Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC WMT25 submissions in Retrospect

What Worked, What Didn’t, and What’s Next

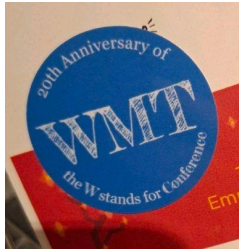
Javier García Gilabert

31 March, 2025

Virtual Talk @ Meta

What is WMT?

WMT (Conference on Machine Translation) is the main annual international conference and shared-task event focused on machine translation.



In practice, “WMT” = the big MT competition + conference where research groups submit systems to be evaluated on standard datasets for many language pairs. This year it included:

TRANSLATION TASKS: GENERAL MT (NEWS) • INDIC MT • TERMINOLOGY • CREOLE MT • MODEL COMPRESSION
EVALUATION TASKS: MT TEST SUITES • (UNIFIED) MT EVALUATION
OTHER TASKS: OPEN DATA
MULTILINGUAL TASKS: MULTILINGUAL INSTRUCTION • LIMITED RESOURCES SLAVIC LLM

Document-level translation
Language pairs: 31 - **we participated in 21 pairs**
Different domains:

- News
- Speech
- Social
- Literary


From SALAMANDRA to SALAMANDRATA:
BSC Submission for WMT25 General Machine Translation Shared Task

Javier Garcia Gilabert[1] Xixian Liao[1] Severino Da Dalt[1] Ella Bohman[1]
Audrey Mash[1] Francesca De Luca Fornaciari[1] Irene Baucells[1] Joan Llop[1]
Miguel Claramunt Argote[1] Carlos Escolano[1,2] Maite Melero[1]
[1]Barcelona Supercomputing Center
[2]Universitat Politècnica de Catalunya

Abstract

In this paper, we present the SALAMANDRATA family of models, an improved iteration of SALAMANDRA LLMs (Gonzalez-Agirre et al., 2025) specifically trained to achieve strong performance in translation-related tasks for 38 European languages. SALAMANDRATA comes in two scales: 2B and 7B parameters. For both versions, we applied the same training recipe with a first step of continual pre-training on parallel data, and a second step of supervised fine-tuning on high-quality instructions.

The BSC submission to the WMT25 General Machine Translation shared task is based on the 7B variant of SALAMANDRATA. We first adapted the model vocabulary to support the additional non-European languages included in the task. This was followed by a second phase of continual pre-training and supervised fine-tuning, carefully designed to optimize performance across all translation directions for this year's shared task. For decoding, we employed two quality-aware strategies: Minimum Bayes Risk Decoding and Tuned Re-ranking using COMET and COMET-KIWI respectively.

We publicly release both the 2B and 7B versions of SALAMANDRATA, along with the newer SALAMANDRATA-V2 model, on Hugging Face[1].

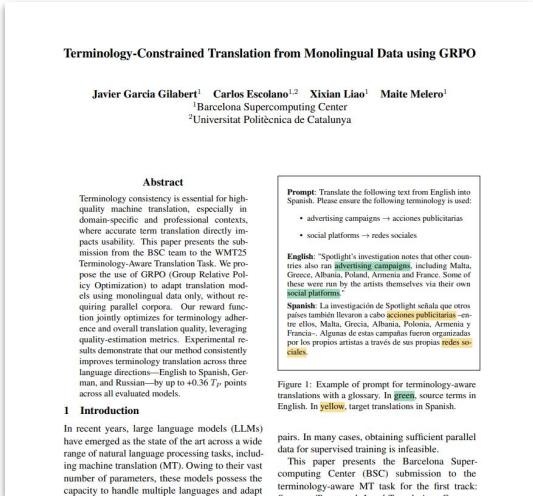
the translation capabilities of a pre-trained LLM model and better align it with human translations (Zhang et al., 2023; Alves et al., 2024; Xu et al., 2024).

One such approach is continual pre-training using a combination of monolingual and parallel corpora followed by supervised fine-tuning (Alves et al., 2024). However, most previous approaches have predominantly relied on English-centric parallel corpora. This has been shown to bias the models towards English-centric latent representations (Zhang et al., 2025) which has been attributed to the language distribution used in the training corpora (Zhong et al., 2024). It is well known that training with only a single bridge language can negatively impact translation performance across zero-shot language pairs, due to limited cross-lingual transfer (Arivazhagan et al., 2019). Unlike previous works, in this paper we rely on parallel corpora only for the continual pre-training stage pivoting on three bridge languages.

When working with pre-trained language models on languages not covered by their original tokenizer, a highly effective solution involves replacing the existing tokenizer with a more comprehensive one that supports such languages. For the newly introduced tokens, embeddings must be initialized. In our work, these new embeddings were

Terminology translation
Language pairs:

- English → German
- English → Russian
- English → Spanish


Terminology-Constrained Translation from Monolingual Data using GRPO

Javier Garcia Gilabert[1] Carlos Escolano[1,2] Xixian Liao[1] Maite Melero[1]
[1]Barcelona Supercomputing Center
[2]Universitat Politècnica de Catalunya

Abstract

Terminology consistency is essential for high-quality machine translation, especially in domain-specific and professional contexts, where accurate term translation directly impacts usability. This paper presents the submission from the BSC team to the WMT25 Terminology-Aware Translation Task. We propose the use of GRPO (Group Relative Policy Optimization) to adapt translation models using monolingual data only, without requiring parallel corpora. Our reward function jointly optimizes for terminology adherence and overall translation quality, leveraging quality-estimation metrics. Experimental results demonstrate that our method consistently improves terminology translation across three language directions—English to Spanish, German, and Russian—by up to +0.36 T_P points across all evaluated models.

Prompt: Translate the following text from English into Spanish. Please ensure the following terminology is used:

- advertising campaigns → acciones publicitarias
- social platforms → redes sociales

English: "Spotlight's investigation notes that other countries also ran advertising campaigns, including Malta, Greece, Albania, Poland, Armenia and France. Some of these were run by the artists themselves via their own social platforms."

Spanish: La investigación de Spotlight señala que otros países también llevaron a cabo acciones publicitarias –entre ellos, Malta, Grecia, Albania, Polonia, Armenia y Francia–. Algunas de estas campañas fueron organizadas por los propios artistas a través de sus propias redes sociales.

Figure 1: Example of prompt for terminology-aware translations with a glossary. In green, source terms in English. In yellow, target translations in Spanish.

1 Introduction

In recent years, large language models (LLMs) have emerged as the state of the art across a wide range of natural language processing tasks, including machine translation (MT). Owing to their vast number of parameters, these models possess the capacity to handle multiple languages and adapt

pairs. In many cases, obtaining sufficient parallel data for supervised training is infeasible.

This paper presents the Barcelona Supercomputing Center (BSC) submission to the terminology-aware MT task for the first track:

Agenda

1. From SALAMANDRA to SALAMANDRA-TA: BSC Submission for WMT25 General Machine Translation Shared Task
    - *Recipe*
    - *CP Dataset and SFT Dataset*
    - *Adding new languages to Salamandra*
    - *Quality-aware decoding*
    - *Results*

2. Terminology-Constrained Translation from Monolingual Data using GRPO
    - *Task*
    - *Proposed Methodology*
    - *Results*

3. Future work

Recipe

Backbone LLM
Salamandra7B
Salamandra2B

35 European languages
Pre-trained from scratch on
12.875T tokens
Parameters: **7,768,117,248**
Vocabulary: **256k**

1/Continual pre-training dataset

Continual Pre-training on **424B tokens** of translation pairs

SalamandraTA-base

Base Model improved for sentence-level translation

SalamandraTA-2B-base
SalamandraTA-7B-base

2/Instruction tuning dataset

Instruction tuning on high-quality instructions

SalamandraTA-instruct

Instructed Model based on SalamandraTA-base for translation related tasks

SalamandraTA-2B-instruct
SalamandraTA-7B-instruct

+ 7 New languages

Arabic
Bhojpuri
Hindi
Icelandic
Chinese
Japanese
Korean

WMT/Vocabulary-Adaptation + Continual pre-training dataset

SalamandraTA-base-WMT25

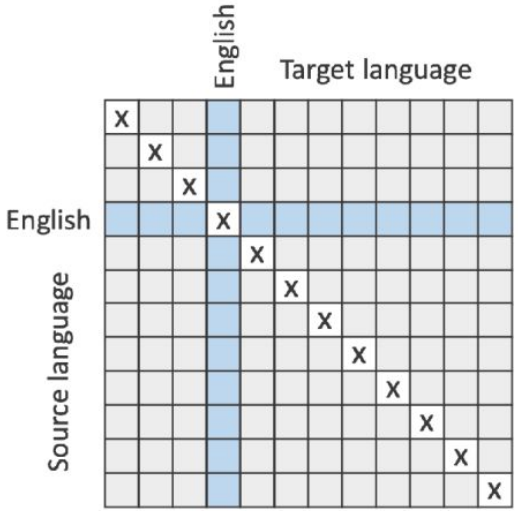
SalamandraTA base model adapted for new languages and scripts in WMT25.

