.467	22.10	0.813	0.605
hi-ja	NLLB-600M-ZS (uncontrolled)	0.500	0.181	0.180	8.42	0.748	0.371
	NLLB-600M-IT	0.920	0.431	0.358	18.66	0.820	0.637
	NLLB-600M-IT + CPO	0.939	0.469	0.362	17.55	0.802	0.578
	NLLB-600M-IT + CPO-SD	0.883	0.411	0.381	17.65	0.830	0.681
hi-nl	NLLB-600M-ZS (uncontrolled)	0.500	0.376	0.463	14.81	0.769	0.533
	NLLB-600M-IT	0.962	0.827	0.758	21.89	0.807	0.605
	NLLB-600M-IT + CPO	1.000	0.856	0.780	20.46	0.802	0.600
	NLLB-600M-IT + CPO-SD	0.967	0.839	0.757	19.17	0.804	0.634
hi-pt	NLLB-600M-ZS (uncontrolled)	0.500	0.312	0.297	16.96	0.785	0.563
	NLLB-600M-IT	0.943	0.703	0.554	23.63	0.813	0.580
	NLLB-600M-IT + CPO	1.00	0.733	0.566	22.73	0.808	0.566
	NLLB-600M-IT + CPO-SD	0.972	0.706	0.541	22.69	0.817	0.613

Table 19: CoCoA-MT hi2all results per language pair.

Lang. Pair	Model	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	BLEU	COMET	KIWI-XXL
it-de	NLLB-600M-ZS (uncontrolled)	0.500	0.150	0.117	8.42	0.579	0.250
	NLLB-600M-IT	0.965	0.744	0.619	27.68	0.765	0.510
	NLLB-600M-IT + CPO	0.992	0.761	0.629	27.18	0.754	0.483
	NLLB-600M-IT + CPO-SD	0.989	0.766	0.639	25.59	0.761	0.536
it-es	NLLB-600M-ZS (uncontrolled)	0.500	0.244	0.254	17.47	0.687	0.400
	NLLB-600M-IT	0.958	0.676	0.557	32.35	0.811	0.621
	NLLB-600M-IT + CPO	0.993	0.675	0.550	30.68	0.801	0.585
	NLLB-600M-IT + CPO-SD	0.969	0.699	0.578	33.44	0.812	0.668
it-fr	NLLB-600M-ZS (uncontrolled)	0.500	0.194	0.182	13.61	0.624	0.263
	NLLB-600M-IT	0.984	0.795	0.649	33.04	0.786	0.522
	NLLB-600M-IT + CPO	0.992	0.796	0.656	31.67	0.780	0.493
	NLLB-600M-IT + CPO-SD	0.982	0.793	0.655	33.39	0.790	0.542
it-hi	NLLB-600M-ZS (uncontrolled)	0.500	0.076	0.179	9.77	0.558	0.266
	NLLB-600M-IT	0.969	0.282	0.711	23.69	0.715	0.537
	NLLB-600M-IT + CPO	0.986	0.289	0.722	22.96	0.710	0.527
	NLLB-600M-IT + CPO-SD	0.982	0.278	0.719	23.69	0.714	0.561
it-ja	NLLB-600M-ZS (uncontrolled)	0.500	0.151	0.112	5.08	0.651	0.258
	NLLB-600M-IT	0.847	0.410	0.314	17.38	0.790	0.603
	NLLB-600M-IT + CPO	0.895	0.458	0.322	17.75	0.792	0.601
	NLLB-600M-IT + CPO-SD	0.866	0.431	0.326	17.37	0.797	0.612
it-nl	NLLB-600M-ZS (uncontrolled)	0500	0.175	0.321	6.87	0.571	0.196
	NLLB-600M-IT	0.950	0.805	0.745	24.48	0.762	0.504
	NLLB-600M-IT + CPO	0.985	0.850	0.767	24.33	0.745	0.450
	NLLB-600M-IT + CPO-SD	0.937	0.803	0.736	24.15	0.758	0.528
it-pt	NLLB-600M-ZS (uncontrolled)	0.500	0.208	0.187	14.04	0.654	0.311
	NLLB-600M-IT	0.962	0.692	0.551	31.35	0.800	0.521
	NLLB-600M-IT + CPO	0.990	0.720	0.574	30.83	0.791	0.477
	NLLB-600M-IT + CPO-SD	0.990	0.726	0.580	31.65	0.800	0.580

