Barcelona Supercomputing Center
Centro Nacional de Supercomputación

SalamandraTA: A European Multilingual Large Language Model for Translation-Related Tasks

Javier García Gilabert, Carlos Escolano, Audrey Mash, Xixian Liao, Francesca De Luca Fornaciari, Miguel Claramunt Argote, Ella Bohman and Maite Melero

31 March, 2025

Virtual Talk @ Meta

Agenda

1. Motivations
2. Recipe
3. Results
4. Takeaways

Motivations

1. Train a high-quality translator for EU languages and Spain's low-resource languages
2. Have a model that is able to perform translation-related tasks

Some questions we investigated;

- Impact of non-MT tasks in instruction tuning
- How can we extend Tower recipe to 37 languages?
- Gender-Bias evaluation after instruction tuning
- Robustness to misspellings

arXiv:2402.17733v1 [cs.CL] 27 Feb 2024

TOWER: An Open Multilingual Large Language Model for Translation-Related Tasks

Duarte M. Alves[†2,4] **José Pombal**[†1] **Nuno M. Guerreiro**[†1,2,4,5] **Pedro H. Martins**[1] **João Alves**[1] **Amin Farajian**[1] **Ben Peters**[2,4] **Ricardo Rei**[1,3] **Patrick Fernandes**[2,4,7] **Sweta Agrawal**[*2] **Pierre Colombo**[5,6] **José G.C. de Souza**[1] **André F.T. Martins**[1,2,4]

[1]Unbabel, [2]Instituto de Telecomunicações, [3]INESC-ID, [4]Instituto Superior Técnico & Universidade de Lisboa (Lisbon ELLIS Unit), [5]MICS, CentraleSupélec, Université Paris-Saclay, [6]Equall, [7]Carnegie Mellon University

[†]Equal contribution, ordered alphabetically by the first name.
[*]Work partially developed during an internship at Unbabel.
duartemalves@tecnico.ulisboa.pt, {jose.pombal, nuno.guerreiro}@unbabel.com.

While general-purpose large language models (LLMs) demonstrate proficiency on multiple tasks within the domain of translation, approaches based on open LLMs are competitive only when specializing on a single task. In this paper, we propose a recipe for tailoring LLMs to multiple tasks present in translation workflows. We perform continued pretraining on a multilingual mixture of monolingual and parallel data, creating TOWERBASE, followed by finetuning on instructions relevant for translation processes, creating TOWERINSTRUCT. Our final model surpasses open alternatives on several tasks relevant to translation workflows and is competitive with general-purpose closed LLMs. To facilitate future research, we release the TOWER models, our specialization dataset, an evaluation framework for LLMs focusing on the translation ecosystem, and a collection of model generations, including ours, on our benchmark.

1 Introduction

Many important tasks within multilingual NLP, such as quality estimation, automatic post-edition, or grammatical error correction, involve analyzing, generating or operating with text in multiple languages, and are relevant to various translation workflows — we call these **translation-related tasks**. Recently, general-purpose large language models (LLMs) challenged the paradigm of *per-task* dedicated systems, achieving state-of-the-art performance on several recent WMT shared tasks (Kocmi et al., 2023; Freitag et al., 2023; Neves et al., 2023). Unfortunately, strong capabilities for *multiple* translation-related tasks have so far been exhibited by *closed* LLMs only (Hendy et al., 2023; Kocmi & Federmann, 2023; Fernandes et al., 2023; Raunak et al., 2023). Perhaps because most *open* LLMs are English-centric, approaches leveraging these models still lag behind, having thus far achieved competitive results only when specializing on a *single* task (Xu et al., 2024a; 2023; Iyer et al., 2023).

In this paper, we bridge this gap with a detailed recipe to develop an LLM for *multiple* translation-related tasks. Our approach, illustrated in Figure 1 and inspired by Xu et al.