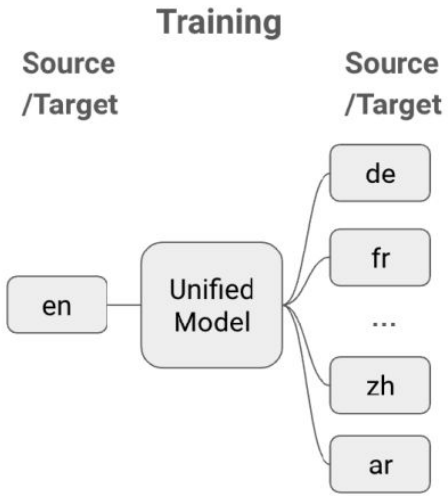
WMT/Instruction tuning for WMT25

SalamandraTA-instruct-WMT25

Motivations CP/ Parallel Data Mining

In real-world, most translation data is English-centric





English-centric Multilingual MT approach

But this can have some drawbacks

- Training with only a single bridge language can negatively impact translation performance across zero-shot language pairs
- or bias the models towards English-centric latent representations

we rely on parallel corpora for the continual pre-training stage pivoting on three bridge languages

We continually pre-train Salamandra-7B using **parallel data only** on 424B tokens

English Pairs

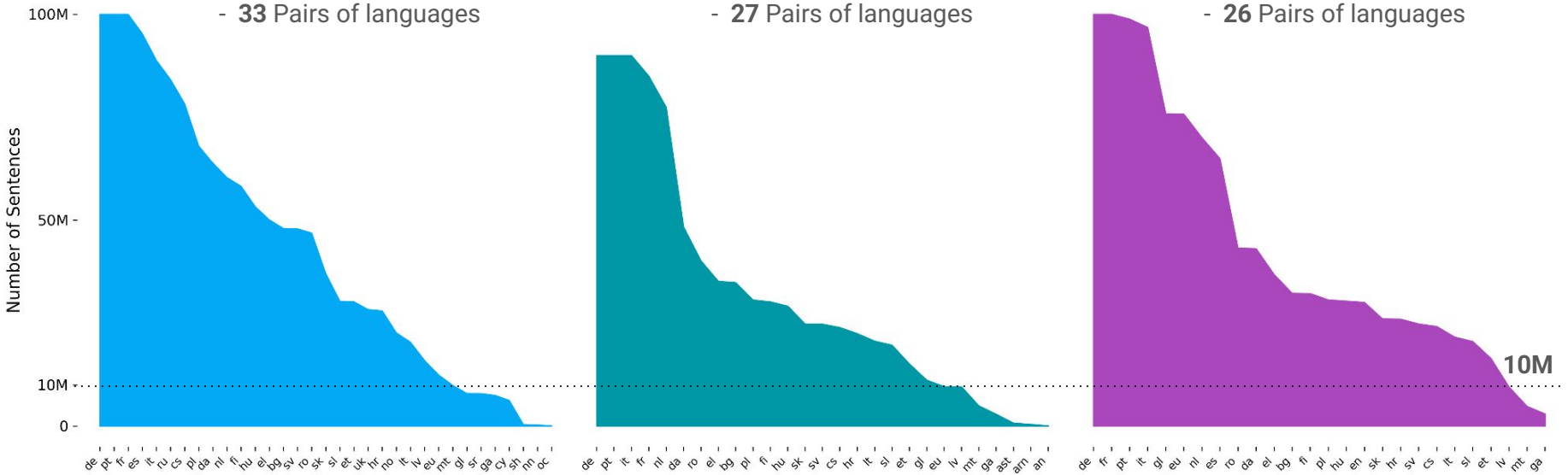
- **1.41B** sentences
- **33** Pairs of languages

Spanish Pairs

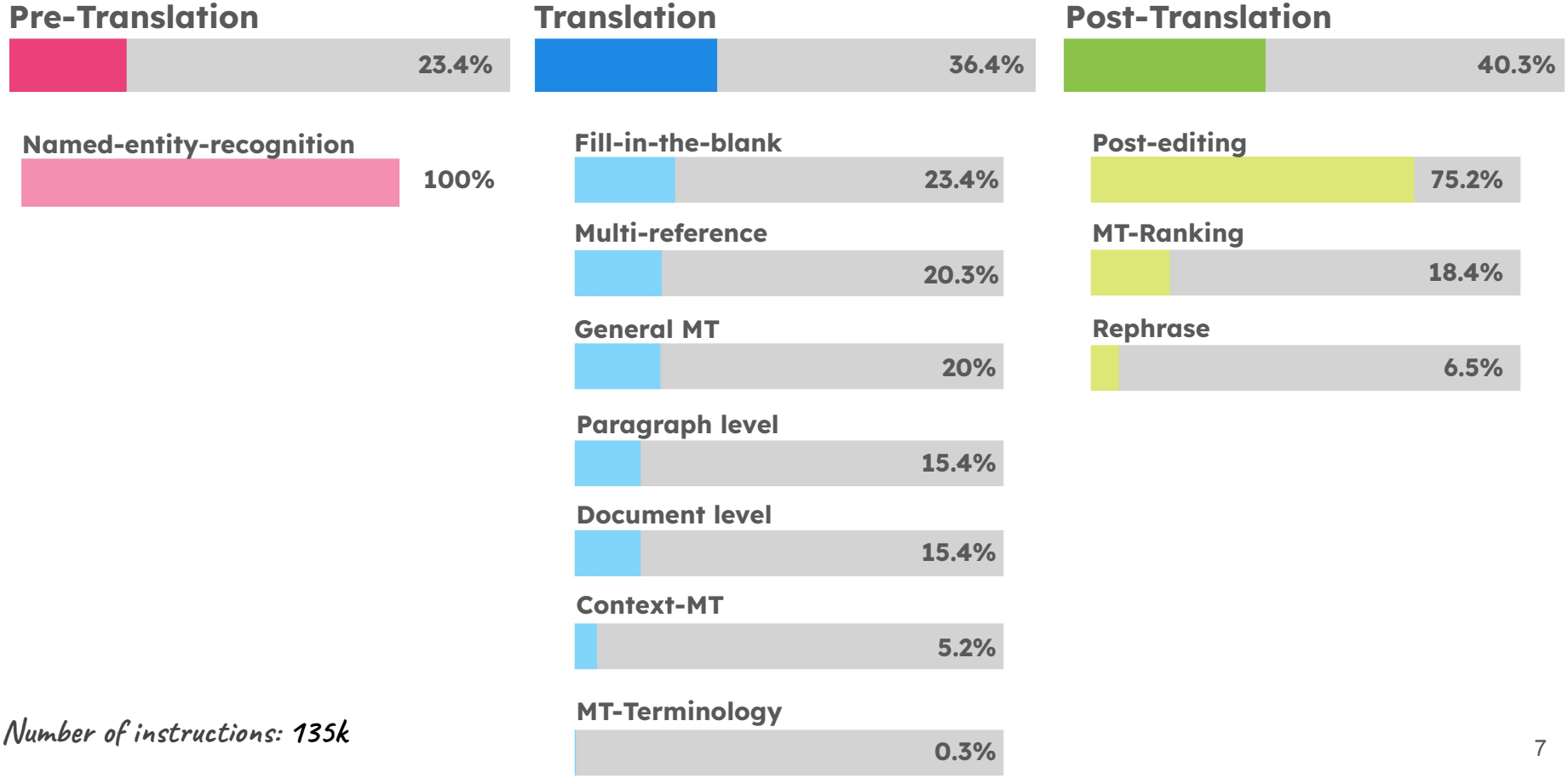
- **0.87B** sentences
- **27** Pairs of languages