Table 20: CoCoA-MT it2all results per language pair

Lang. Pair	**Model**	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	**BLEU**	**COMET**	**Kiwi-XXL**
ja-de	NLLB-600M-ZS (uncontrolled)	0.500	0.152	0.116	4.91	0.610	0.409
	NLLB-600M-IT	0.965	0.615	0.478	14.50	0.707	0.514
	NLLB-600M-IT + CPO	1.000	0.631	0.496	13.21	0.695	0.476
	NLLB-600M-IT + CPO-SD	0.974	0.610	0.476	14.41	0.708	.531
ja-es	NLLB-600M-ZS (uncontrolled)	0.500	0.155	0.186	8.62	0.671	0.533
	NLLB-600M-IT	0.941	0.570	0.477	16.76	0.753	0.639
	NLLB-600M-IT + CPO	0.987	0.615	0.519	16.45	0.744	0.607
	NLLB-600M-IT + CPO-SD	0.957	0.567	0.467	17.77	0.757	0.647
ja-fr	NLLB-600M-ZS (uncontrolled)	0.500	0.159	0.167	8.35	0.632	0.369
	NLLB-600M-IT	0.946	0.625	0.464	18.00	0.725	0.458
	NLLB-600M-IT + CPO	0.987	0.654	0.493	17.80	0.716	0.451
	NLLB-600M-IT + CPO-SD	0.968	0.635	0.475	18.67	0.728	0.486
ja-hi	NLLB-600M-ZS (uncontrolled)	0.500	0.080	0.136	4.46	0.541	0.418
	NLLB-600M-IT	0.882	0.275	0.604	15.12	0.687	0.557
	NLLB-600M-IT + CPO	0.918	0.285	0.636	14.98	0.685	0.547
	NLLB-600M-IT + CPO-SD	0.862	0.262	0.602	16.33	0.693	0.581
ja-it	NLLB-600M-ZS (uncontrolled)	0.500	0.122	0.103	6.91	0.655	0.433
	NLLB-600M-IT	0.926	0.508	0.388	15.09	0.750	0.525
	NLLB-600M-IT + CPO	0.938	0.529	0.410	15.46	0.746	0.507
	NLLB-600M-IT + CPO-SD	0.910	0.505	0.390	9.22	0.729	0.508
ja-nl	NLLB-600M-ZS (uncontrolled)	0.500	0.200	0.321	4.19	0.669	0.441
	NLLB-600M-IT	0.973	0.694	0.693	13.86	0.767	0.591
	NLLB-600M-IT + CPO	0.996	0.743	0.716	13.99	0.766	0.576
	NLLB-600M-IT + CPO-SD	0.991	0.722	0.703	14.39	0.769	0.613
ja-pt	NLLB-600M-ZS (uncontrolled)	0.500	0.207	0.181	9.48	0.705	0.507
	NLLB-600M-IT	0.970	0.660	0.496	16.78	0.789	0.601
	NLLB-600M-IT + CPO	0.988	0.683	0.492	16.15	0.786	0.588
	NLLB-600M-IT + CPO-SD	0.949	0.671	0.507	16.84	0.790	0.607

Table 21: CoCoA-MT ja2all results per language pair.