Catalan Pairs

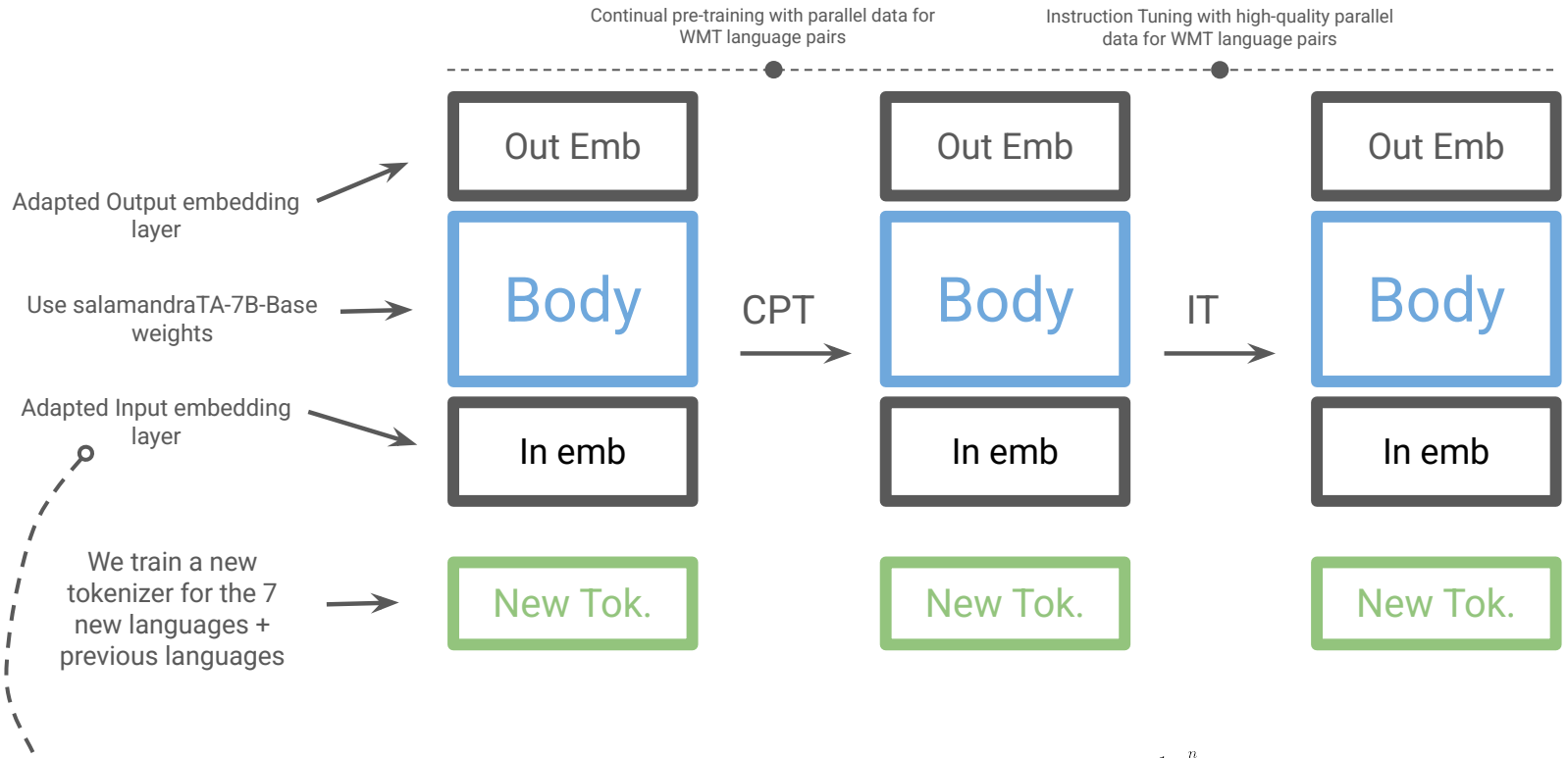
- **1.14B** sentences
- **26** Pairs of languages



In total, 37 languages; official EU + Low-Resource Languages of Spain



Number of instructions: 135k



Adaptation strategy: - *new tokens were initialized to the* ***average of previous embeddings*** $\mu = \frac{1}{n}\sum_{i=1}^{n} e_i$

- *tokens* ***common to both the old and new tokenizers*** *retained their* ***original embeddings***

Results

Instruction tuning yields significant gains over the CPT baselines, ***improving models by an average of 4 COMET points***

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	71.7	-	79.7	-	-	-	-	81.9	-	**84.1**	76.8	-	-
MADLAD400 7B	82.7	83.2	76.8	-	82.1	71.1	72.4	73.7	81.7	78.3	81.8	82.8	76.4
NLLB 3.3B	79.5	80.4	76.6	-	78.3	70.1	72.7	70.3	77.9	80.3	76.9	78.9	68.4
SALAMANDRA**TA**2B													
BASE + CPT-V1	80.3	80.1	76.0	-	69.6	-	-	-	-	-	80.1	57.0	-
+ INSTRUCT-V1	80.7	80.3	76.5	-	78.0	-	-	-	-	-	76.0	78.0	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	81.9	79.8	76.6	-	78.0	-	-	-	-	-	81.5	82.2	-
+ INSTRUCT-V1	85.3	86.6	80.3	-	83.8	-	-	-	-	-	81.6	83.4	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	81.1	79.3	76.2	79.4	77.0	69.3	70.6	74.7	75.5	75.9	81.5	82.5	77.3
+ INSTRUCT-V2	83.1	85.3	79.3	83.9	84.1	77.4	71.3	81.1	80.9	80.2	80.4	82.3	77.8

COMET scores on the WMT24++ test set, comparing our SalamandraTA models against several baselines.

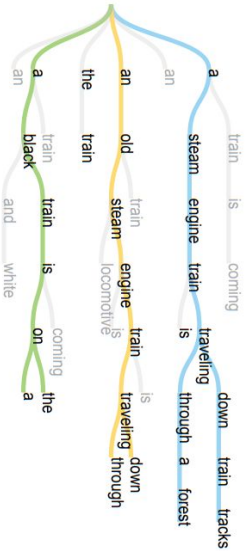
We lose some translation quality for previous language pairs →

For new language pairs we used SalamandraTA-WMT model, for previous language pairs we used original salamandraTA7B.

Quality-aware decoding \MBR and Re-ranking

Let's spend a bit more time on inference to align SalamandraTA machine translations with human translations!!

Generate diverse set of translations



We use **diverse beam search** strategy with a beam size of 20 and 5 beam groups.

Quality - aware decoding

In **reranking**, we find the translation that maximizes a given metric of interest

Relies on Comet-kiwi;
reference-less metric

$$\hat{y} = \arg\max_{h \in H} \mathrm{QE}(x, h)$$

In **minimum Bayes Risk (MBR)**, we find the translation that maximizes the expected utility

Relies on Comet;
reference-based metric

$$\hat{y}_i = \arg\max_{c_i \in C} \frac{1}{|C|} \sum_{j=1}^{|C|} \mathrm{RefMetric}(c_i, c_j)$$

Hypothesis used as references

Results \MBR and Re-ranking

QAD strategies yield ***significant gains*** *over Instruction tuning*

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
SALAMANDRA**TA**2B													
BASE + CPT-V1	80.3	80.1	76.0	-	69.6	-	-	-	-	-	80.1	57.0	-
+ INSTRUCT-V1	80.7	80.3	76.5	-	78.0	-	-	-	-	-	76.0	78.0	-
+ TRR	84.3	86.0	80.5	-	83.3	-	-	-	-	-	80.4	81.8	-
+ MBR	85.6	87.0	81.4	-	84.0	-	-	-	-	-	81.5	83.5	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	81.9	79.8	76.6	-	78.0	-	-	-	-	-	81.5	82.2	-
+ INSTRUCT-V1	85.3	86.6	80.3	-	83.8	-	-	-	-	-	81.6	83.4	-
+ TRR	85.9	87.6	82.0	-	85.0	-	-	-	-	-	81.3	84.0	-
+ MBR	**87.2**	**88.7**	**82.9**	-	85.9	-	-	-	-	-	**82.6**	**85.1**	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	81.1	79.3	76.2	79.4	77.0	69.3	70.6	74.7	75.5	75.9	81.5	82.5	77.3
+ INSTRUCT-V2	83.1	85.3	79.3	83.9	84.1	77.4	71.3	81.1	80.9	80.2	80.4	82.3	77.8
+ TRR	85.3	87.3	81.8	84.9	85.1	79.7	74.2	82.7	83.3	82.5	81.3	84.2	79.6
+ MBR	86.6	88.5	82.4	86.3	**86.1**	**80.7**	**75.5**	**83.4**	**84.1**	83.6	82.5	**85.1**	**80.4**

SALAMANDRATA-7B and SALAMANDRATA-V2 gap is largely mitigated when employing quality-aware decoding strategies

COMET scores on the WMT24++ test set, comparing our SalamandraTA models against several baselines.

Results \Ablation study; what is the cost for MT of adding additional tasks?

We can add Post-MT and Pre-MT tasks *"without"* hurting MT quality

The reduced number of tasks encourages the model to "split" its capacity, independently learning each task

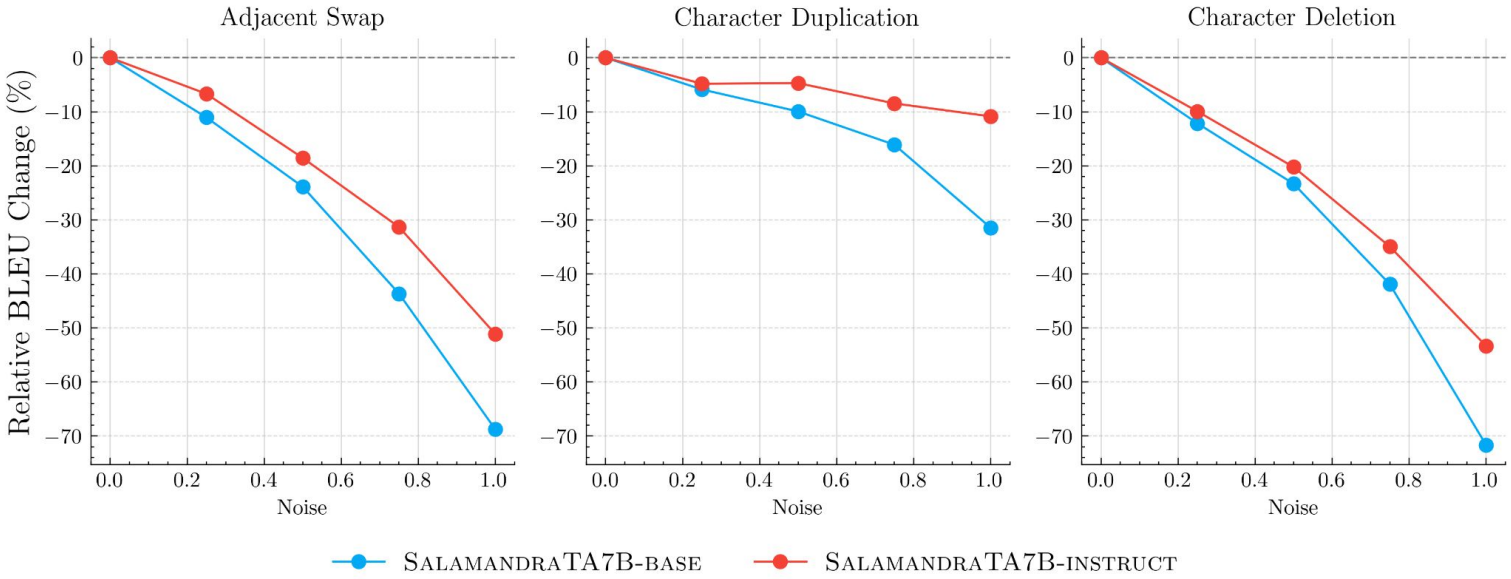
	en→xx		xx→en	
	COMET	BLEU	COMET	BLEU
SALAMANDRATA 7B BASE	0.85	33.33	0.88	43.01
Supervised Finetuning				
MT	0.87	35.55	0.88	44.22
+ Pre-MT + Post-MT	0.87	35.04	0.88	43.76
+ Chat + Code	0.87	34.45	0.88	43.98
MT + Post-MT	0.87	35.44	0.88	44.08
MT + Pre-MT	0.87	35.18	0.88	43.88

Towerblocks data! Synthetic Chat data and Code instructions in English

Table: Ablation results for the components of the instruction tuning dataset. We consider FLORES-200-devtest to evaluate translation quality.

Instruction tuning data improves robustness to misspellings

We evaluate **English -> Spanish** on three types of synthetic noise that have been previously used to stress NMT systems: **swap, chardupe, chardrop** on BLEU.



Performance in the Shared Task \RoCS-MT v2 Test suite

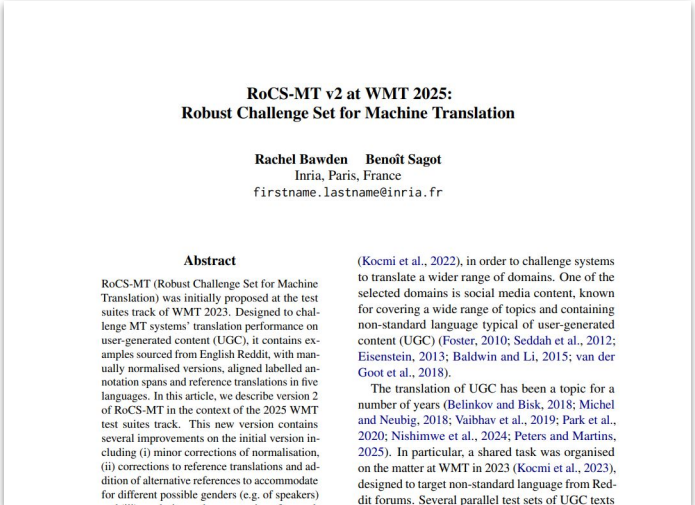
They evaluated our SalamandraTA submissions in *RoCS-MT v2 Test suite*: Designed to challenge MT systems' translation performance on **user-generated content (UGC)**, it contains examples sourced from English Reddit. Directions:

English -> {French, German, Czech, Ukrainian, Russian}

System	#lp	CometKiwi Rank	CometKiwi "Medals"	MetricX Rank	MetricX "Medals"	Overall rank	Δrank
Yolu	12	1	9, 1, 0	3	9, 1, 1	1	2
Shy-hunyuan-MT	12	4	6, 4, 2	1	11, 0, 1	2	3
CommandA-WMT	12	2	7, 3, 1	4	8, 2, 0	3	2
Laniqo°	7	5	6, 1, 0	5	4, 3, 0	4	0
SalamandraTA	11	3	7, 2, 1	8	1, 0, 1	5	5
GemTrans	12	12	1, 2, 1	2	10, 1, 0	6	10
UvA-MT	12	7	2, 4, 4	16	0, 5, 3	7	9
*GPT-4.1	12	17	0, 4, 1	6	1, 8, 1	7	11
*ONLINE-B	11	8	2, 1, 4	20	0, 2, 2	9	12
*TowerPlus 9B	12	9	1, 5, 2	21	0, 2, 0	10	12
*TowerPlus 72B	12	11	1, 2, 3	22	0, 2, 0	11	11
SRPOL	7	6	3, 4, 0	28	0, 1, 0	12	22
IR-MultiagentMT	12	23	0, 1, 1	17	0, 4, 4	13	6
TranssionTranslate	12	10	1, 2, 4	33	0, 0, 2	14	23
*CommandA	12	18	0, 2, 4	26	0, 1, 1	15	8
NNTSU	1	32	0, 0, 1	12	1, 0, 0	15	20
Erlendur	1	32	0, 0, 1	12	1, 0, 0	15	20
In2x	1	15	1, 0, 0	30	0, 1, 0	18	15
*Claude4	12	21	0, 1, 2	24	0, 1, 2	18	3
*DeepSeek V3	12	22	0, 1, 2	23	0, 1, 7	18	1
Algharb°	12	26	0, 1, 0	19	0, 2, 3	18	7
*Gemini 2.5 Pro	12	38	0, 0, 0	7	1, 1, 7	18	31
*Gemma 3 27B	12	28	0, 0, 1	18	0, 3, 4	23	10
*AyaExpanse 32B	12	13	1, 0, 2	35	0, 0, 1	24	22
*AyaExpanse 8B	12	20	0, 2, 0	29	0, 1, 0	25	9
KIKIS	1	39	0, 0, 0	12	1, 0, 0	26	27
*EuroLLM 22B	12	19	0, 2, 0	36	0, 0, 1	27	17
*Gemma 3 12B	12	25	0, 1, 0	32	0, 0, 2	28	7
Yandex	1	46	0, 0, 0	12	1, 0, 0	29	34
Systran°	1	15	1, 0, 0	44	0, 0, 0	30	29
*Llama 3.1 8B	12	14	1, 0, 0	47	0, 0, 0	31	33
Kaze-MT°	12	52	0, 0, 0	9	1, 0, 0	31	43
KYUoM°	12	52	0, 0, 0	10	1, 0, 0	33	42
ctpc_nlp	12	52	0, 0, 0	11	1, 0, 0	34	41
Wenyiil°	12	30	0, 0, 1	34	0, 0, 2	35	4
*Mistral-Medium	9	40	0, 0, 0	25	0, 1, 1	36	15
*CommandR	12	24	0, 1, 1	43	0, 0, 0	37	19
*Qwen3 235B	12	41	0, 0, 0	27	0, 1, 1	38	14
*ONLINE-W	8	27	0, 1, 0	42	0, 0, 0	39	15
AMI°	1	32	0, 0, 1	37	0, 0, 1	39	5
*EuroLLM 9B	12	29	0, 0, 1	41	0, 0, 0	41	12
IRB-MT	12	42	0, 0, 0	31	0, 0, 3	42	11
*Llama-4-Maverick	12	37	0, 0, 0	38	0, 0, 0	43	1
CUNI-MH-v2	1	32	0, 0, 1	46	0, 0, 0	44	14
bb88	1	32	0, 0, 1	49	0, 0, 0	45	17
*NLLB	12	44	0, 0, 0	39	0, 0, 0	46	5
*Mistral 7B	12	31	0, 0, 1	53	0, 0, 0	47	22
DLUT_GTCOM	2	45	0, 0, 0	40	0, 0, 0	48	5
CUNI-SFT	3	48	0, 0, 0	45	0, 0, 0	49	3
TranssionMT	8	43	0, 0, 0	51	0, 0, 0	50	8
*Qwen 2.5	12	47	0, 0, 0	48	0, 0, 0	51	1
CGFOKUS	1	51	0, 0, 0	49	0, 0, 0	52	2
*ONLINE-G	10	49	0, 0, 0	52	0, 0, 0	53	3
SH	1	50	0, 0, 0	54	0, 0, 0	54	4
CUNI-DocTransformer	1	55	0, 0, 0	55	0, 0, 0	55	0
COILD-BHO	1	55	0, 0, 0	55	0, 0, 0	55	0

5th place in this set

Several larger LLM-based baselines, such as Claude-4 (18th) and both Gemini models **appear in the middle of the table**



RoCS-MT v2 at WMT 2025:
Robust Challenge Set for Machine Translation

Rachel Bawden **Benoît Sagot**
Inria, Paris, France
firstname.lastname@inria.fr

Abstract

RoCS-MT (Robust Challenge Set for Machine Translation) was initially proposed at the test suites track of WMT 2023. Designed to challenge MT systems' translation performance on user-generated content (UGC), it contains examples sourced from English Reddit, with manually normalised versions, aligned labelled annotation spans and reference translations in five languages. In this article, we describe version 2 of RoCS-MT in the context of the 2025 WMT test suites track. This new version contains several improvements on the initial version including (i) minor corrections of normalisation, (ii) corrections to reference translations and addition of alternative references to accommodate for different possible genders (e.g. of speakers)

(Kocmi et al., 2022), in order to challenge systems to translate a wider range of domains. One of the selected domains is social media content, known for covering a wide range of topics and containing non-standard language typical of user-generated content (UGC) (Foster, 2010; Seddah et al., 2012; Eisenstein, 2013; Baldwin and Li, 2015; van der Goot et al., 2018).