Lang. Pair	**Model**	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	**BLEU**	**COMET**	**Kiwi-XXL**
nl-de	NLLB-600M-ZS (uncontrolled)	0.500	0.150	0.103	10.10	0.594	0.296
	NLLB-600M-IT	0.974	0.694	0.595	25.00	0.755	0.531
	NLLB-600M-IT + CPO	0.983	0.711	0.579	23.12	0.738	0.487
	NLLB-600M-IT + CPO-SD	0.991	0.720	0.621	21.17	0.759	0.589
nl-es	NLLB-600M-ZS (uncontrolled)	0.500	0.143	0.184	7.53	0.606	0.267
	NLLB-600M-IT	0.938	0.681	0.575	23.03	0.762	0.561
	NLLB-600M-IT + CPO	0.983	0.686	0.570	24.61	0.755	0.545
	NLLB-600M-IT + CPO-SD	0.954	0.698	0.583	20.96	0.764	0.583
nl-fr	NLLB-600M-ZS (uncontrolled)	0.500	0.150	0.127	10.69	0.598	0.243
	NLLB-600M-IT	0.919	0.692	0.546	26.11	0.749	0.488
	NLLB-600M-IT + CPO	0.987	0.750	0.586	24.28	0.729	0.422
	NLLB-600M-IT + CPO-SD	0.935	0.707	0.563	26.26	0.752	0.510
nl-hi	NLLB-600M-ZS (uncontrolled)	0.500	0.066	0.169	7.43	0.542	0.265
	NLLB-600M-IT	0.976	0.275	0.722	18.75	0.706	0.512
	NLLB-600M-IT + CPO	0.976	0.290	0.726	17.35	0.701	0.493
	NLLB-600M-IT + CPO-SD	0.976	0.281	0.717	18.71	0.706	0.546
nl-it	NLLB-600M-ZS (uncontrolled)	0.500	0.104	0.107	9.18	0.607	0.225
	NLLB-600M-IT	0.937	0.657	0.532	25.07	0.770	0.472
	NLLB-600M-IT + CPO	0.997	0.671	0.527	24.81	0.755	0.414
	NLLB-600M-IT + CPO-SD	0.936	0.670	0.542	21.39	0.769	0.496
nl-ja	NLLB-600M-ZS (uncontrolled)	0.500	0.143	0.112	5.24	0.647	0.258
	NLLB-600M-IT	0.891	0.366	0.308	18.73	0.804	0.652
	NLLB-600M-IT + CPO	0.935	0.397	0.296	17.40	0.791	0.625
	NLLB-600M-IT + CPO-SD	0.933	0.388	0.325	19.02	0.810	0.674
nl-pt	NLLB-600M-ZS (uncontrolled)	0.500	0.152	0.154	8.90	0.606	0.229
	NLLB-600M-IT	0.975	0.738	0.579	30.16	0.801	0.601
	NLLB-600M-IT + CPO	0.994	0.741	0.576	29.83	0.796	0.581
	NLLB-600M-IT + CPO-SD	0.985	0.750	0.591	30.35	0.804	0.619

Table 22: CoCoA-MT nl2all results per language pair.