The translation of UGC has been a topic for a number of years (Belinkov and Bisk, 2018; Michel and Neubig, 2018; Vaibhav et al., 2019; Park et al., 2020; Nishimwe et al., 2024; Peters and Martins, 2025). In particular, a shared task was organised on the matter at WMT in 2023 (Kocmi et al., 2023), designed to target non-standard language from Reddit forums. Several parallel test sets of UGC texts

Including English <-> Hindi directions benefits Bhojpuri

We find that during CPT, removing the EN→HI parallel data causes performance to drop from 9.32 to 0.35 BLEU and from 35.43 to 9.83 CHRF. This result provides clear evidence that the model relies on cross-lingual transfer from Hindi for translating to Bhojpuri.

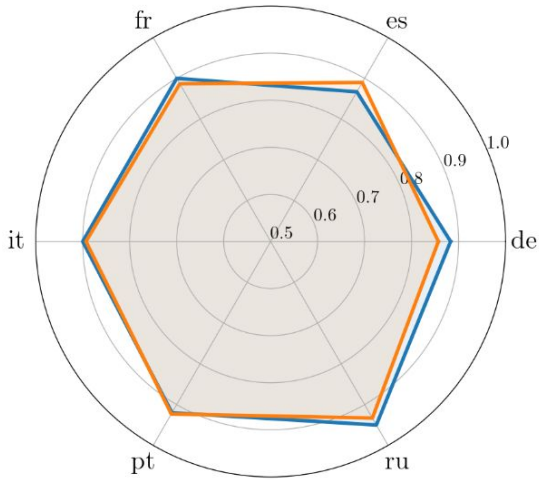
	BLEU	CHRF
Continual pre-training		
CPT-v2	9.32	35.43
CPT-v2 (no EN→HI)	0.35	9.83
Supervised Finetuning		
CPT-v2 + IT-v2	11.67	37.75

Knowledge transfer benefits **low-resource** languages

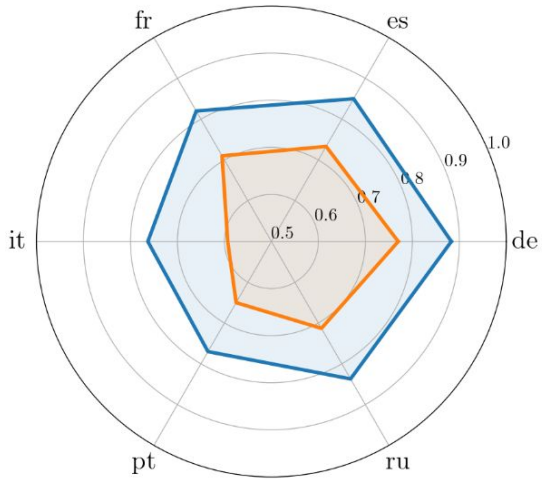
Instruction tuning data improves feminine and full-pair accuracy without sacrificing masculine accuracy (we included gender bias datasets in SFT)

We evaluate translation accuracy on gender-balanced sentence pairs using MT-GenEval-test

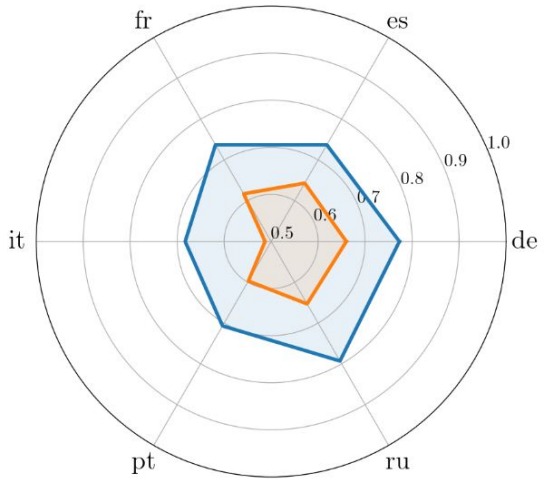
Masculine



Feminine



Pair



SalamandraTA-7b-instruct SalamandraTA-7b-base

Performance in the Shared Task \GENDER1PERSON

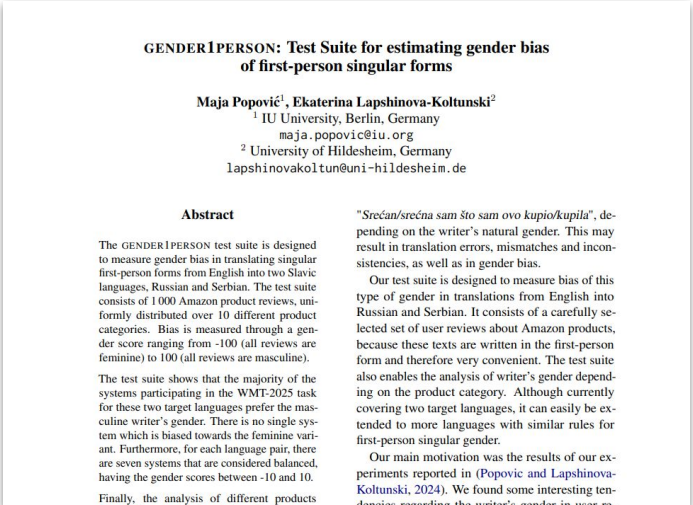
The GENDER1PERSON test suite is designed to measure gender bias in translating singular first-person forms from English into two Slavic languages, Russian and Serbian.

English -> {Russian, Serbian}

Bias towards masculine translations in most categories

And biased toward feminine translations in categories such as:

- *BEAUTY AND PERSONAL CARE*
- *BABY PRODUCTS*



GENDER1PERSON: Test Suite for estimating gender bias of first-person singular forms

Maja Popović[1], **Ekaterina Lapshinova-Koltunski**[2]
[1] IU University, Berlin, Germany
maja.popovic@iu.org
[2] University of Hildesheim, Germany
lapshinovakoltun@uni-hildesheim.de

Abstract

The GENDER1PERSON test suite is designed to measure gender bias in translating singular first-person forms from English into two Slavic languages, Russian and Serbian. The test suite consists of 1 000 Amazon product reviews, uniformly distributed over 10 different product categories. Bias is measured through a gender score ranging from -100 (all reviews are feminine) to 100 (all reviews are masculine).

The test suite shows that the majority of the systems participating in the WMT-2025 task for these two target languages prefer the masculine writer's gender. There is no single system which is biased towards the feminine variant. Furthermore, for each language pair, there are seven systems that are considered balanced, having the gender scores between -10 and 10.

Finally, the analysis of different products "*Srećan/srećna sam što sam ovo kupio/kupila*", depending on the writer's natural gender. This may result in translation errors, mismatches and inconsistencies, as well as in gender bias.

Our test suite is designed to measure bias of this type of gender in translations from English into Russian and Serbian. It consists of a carefully selected set of user reviews about Amazon products, because these texts are written in the first-person form and therefore very convenient. The test suite also enables the analysis of writer's gender depending on the product category. Although currently covering two target languages, it can easily be extended to more languages with similar rules for first-person singular gender.

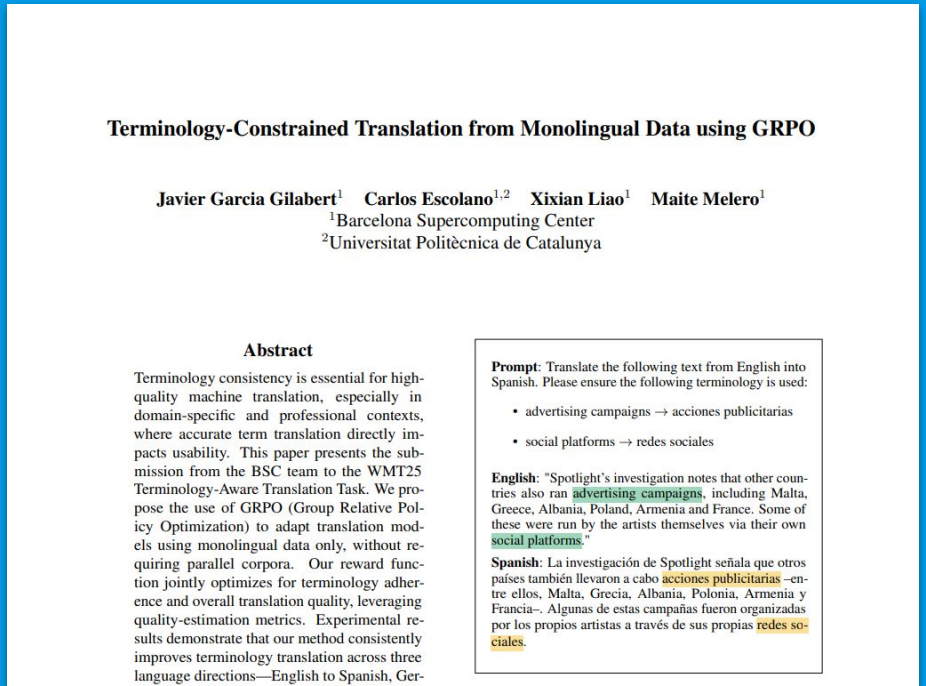
Our main motivation was the results of our experiments reported in (Popovic and Lapshinova-Koltunski, 2024). We found some interesting ten-