Lang. Pair	Model	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	BLEU	COMET	KIWI-XXL
pt-de	NLLB-600M-ZS (uncontrolled)	0.500	0.192	0.136	8.51	0.608	0.296
	NLLB-600M-IT	0.932	0.692	0.643	26.46	0.758	0.502
	NLLB-600M-IT + CPO	0.973	0.707	0.670	26.30	0.750	0.479
	NLLB-600M-IT + CPO-SD	0.970	0.721	0.670	25.19	0.758	0.544
pt-es	NLLB-600M-ZS (uncontrolled)	0.500	0.240	0.252	18.91	0.721	0.458
	NLLB-600M-IT	0.932	0.718	0.588	33.79	0.817	0.574
	NLLB-600M-IT + CPO	0.983	0.754	0.613	32.99	0.810	0.556
	NLLB-600M-IT + CPO-SD	0.933	0.713	0.592	33.98	0.817	0.654
pt-fr	NLLB-600M-ZS (uncontrolled)	0.500	0.200	0.183	13.57	0.635	0.310
	NLLB-600M-IT	0.942	0.694	0.604	28.35	0.761	0.483
	NLLB-600M-IT + CPO	0.994	0.729	0.625	28.57	0.756	0.453
	NLLB-600M-IT + CPO-SD	0.932	0.706	0.626	28.98	0.765	0.526
pt-hi	NLLB-600M-ZS (uncontrolled)	0.500	0.055	0.177	10.70	0.584	0.322
	NLLB-600M-IT	0.938	0.194	0.666	21.07	0.714	0.517
	NLLB-600M-IT + CPO	1.000	0.206	0.680	20.07	0.712	0.510
	NLLB-600M-IT + CPO-SD	0.952	0.194	0.670	20.87	0.713	0.548
pt-it	NLLB-600M-ZS (uncontrolled)	0.500	0.194	0.185	16.42	0.679	0.311
	NLLB-600M-IT	0.982	0.730	0.610	33.54	0.803	0.470
	NLLB-600M-IT + CPO	0.991	0.735	0.601	30.64	0.794	0.449
	NLLB-600M-IT + CPO-SD	0.984	0.746	0.618	27.85	0.797	0.483
pt-ja	NLLB-600M-ZS (uncontrolled)	0.500	0.171	0.120	5.54	0.688	0.281
	NLLB-600M-IT	0.813	0.304	0.370	20.25	0.808	0.619
	NLLB-600M-IT + CPO	0.920	0.332	0.379	29.28	0.804	0.605
	NLLB-600M-IT + CPO-SD	0.868	0.332	0.384	19.67	0.815	0.631
pt-nl	NLLB-600M-ZS (uncontrolled)	0.500	0.270	0.401	8.46	0.633	0.283
	NLLB-600M-IT	0.953	0.820	0.781	29.70	0.800	0.573
	NLLB-600M-IT + CPO	0.948	0.807	0.776	29.00	0.789	0.546
	NLLB-600M-IT + CPO-SD	0.982	0.836	0.803	29.74	0.796	0.598

Table 23: CoCoA-MT pt2all results per language pair.