English→Russian						English→Serbian					
		distribution						distribution			
system	score	m	f	x	mix	system	score	m	f	x	mix
Algharb	-0.5	485	490	16	9	Gemini-2.5-Pro	-3.2	483	515	1	1
Yolu	-1.2	482	494	13	11	Algharb	-2.3	485	508	1	6
Yandex	-1.8	481	499	17	3	ONLINE-B	3.7	506	469	1	24
Gemini-2.5-Pro	1.7	499	482	16	3	Yolu	4.4	510	466	2	22
ONLINE-G	2.3	412	389	10	189	GemTrans	6.8	527	459	2	12
Wenyiil	8.3	528	445	19	8	Wenyiil	7.0	531	461	1	7
Shy	9.9	533	434	27	6	Shy	7.6	535	459	1	5
Laniqo	11.6	537	421	31	11	CUNI-SFT	16.0	549	389	5	57
GemTrans	14.9	568	419	9	4	GPT-4.1	20.7	601	394	1	4
SalamandraTA	16.8	561	393	15	31	Claude-4	21.4	604	390	1	5
TowerPlus-9B	17.1	579	408	9	4	EuroLLM-22B	22.9	589	360	2	49
IRB-MT	18.3	580	397	18	5	hybrid	22.9	609	380	3	8
hybrid	20.3	580	377	34	9	IRB-MT	23.3	608	375	2	15
Claude-4	20.4	590	386	20	4	Gemma-3-12B	27.9	606	327	2	65
Gemma-3-12B	23.1	603	372	15	10	AyaExpanse-32B	29.9	622	323	3	52
GPT-4.1	23.3	607	374	18	1	Gemma-3-27B	32.4	645	321	3	31
DeepSeek-V3	24.4	607	363	26	4	UvA-MT	32.5	656	331	1	12
ONLINE-W	25.5	528	273	11	188	DeepSeek-V3	34.5	668	323	1	8
DLUT_GTCOM	26.5	614	349	20	17	TowerPlus-9B	36.4	628	264	22	86
UvA-MT	27.4	630	356	10	4	AyaExpanse-8B	37.8	626	248	5	121
AyaExpanse-32B	31.7	649	332	11	8	Qwen3-235B	43.8	715	277	1	7
Qwen3-235B	34.4	662	318	12	8	CommandR7B	45.1	643	192	54	111
EuroLLM-22B	35.5	665	310	12	13	Llama-3.1-8B	46.7	707	240	3	50
Gemma-3-27B	38.2	681	299	13	7	CommandA	52.4	747	223	1	29
CommandA	39.3	686	293	14	7	IR-MultiagentMT	52.6	753	227	6	14
TowerPlus-72B	40.6	693	287	9	11	EuroLLM-9B	56.9	750	181	2	67
TranssionTranslate	44.7	645	198	10	147	Qwen2.5-7B	58.7	732	145	12	111
ONLINE-B	46.9	715	246	10	29	SalamandraTA	59.9	765	166	3	66
AyaExpanse-8B	47.6	728	252	9	11	Llama-4-Maverick	62.5	804	179	2	15
IR-MultiagentMT	47.6	719	243	30	8	CommandA-MT	69.6	841	145	1	13
Qwen2.5-7B	48.0	681	201	60	58	TowerPlus-72B	70.8	835	127	3	35
SRPOL	49.2	733	241	12	14	Mistral-7B	86.1	898	37	4	61
Llama-4-Maverick	54.0	762	222	13	3	ONLINE-G	86.7	878	11	3	108
CommandA-MT	54.8	767	219	8	6	TranssionMT	90.8	916	8	2	74
CommandR7B	55.0	753	203	17	27	TranssionTranslate	98.3	983	0	2	15
Llama-3.1-8B	55.2	750	198	5	47						
EuroLLM-9B	66.7	822	155	17	6						
NLLB	83.6	896	60	29	15						
Mistral-7B	90.9	938	29	3	30						
TranssionMT	98.5	985	0	6	9						

Overwhelmingly biased towards masculine translations!!

Terminology-Constrained Translation from Monolingual Data using GRPO

Javier Garcia Gilabert, Carlos Escolano, Xixian Liao, Maite Melero



Terminology-Constrained Translation from Monolingual Data using GRPO

Javier Garcia Gilabert[1] Carlos Escolano[1,2] Xixian Liao[1] Maite Melero[1]
[1]Barcelona Supercomputing Center
[2]Universitat Politècnica de Catalunya

Abstract

Terminology consistency is essential for high-quality machine translation, especially in domain-specific and professional contexts, where accurate term translation directly impacts usability. This paper presents the submission from the BSC team to the WMT25 Terminology-Aware Translation Task. We propose the use of GRPO (Group Relative Policy Optimization) to adapt translation models using monolingual data only, without requiring parallel corpora. Our reward function jointly optimizes for terminology adherence and overall translation quality, leveraging quality-estimation metrics. Experimental results demonstrate that our method consistently improves terminology translation across three language directions—English to Spanish, Ger-

Prompt: Translate the following text from English into Spanish. Please ensure the following terminology is used:

- advertising campaigns → acciones publicitarias
- social platforms → redes sociales

English: "Spotlight's investigation notes that other countries also ran advertising campaigns, including Malta, Greece, Albania, Poland, Armenia and France. Some of these were run by the artists themselves via their own social platforms."

Spanish: La investigación de Spotlight señala que otros países también llevaron a cabo acciones publicitarias –entre ellos, Malta, Grecia, Albania, Polonia, Armenia y Francia–. Algunas de estas campañas fueron organizadas por los propios artistas a través de sus propias redes sociales.

Task \Terminology-constrained translation

We participated in Track 1 of WMT25 Terminology Translation Task

- Sentence-level translation
- Domain: Information Technology (IT), SAP
- Translation directions: **English→German, English →Russian, and English→Spanish**
- Setup: Participants are provided with sentence segments, each with a small terminology dictionary containing only the terms present in the segment, usually 1-2 entries.

Example terminology translation:
`Translate the following text from English into Spanish. Please ensure the following terminology is used:`
clause -> cláusula.

English: To this end, it must comply with the WTO requirements and, in particular, with the GATT enabling clause of 1979.

Translation: **A tal fin, debe cumplir con los requisitos de la OMC y, en especial, con la cláusula de habilitación del GATT de 1979.**

State-of-the art techniques for terminology-translation usually use parallel data and post-training strategies. But…
what can we do if we don't have terminology parallel data for instance between Catalan and Russian?

Proposed method: Improve performance in terminology-translation without relying on parallel data and using only monolingual data and pseudo-terminology mined heuristically

Generating pseudo-terminology from monolingual data

\1st step: extract terms

We extract terms **automatically** using spacy: named entities, noun phrases, adverbial constructions.

Spotlight’s investigation notes that other countries also ran advertising campaigns, including Malta, Greece, Albania, Poland, Armenia and France. Some of these were run by the artists themselves via their own social platforms.

"advertising campaigns"
"social platforms"

Generating pseudo-terminology from monolingual data

\2nd step: translate terms using an MT model

For each detected term, we translate it **individually** using an MT model to our target language

Spotlight's investigation notes that other countries also ran advertising campaigns, including Malta, Greece, Albania, Poland, Armenia and France. Some of these were run by the artists themselves via their own social platforms.

"advertising campaigns": "acciones publicitarias"
"social platforms": "redes sociales"

Ideally we would like a translation like the following:

La investigación de Spotlight señala que otros países también llevaron a cabo acciones publicitarias –entre ellos, Malta, Grecia, Albania, Polonia, Armenia y Francia–. Algunas de estas campañas fueron organizadas por los propios artistas a través de sus propias redes sociales.

Generating pseudo-terminology from monolingual data

For each direction in; en->es, en->de, en->zh, we generate 5k of synthetic data using only **monolingual english data.**

Prompt format:

Translate the following text from English into Spanish.
Please ensure the following terminology is used:
- advertising campaigns → acciones publicitarias
- social platforms → redes sociales
English: Spotlight's investigation notes that other countries also ran advertising campaigns, including Malta, Greece, Albania, Poland, Armenia and France. Some of these were run by the artists themselves via their own social platforms, and some were promoted via the competing broadcaster in the territory — but with no link to their respective governments.
Spanish:

But...

How can we evaluate if the model is doing good in terminology consistency and also in translation quality if we have no access to parallel data???

We can use a quality estimation metric for translation quality!!!!

And we can check terminology accuracy for terminology consistency!!