Lang. Pair	Model	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	BLEU	COMET	KIWI-XXL
en-ar	NLLB-600M-ZS (uncontrolled)	0.500	0.328	0.299	18.05	0.779	0.596
	NLLB-600M-IT	0.970	0.670	0.446	21.85	0.792	0.541
	NLLB-600M-IT + CPO	0.981	0.696	0.454	22.03	0.792	0.543
	NLLB-600M-IT + CPO-SD	0.981	0.698	0.459	21.83	0.794	0.581
	Qwen3 8B-IT	0.962	0.605	0.386	16.95	0.760	0.539
	Qwen3 8B-IT + CPO	0.955	0.607	0.386	16.95	0.760	0.539
	Qwen3 8B-IT + CPO-SD	0.963	0.632	0.409	17.54	0.791	0.601
	Claude 3 Sonnet-ZS	0.978	0.606	0.411	18.39	0.758	0.558
	Claude 3 Sonnet-IC (2-shot)	0.971	0.698	0.526	26.82	0.843	0.761
	Claude 3 Sonnet-IC (8-shot)	0.978	0.714	0.536	27.06	0.846	0.761
	Claude 3 Sonnet-IC (16-shot)	0.973	0.700	0.529	27.13	0.844	0.762
en-de	NLLB-600M-ZS (uncontrolled)	0.500	0.389	0.450	34.83	0.810	0.619
	NLLB-600M-IT	0.982	0.790	0.710	38.81	0.823	0.558
	NLLB-600M-IT + CPO	0.988	0.798	0.714	38.14	0.821	0.550
	NLLB-600M-IT + CPO-SD	0.986	0.803	.723	39.81	0.828	0.623
	Qwen3 8B-IT	0.985	0.756	0.691	36.17	0.825	0.626
	Qwen3 8B-IT + CPO	0.985	0.759	0.693	36.33	0.827	0.629
	Qwen3 8B-IT + CPO-SD	0.991	0.777	0.706	38.13	0.842	0.660
	Claude 3 Sonnet-ZS	0.978	0.720	0.638	31.15	0.721	0.473
	Claude 3 Sonnet-IC (2-shot)	0.969	0.750	0.741	45.88	0.865	0.721
	Claude 3 Sonnet-IC (8-shot)	0.972	0.771	0.752	45.61	0.866	0.727
	Claude 3 Sonnet-IC (16-shot)	0.957	0.754	0.741	44.12	0.856	0.717
en-es	NLLB-600M-ZS (uncontrolled)	0.500	0.343	0.338	48.06	0.836	0.698
	NLLB-600M-IT	0.965	0.746	0.646	52.71	0.851	0.630
	NLLB-600M-IT + CPO	0.981	0.772	0.664	52.60	0.849	0.621
	NLLB-600M-IT + CPO-SD	.984	0.793	0.683	52.87	0.850	0.662
	Qwen3 8B-IT	0.986	.767	0.653	49.28	0.846	0.682
	Qwen3 8B-IT + CPO	0.986	0.768	0.652	49.31	0.846	0.682
	Qwen3 8B-IT + CPO-SD	0.987	0.784	0.665	50.36	0.851	0.691
	Claude 3 Sonnet-ZS	0.968	0.682	0.551	40.08	0.770	0.518
	Claude 3 Sonnet-IC (2-shot)	0.966	0.752	0.666	53.77	0.868	0.731
	Claude 3 Sonnet-IC (8-shot)	0.964	0.760	0.677	52.97	0.867	0.735
	Claude 3 Sonnet-IC (16-shot)	0.962	0.750	0.672	52.19	0.863	0.734
en-fr	NLLB-600M-ZS (uncontrolled)	0.500	0.401	0.439	35.97	0.812	0.620
	NLLB-600M-IT	0.983	0.823	0.682	40.17	0.823	0.532
	NLLB-600M-IT + CPO	0.989	0.840	0.693	39.81	0.823	0.525
	NLLB-600M-IT + CPO-SD	0.991	0.840	0.707	41.30	0.828	0.605
	Qwen3 8B-IT	0.984	0.780	0.648	36.35	0.822	0.616
	Qwen3 8B-IT + CPO	0.984	0.783	0.649	36.42	0.823	0.618
	Qwen3 8B-IT + CPO-SD	0.990	0.822	0.679	38.51	0.832	0.633
	Claude 3 Sonnet-ZS	0.964	0.756	0.595	30.33	0.728	0.414
	Claude 3 Sonnet-IC (2-shot)	0.985	0.807	0.723	42.46	0.835	0.666
	Claude 3 Sonnet-IC (8-shot)	0.991	0.808	0.733	38.74	0.810	0.652
	Claude 3 Sonnet-IC (16-shot)	0.990	0.771	0.721	34.09	0.774	0.612

Table 24: MT-GenEval en2all results per language pair.