Generating pseudo-terminology from monolingual data

Translate the following text from English into Spanish.
Please ensure the following terminology is used:
- advertising campaigns → acciones publicitarias
- social platforms → redes sociales

English: Spotlight's investigation notes that other countries also ran advertising campaigns, including Malta, Greece, Albania, Poland, Armenia and France. Some of these were run by the artists themselves via their own social platforms, and some were promoted via the competing broadcaster in the territory — but with no link to their respective governments.
Spanish:

	Term. Acc	Comet-kiwi	Total
Translation 1: 'La investigación de Spotlight señala que otros países también llevaron a cabo campañas publicitarias –entre ellos, Malta, Grecia, Albania, Polonia, Armenia y Francia–. Algunas de estas campañas fueron organizadas por los propios artistas a través de sus propias redes sociales.	0.5	+ 0.7	= 1.2
Translation 2: 'La investigación de Spotlight señala que otros países también llevaron a cabo acciones publicitarias –entre ellos, Malta, Grecia, Albania, Polonia, Armenia y Francia–. Algunas de estas campañas fueron organizadas por los propios artistas a través de sus propias redes sociales.	1.0	+ 0.7	= **1.7**

Second one is better

Reward definition

In each training step, for each source sentence q, we sample G candidate translations $\{o_1, o_2, \cdots, o_G\}$ from the current policy model π, then we optimize GRPO objective:

The reward function used to guide the reinforcement learning process combines two components:

Terminology adherence score (e.g., proportion of required target terms correctly present in the output)

$$S_i = \frac{1}{|T|} \sum_{i=1}^{|T|} \delta(t_i \in o_i)$$

$$r_i = S_i + \gamma(o_i, q)$$

Translation faithfulness score (reference-free quality metric)

We set $G = 4$

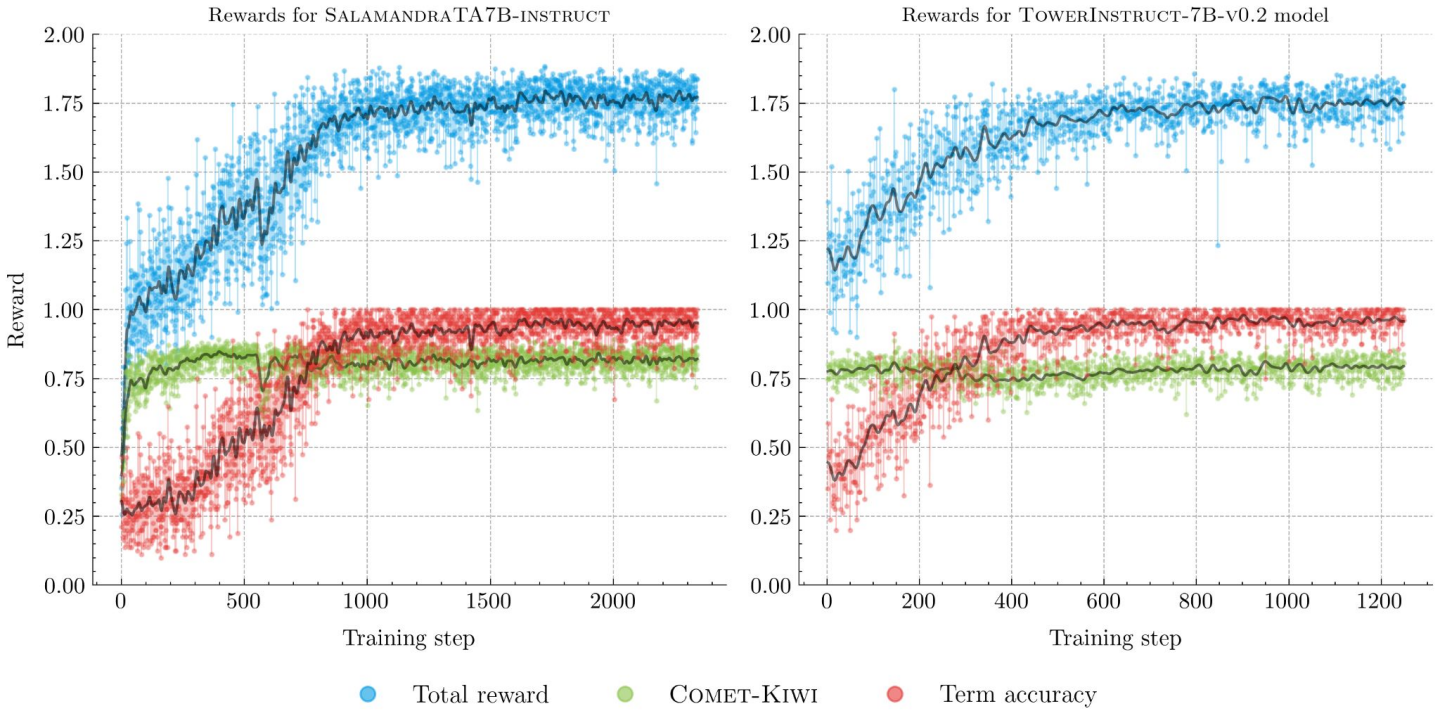
$$\frac{1}{G} \sum_{1}^{G} (\min(\nabla_\pi A_i, \text{clip}(\nabla_\pi, 1-\epsilon, 1+\epsilon) A_i) - \beta \mathcal{D}$$

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}.$$

We set beta to 0. Removing the KL constraint against the reference policy

Results \Reward curve

We experiment with two Translation-LLM models: **SalamandraTA7B-Instruct** and **Tower-Instruct-7B-v0.2.**
Both models have seen terminology instruction data in SFT stage.



Results

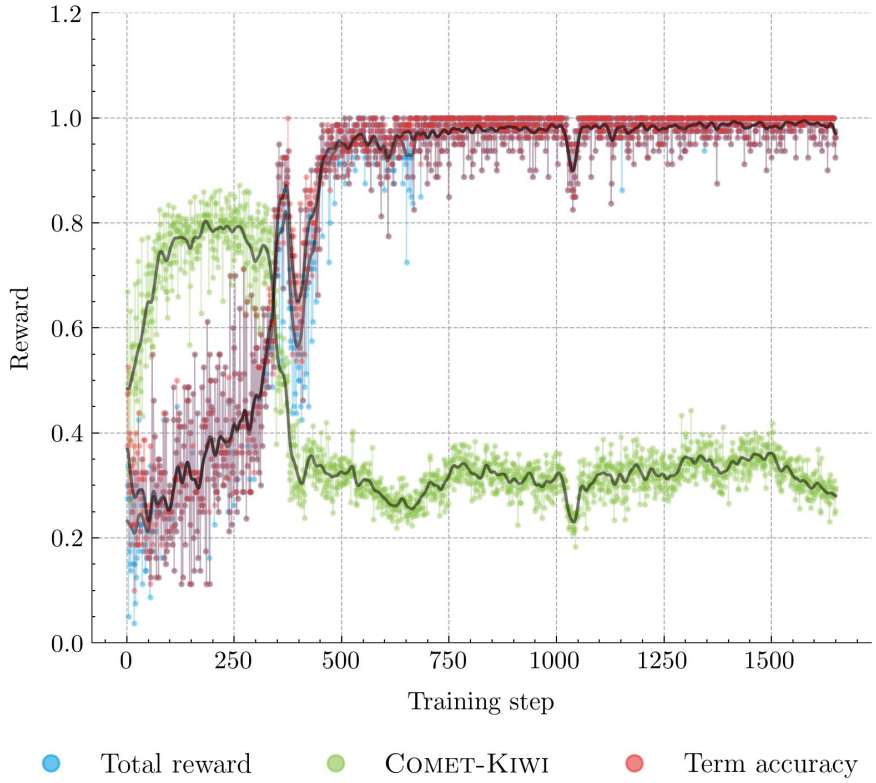
Direction	Model	T_P	T_F	BLEU	CHRF	COMET
En → Es	TowerInstruct-7B-v0.2	0.48	0.49	27.43	45.83	0.74
	+ GRPO	**0.93**	**0.91**	**51.27**	**74.21**	0.89
	SalamandraTA7B-instruct	0.54	0.54	43.64	62.82	0.79
	+ GRPO	0.90	0.88	47.46	73.75	**0.90**
En → De	TowerInstruct-7B-v0.2	0.60	0.59	38.81	65.43	0.86
	+ GRPO	**0.90**	**0.88**	39.40	68.33	0.87
	SalamandraTA7B-instruct	0.66	0.66	24.57	46.09	0.70
	+ GRPO	0.89	0.87	**44.46**	**71.26**	**0.89**
En → Ru	TowerInstruct-7B-v0.2	0.54	0.57	27.64	58.90	0.87
	+ GRPO	**0.87**	**0.86**	26.08	60.58	0.85
	SalamandraTA7B-instruct	0.66	0.68	20.70	45.10	0.72
	+ GRPO	0.84	0.85	**30.91**	**63.17**	**0.88**

*TowerInstruct improves +0.36 TP and salamandraTA improves +0.29 TP after GRPO on average **while also improving the translation quality!!***

Performance of TowerInstruct-7B-v0.2 and SalamandraTA7B-instruct models on terminology-aware translation for English-to-Spanish (En→Es), English-to-German (En→De), and English-to-Russian (En→Ru) directions. Results are reported for both base models and models aligned with GRPO.