Lang. Pair	Model	$\mathbf{M}_{Acc}$	$\mathbf{M}_{Acc\text{-}Strict}$	$\mathbf{T}_{recall}$	BLEU	COMET	KIWI-XXL
en-hi	NLLB-600M-ZS (uncontrolled)	0.500	0.143	0.292	28.23	0.766	0.645
	NLLB-600M-IT	0.715	0.237	0.550	28.02	0.773	0.603
	NLLB-600M-IT + CPO	0.786	0.249	0.570	28.29	0.771	0.595
	NLLB-600M-IT + CPO-SD	0.727	0.236	0.554	28.04	0.775	0.608
	Qwen3 8B-IT	0.712	0.188	0.491	19.32	0.718	0.497
	Qwen3 8B-IT + CPO	0.714	0.188	0.495	19.38	0.718	0.496
	Qwen3 8B-IT + CPO-SD	0.736	0.194	0.503	19.82	0.719	0.511
	Claude 3 Sonnet-ZS	0.690	0.160	0.391	22.08	0.712	0.525
	Claude 3 Sonnet-IC (2-shot)	0.666	0.186	0.459	27.89	0.792	0.703
	Claude 3 Sonnet-IC (8-shot)	0.680	0.194	0.488	28.56	0.793	0.698
	Claude 3 Sonnet-IC (16-shot)	0.675	0.197	0.481	27.08	0.781	0.681
en-it	NLLB-600M-ZS (uncontrolled)	0.500	0.350	0.305	35.62	0.837	0.689
	NLLB-600M-IT	0.965	0.677	0.525	37.41	0.843	0.594
	NLLB-600M-IT + CPO	0.975	0.693	0.531	36.97	0.841	0.584
	NLLB-600M-IT + CPO-SD	0.987	0.716	0.547	37.60	0.846	0.653
	Qwen3 8B-IT	0.937	0.644	0.482	34.28	0.846	0.679
	Qwen3 8B-IT + CPO	0.936	0.643	0.482	34.35	0.846	0.681
	Qwen3 8B-IT + CPO-SD	0.953	0.682	0.499	36.05	0.850	0.692
	Claude 3 Sonnet-ZS	0.959	0.626	0.471	31.23	0.776	0.500
	Claude 3 Sonnet-IC (2-shot)	0.956	0.685	0.408	40.81	0.865	0.740
	Claude 3 Sonnet-IC (8-shot)	0.956	0.703	0.575	32.30	0.815	0.710
	Claude 3 Sonnet-IC (16-shot)	0.964	0.704	0.581	29.02	0.774	0.671
en-pt	NLLB-600M-ZS (uncontrolled)	0.500	0.377	0.376	44.57	0.849	0.701
	NLLB-600M-IT	0.970	0.757	0.661	47.62	0.858	0.604
	NLLB-600M-IT + CPO	0.983	0.783	0.677	47.70	0.858	0.601
	NLLB-600M-IT + CPO-SD	0.981	0.774	0.672	48.51	0.858	0.666
	Qwen3 8B-IT	0.983	0.778	0.650	42.75	0.857	0.681
	Qwen3 8B-IT + CPO	0.983	0.776	0.650	42.72	0.857	0.681
	Qwen3 8B-IT + CPO-SD	0.981	0.781	0.661	43.63	0.863	0.698
	Claude 3 Sonnet-ZS	0.985	0.745	0.625	39.16	0.801	0.554
	Claude 3 Sonnet-IC (2-shot)	0.985	0.807	0.720	52.06	0.883	0.748
	Claude 3 Sonnet-IC (8-shot)	0.977	0.791	0.709	50.67	0.875	0.737
	Claude 3 Sonnet-IC (16-shot)	0.968	0.770	0.694	46.60	0.856	0.718
en-ru	NLLB-600M-ZS (uncontrolled)	0.500	0.348	0.313	29.26	0.823	0.604
	NLLB-600M-IT	0.968	0.685	0.502	33.60	0.837	0.563
	NLLB-600M-IT + CPO	0.985	0.705	0.519	32.99	0.837	0.565
	NLLB-600M-IT + CPO-SD	0.980	0.694	0.518	33.74	0.839	0.601
	Qwen3 8B-IT	0.982	0.681	0.481	29.01	0.843	0.620
	Qwen3 8B-IT + CPO	0.982	0.680	0.482	29.01	0.843	0.620
	Qwen3 8B-IT + CPO-SD	0.984	0.702	0.505	30.04	0.851	0.636
	Claude 3 Sonnet-ZS	0.959	0.599	0.409	21.25	0.738	0.454
	Claude 3 Sonnet-IC (2-shot)	0.961	0.696	0.574	38.05	0.882	0.748
	Claude 3 Sonnet-IC (8-shot)	0.979	0.713	0.582	36.66	0.874	0.735
	Claude 3 Sonnet-IC (16-shot)	0.968	0.701	0.576	34.89	0.859	0.723

Table 25: MT-GenEval en2all results per language pair.

Lang. Pair	Model	M_{Acc}	$\mathrm{M}_{Acc\text{-}Strict}$	T_{recall}	BLEU	COMET	KIWI-XXL
da-es	NLLB-600M-ZS (uncontrolled)	0.500	0.171	0.184	15.72	0.634	0.385
	NLLB-600M-IT (orig. FAME-MT)	0.776	0.402	0.356	34.01	0.833	0.843
	NLLB-600M-IT	0.861	0.490	0.440	35.60	0.841	0.858
	NLLB-600M-IT + CPO	0.874	0.486	0.445	34.77	0.839	0.854
	NLLB-600M-IT + CPO-SD	0.870	0.500	0.442	35.37	0.843	0.862
	Qwen3 8B-IT	0.868	0.490	0.484	36.94	0.846	0.869
	Qwen3 8B-IT + CPO	0.868	0.490	0.486	37.09	0.847	0.868
	Qwen3 8B-IT + CPO-SD	0.868	0.490	0.487	37.15	0.847	0.870
	EuroLLM 9B-IT	0.839	0.490	0.451	36.85	0.858	0.903
	EuroLLM 9B-IT + CPO	0.833	0.490	0.455	37.10	0.859	0.903
	EuroLLM 9B-IT + CPO-SD	0.839	0.490	0.451	36.89	0.858	0.903
de-fr	NLLB-600M-ZS (uncontrolled)	0.500	0.184	0.156	9.82	0.598	0.274
	NLLB-600M-IT (orig. FAME-MT)	0.889	0.543	0.441	32.17	0.812	0.770
	NLLB-600M-IT	0.931	0.611	0.490	34.49	0.817	0.791
	NLLB-600M-IT + CPO	0.948	0.640	0.498	33.79	0.810	0.769
	NLLB-600M-IT + CPO-SD	0.932	0.616	0.496	34.76	0.819	0.792
	Qwen3 8B-IT	0.904	0.601	0.485	35.59	0.832	0.837
	Qwen3 8B-IT + CPO	0.904	0.601	0.485	35.65	0.832	0.840
	Qwen3 8B-IT + CPO-SD	0.911	0.606	0.489	35.44	0.832	0.835
	EuroLLM 9B-IT	0.949	0.553	0.439	31.81	0.818	0.827
	EuroLLM 9B-IT + CPO	0.942	0.572	0.453	30.11	0.825	0.847
	EuroLLM 9B-IT + CPO-SD	0.934	0.567	0.453	30.50	0.827	0.852
en-de	NLLB-600M-ZS (uncontrolled)	0.500	0.267	0.262	28.40	0.842	0.793
	NLLB-600M-IT (orig. FAME-MT)	0.878	0.460	0.433	33.17	0.850	0.802
	NLLB-600M-IT	0.919	0.507	0.471	34.42	0.849	0.800
	NLLB-600M-IT + CPO	0.910	0.440	0.405	28.12	0.829	0.734
	NLLB-600M-IT + CPO-SD	0.927	0.515	0.475	34.34	0.851	0.802
	Qwen3 8B-IT	0.938	0.544	0.522	36.87	0.874	0.857
	Qwen3 8B-IT + CPO	0.937	0.540	0.522	36.83	0.874	0.857
	Qwen3 8B-IT + CPO-SD	0.938	0.544	0.522	36.87	0.874	0.858
	EuroLLM 9B-IT	0.907	0.507	0.482	36.57	0.885	0.878
	EuroLLM 9B-IT + CPO	0.907	0.507	0.482	36.70	0.885	0.878
	EuroLLM 9B-IT + CPO-SD	0.906	0.504	0.477	36.76	0.885	0.878
es-en	NLLB-600M-ZS (uncontrolled)	0.500	0.280	0.285	29.47	0.798	0.693
	NLLB-600M-IT (orig. FAME-MT)	0.514	0.317	0.311	43.17	0.870	0.892
	NLLB-600M-IT	0.689	0.368	0.372	44.25	0.869	0.879
	NLLB-600M-IT + CPO	0.521	0.324	0.294	30.86	0.798	0.726
	NLLB-600M-IT + CPO-SD	0.691	0.364	0.373	44.38	0.869	0.878