Ablations

What happens when we remove comet-kiwi from the reward?



As the model gets better in terminology accuracy, translation quality drops.

The model learns to only copy-paste terms specified in the prompt (reward hacking)

Performance in the Shared Task

System	Proper, ChrF Avg	Es	De	Ru	Proper, Acc. Avg	Es	De	Ru	Proper, Cons. Avg	Es	De	Ru	Random, ChrF Avg	Es	De	Ru	Random, Acc. Avg	Es	De	Ru	NoTerm, ChrF Avg	Es	De	Ru
o3-term-guide	71.0	75.9	71.6	65.6	99.1	99.1	99.1	99.0	87.7	86.7	86.1	90.4	68.1	72.4	69.4	62.4	49.2	50.7	52.3	44.6	63.6	69.5	64.7	56.6
duterm	70.1	76.1	70.7	63.6	98.2	98.7	98.2	97.6	87.3	86.0	86.3	89.5	66.4	72.1	67.2	59.8	46.6	48.8	48.4	42.4	61.6	67.0	62.6	55.3
Erlendur	69.3	74.8	69.9	63.3	92.9	94.4	93.2	91.2	86.7	83.8	86.3	90.0	66.4	71.6	67.6	59.8	44.4	47.1	47.1	38.9	62.6	68.1	64.0	55.6
TiUTermV1	68.9	77.1	65.7	63.8	87.6	89.4	87.3	86.1	86.7	85.7	85.9	88.5	66.8	74.2	64.4	61.8	54.6	59.2	56.7	47.9	64.4	72.4	61.9	58.9
GPT-4.1-nano■	67.4	72.4	67.4	62.3	90.7	95.2	89.0	88.0	87.5	86.3	86.3	90.0												
salamandrata	67.3	72.0	69.6	60.4	91.3	92.7	91.7	89.4	87.4	87.3	86.4	88.6	64.7	69.3	66.2	58.5	48.2	53.1	48.1	43.4	62.0	67.2	64.0	54.7
MeGuMa	67.2	72.0	67.7	61.9	97.4	97.0	96.3	98.8	88.6	86.9	88.6	90.2	64.5	70.3	64.2	59.0	46.7	53.1	46.4	40.5	58.9	65.2	59.4	52.1
tower	66.0	74.0	65.9	58.1	93.7	95.0	94.8	91.2	88.4	87.6	86.8	90.7	63.8	71.2	63.0	57.1	44.3	48.6	45.7	38.5	60.9	68.6	61.2	53.0
CommandA$_{WMT}$	65.9	70.7	67.6	59.3	79.9	81.9	86.9	70.7	86.6	84.5	87.5	87.8	63.7	68.4	65.0	57.6	45.8	49.3	48.1	40.1	60.7	65.5	62.2	54.4
BIT	63.7	69.8	62.4	58.9	97.0	96.3	98.0	96.7	87.8	86.8	86.9	89.8	65.7	67.2	66.3	63.5	80.5	47.5	97.4	96.5	66.5	69.8	66.3	63.5
TiUTermV0	62.7	69.0	61.0	58.3	74.4	75.2	71.1	76.8	86.4	85.0	85.6	88.6	61.0	68.1	59.1	55.8	49.6	54.2	49.9	44.8	60.2	68.0	57.9	54.6
laniqo	61.7	68.5	59.8	56.9	99.3	98.7	99.4	99.6	87.6	85.6	89.3	87.9	60.2	66.3	59.5	54.8	42.7	46.9	43.5	37.7	55.0	60.3	55.5	49.4
LC-primary	61.4	68.9	61.2	54.2	70.2	74.1	70.7	65.8	85.4	83.6	85.8	87.0	61.0	68.1	59.7	55.2	38.6	43.8	37.4	34.6	57.5	65.0	56.9	50.5
LC-2	60.8	67.7	61.0	53.7	70.0	73.6	70.7	65.6	85.8	85.4	85.7	86.2	60.5	67.1	59.5	54.9	38.5	43.4	37.4	34.6	56.9	64.1	56.8	49.9
LC-3	60.8	67.7	61.0	53.7	70.0	73.6	70.7	65.6	86.0	85.6	85.7	86.7	60.5	67.1	59.5	54.9	38.5	43.4	37.4	34.6	56.9	64.1	56.8	49.9
CurTermNLLB	60.1	69.1	60.3	51.0	63.4	76.5	79.0	34.6	88.0	87.5	87.6	88.8	58.8	67.4	58.0	50.8	36.1	44.1	31.7	32.6	55.6	65.6	52.8	48.4
ContexTerm	48.5	53.7	40.2	51.5	72.0	68.5	79.9	67.6	81.9	75.6	85.8	84.4	48.2	52.0	40.7	51.7	24.6	20.5	18.6	34.8	45.7	50.2	37.4	49.4
Systran-ft		71.1				44.1				88.1				71.1				44.1				71.1		
MarianMT-ft		65.6				17.5				54.1				68.9				48.8				68.9		
EuroLLM-ft		63.5				38.9				82.5				63.5				38.9				63.5		
TranssionMT				47.8				33.2				90.1				47.8				33.2				47.8

Table 2: Main results for Track 1: sentence-level IT documentation terminology-informed translation.

5 top submissions were using Closed-Propietary models: GPT4.1, Claude 3.5, etc.

Very similar performance to GPT4.1-nano

Future plans: SalamandraTA-Extended, ALIA-TA

What have we learned from WMT25?

Continual pre-training

Out Emb

Body

In emb

Vocab Adaptation

Before we were using parallel data **only** for CPT; **CPT WMT25**

Now we have two stages

Monolingual CPT

zh en de ko jp el
ca es hi cs fi bu
ar pt it ar an 43 langs. ...

Monolingual + Parallel data CPT

es-zh es-de es-ja es-ko
en-zh en-de en-ja en-ko
ca-zh ca-de ca-ja ca-ko
zh de ja ko 43 langs. ...

Out Emb

Body

In emb

New Tok.

Before we were initializing the new embeddings to the mean of previous embeddings, now we initialize new embeddings from a multivariate Gaussian distribution

$$e_{n+1} \sim \mathcal{N}(\mu, \Sigma)$$

$$\Sigma = (E - \mu)^\top (E - \mu)/n$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} e_i$$

Embedding matrix

SalamandraTA-7B can be used as a teacher model

Online Distillation. We explored Word-Level knowledge distillation [Hinton, et al.]

We ran a continual pre-training on Salamandra2B (student model) on the training data but with an additional objective: to minimize the cross-entropy with respect to the word-level distribution of the teacher model

Teacher has seen 424B tokens

Experiments have only been trained on 2B tokens of parallel data

	en→xx		xx→en	
	COMET	BLEU	COMET	BLEU
SALAMANDRATA 7B BASE **(Teacher)**	0.85	33.33	0.88	43.01
SALAMANDRATA 2B DISTILLED INSTRUCT **(Student)**	**0.85**	**28.92**	**0.86**	**38.23**
SALAMANDRATA 2B INSTRUCT	0.83	26.59	**0.86**	37.46
SALAMANDRATA 7B INSTRUCT	0.87	35.04	0.88	43.76

Table: Translation performance (BLEU) on Flores+200-devtest for SalamandraTA models in both en→xx and xx→en directions.

Future plans: **Train ALIA-TA (40B) and use it for offline knowledge distillation**

We get significant BLEU improvements in zero-shot directions for low-resource languages

When we remove multi-parallel-data, we can't get improvements [Wu, et al.]

	Aranese		Aragonese	
	EN→ARN	ARN→EN	EN→ARG	ARG→EN
SALAMANDRATA 7B BASE	8.36	17.92	12.24	31.26
(SFT) MT + Pre-MT + Post-MT	**13.04** (+4.68)	**21.15** (+3.23)	**20.43** (+8.19)	**36.45** (+5.19)
- Multi Parallel Data	8.98 (+0.62)	18.52 (+0.60)	8.00 (-4.24)	31.63 (+0.37)

Table: Translation performance (BLEU) of SalamandraTA-7B Base model and its SFT on low-resource language pairs involving Aranese and Aragonese. SFT significantly improves translation quality with gains of up to +8.19 BLEU points but removing multi-parallel data narrows the performance gap.

[Wu, et al.] Wu, D., Tan, S., Meng, Y., Stap, D., & Monz, C. (2024, August). How Far can 100 Samples Go? Unlocking Zero-Shot Translation with Tiny Multi-Parallel Data. In Findings of the Association for Computational Linguistics ACL 2024 (pp. 15092-15108).

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thanks!

javier.garcia1@bsc.es

Appendices

Continual Pre-Training

- 64 nodes - 4 H100 (64GB) per node = **256 GPUs**
- Batch size: **512**, Context: **8192**
- Epochs: **1**, LR: **3.0e-05**, Optim: **Fused Adam optimizer**
- Framework: **Nemo-Nvidia**

Supervised Finetuning

- 4 nodes - 4 H100 (64GB) per node = **16 GPUs**
- Batch size: **16**, Context: **8192**
- Epochs: **1**, LR: **1e-5**, Optim: **AdamW optimizer**
- Chat Template: **ChatML template**
- Framework: **FastChat + Deepspeed**