	Qwen3 8B-IT	0.763	0.445	0.445	48.47	0.881	0.897
	Qwen3 8B-IT + CPO	0.763	0.445	0.444	48.30	0.880	0.896
	Qwen3 8B-IT + CPO-SD	0.763	0.445	0.444	48.33	0.880	0.896
	EuroLLM 9B-IT	0.666	0.378	0.404	49.99	0.883	0.908
	EuroLLM 9B-IT + CPO	0.695	0.385	0.413	50.34	0.884	0.909
	EuroLLM 9B-IT + CPO-SD	0.691	0.385	0.416	50.09	0.884	0.908
it-nl	NLLB-600M-ZS (uncontrolled)	0.500	0.202	0.238	12.98	0.701	0.433
	NLLB-600M-IT (orig. FAME-MT)	0.915	0.618	0.583	31.59	0.850	0.776
	NLLB-600M-IT	0.966	0.683	0.667	33.71	0.845	0.759
	NLLB-600M-IT + CPO	0.965	0.683	0.672	32.51	0.841	0.747
	NLLB-600M-IT + CPO-SD	0.966	0.683	0.666	33.43	0.846	0.763
	Qwen3 8B-IT	0.965	0.670	0.652	36.41	0.854	0.806
	Qwen3 8B-IT + CPO	0.965	0.670	0.652	36.44	0.854	0.807
	Qwen3 8B-IT + CPO-SD	0.965	0.670	0.652	36.43	0.854	0.807
	EuroLLM 9B-IT	0.952	0.615	0.606	33.73	0.867	0.856
	EuroLLM 9B-IT + CPO	0.952	0.615	0.606	33.71	0.867	0.856
	EuroLLM 9B-IT + CPO-SD	0.929	0.587	0.586	34.24	0.870	0.867
pl-it	NLLB-600M-ZS (uncontrolled)	0.500	0.180	0.137	8.26	0.647	0.313
	NLLB-600M-IT (orig. FAME-MT)	0.682	0.360	0.291	28.96	0.821	0.722
	NLLB-600M-IT	0.705	0.388	0.319	28.60	0.822	0.723
	NLLB-600M-IT + CPO	0.806	0.430	0.340	27.30	0.813	0.696
	NLLB-600M-IT + CPO-SD	0.713	0.395	0.322	28.44	0.822	0.718
	Qwen3 8B-IT	0.772	0.423	0.353	29.22	0.835	0.773
	Qwen3 8B-IT + CPO	0.777	0.423	0.355	29.25	0.835	0.773
	Qwen3 8B-IT + CPO-SD	0.776	0.430	0.356	29.32	0.836	0.775
	EuroLLM 9B-IT	0.653	0.342	0.296	26.46	0.826	0.781
	EuroLLM 9B-IT + CPO	0.711	0.374	0.310	27.30	0.835	0.787
	EuroLLM 9B-IT + CPO-SD	0.642	0.353	0.301	27.44	0.832	0.793
ru-pt	NLLB-600M-ZS (uncontrolled)	0.500	0.275	0.313	24.57	0.797	0.675
	NLLB-600M-IT (orig. FAME-MT)	0.844	0.377	0.323	26.55	0.819	0.715
	NLLB-600M-IT	0.816	0.403	0.39	28.10	0.814	0.687
	NLLB-600M-IT + CPO	0.914	0.424	0.362	25.13	0.787	0.637
	NLLB-600M-IT + CPO-SD	0.848	0.429	0.367	24.89	0.810	0.678
	Qwen3 8B-IT	0.813	0.420	0.372	26.36	0.826	0.764
	Qwen3 8B-IT + CPO	0.815	0.424	0.375	26.62	0.826	0.764
	Qwen3 8B-IT + CPO-SD	0.810	0.428	0.380	26.54	0.828	0.769
	EuroLLM 9B-IT	0.812	0.390	0.350	30.69	0.833	0.780
	EuroLLM 9B-IT + CPO	0.838	0.398	0.352	30.63	0.832	0.775
	EuroLLM 9B-IT + CPO-SD	0.815	0.394	0.352	30.82	0.834	0.781

Table 26: All PREF-FAME-MT results per language pair.