Recipe

1/Continual pre-training dataset

2/Instruction tuning dataset

Backbone LLM
Salamandra7B

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

SalamandraTA-base

Base Model improved for sentence-level translation

SalamandraTA-7B-base

SalamandraTA-instruct

Instructed Model based on SalamandraTA-base for translation related tasks

SalamandraTA-7B-instruct

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

Instruction tuning on high-quality instructions

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

English Pairs

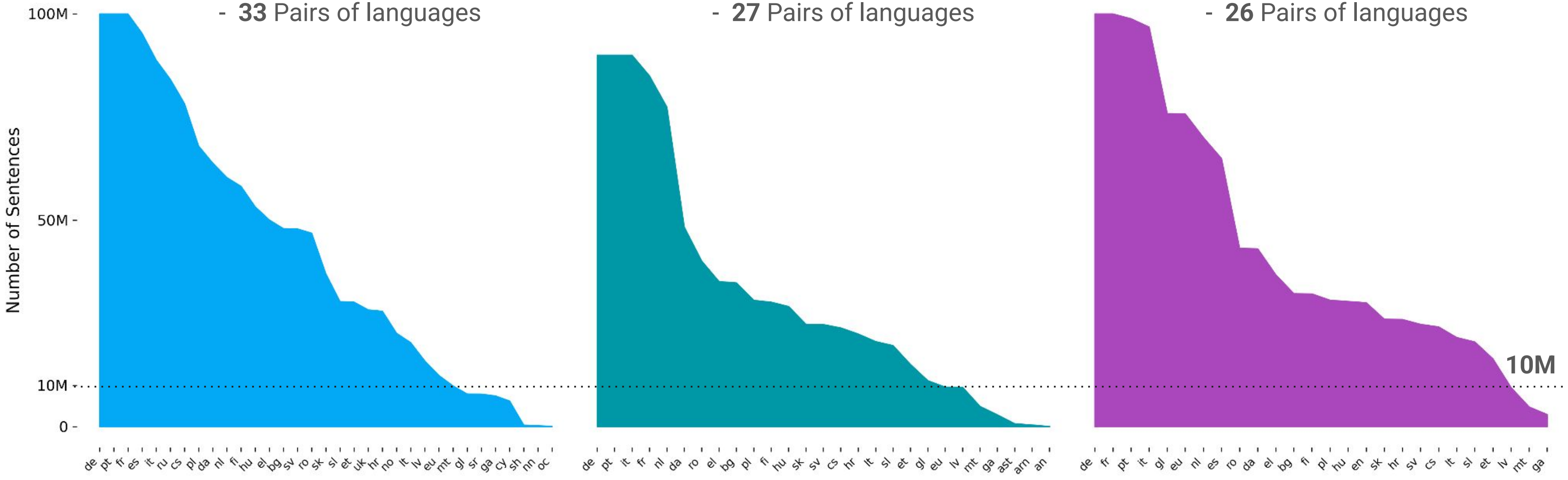
- **1.41B** sentences
- **33** Pairs of languages

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

2/Instruction tuning dataset

Pre-Translation — 23.4%

- Named-entity-recognition: 100%

Translation — 36.4%

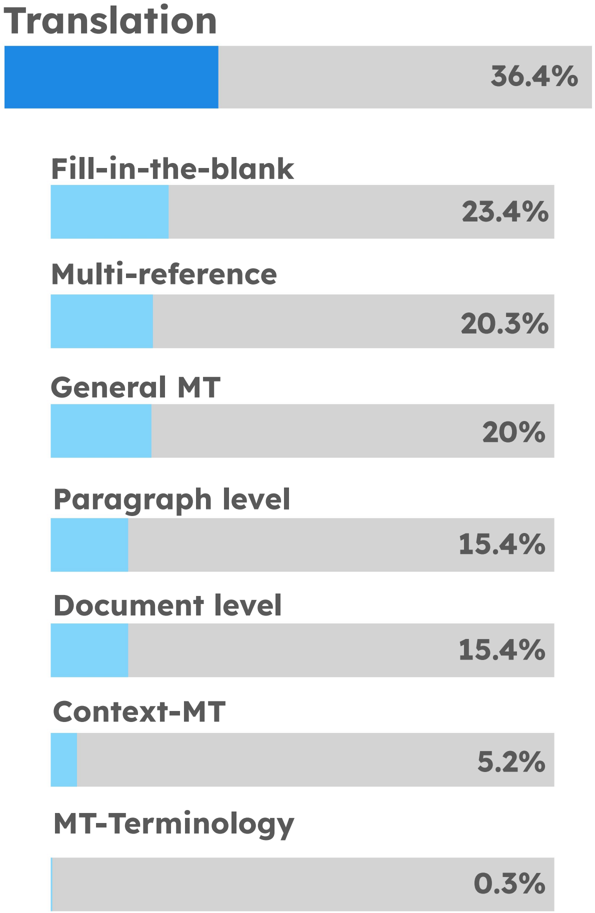
- Fill-in-the-blank: 23.4%
- Multi-reference: 20.3%
- General MT: 20%
- Paragraph level: 15.4%
- Document level: 15.4%
- Context-MT: 5.2%
- MT-Terminology: 0.3%

Post-Translation — 40.3%

- Post-editing: 75.2%
- MT-Ranking: 18.4%
- Rephrase: 6.5%

Number of instructions: 135k

2/Instruction tuning dataset: **Translation**



TowerBlocks

- 👉 Tatoeba Dev
- 👉 OPUS, Flores-dev, Ntrex, WMT, FRMT
- 👉 MT-GenEval-dev
- 👉 WMT-23

Ours

- 👉 Random word masking using Flores
- 👉 Multi-Parallel data using Flores
- 👉 Europarl, News-Commentary
- 👉 Europarl, News-Commentary

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Pre-Translation

23.4%

Named-entity-recognition 100%

TowerBlocks

- MultiCoNER-Dev 2022 and 2023

Ours

- SLI NERC **Galician**
- EusIE **Basque**
- BasqueGLUE **Basque**
- AnCora-Ca-NER **Catalan**

Post-Translation

40.3%

Post-editing 75.2%

MT-Ranking 18.4%

Rephrase 6.5%

TowerBlocks

- QT21 project, ApeQuest project
- WMT17 to WMT22 Metrics Assessments
- PAWS-X Dev

Ours

- Translations from Google Translate, NLLB and MADLAD ranked using automatic metrics

Results \Aggregated results

We evaluated SalamandraTA models on English, Spanish, Catalan and Basque translation directions

SalamandraTA-7B-Instruct model obtains state-of-the-art performance across all language pairs outperforming strong baselines and improving translation quality compared to SalamandraTA-7B base

	English		Spanish		Catalan		Basque	
	EN→XX	XX→EN	ES→XX	XX→ES	CA→XX	XX→CA	EU→XX	XX→EU
SalamandraTA 7B Base	34.99 3	44.12 2	21.63 3	24.71 3	29.06 3	32.75 3	22.87 2	17.01 2
SalamandraTA 7B-Instruct	**36.29** 1	**44.69** 1	**23.67** 1	**25.56** 1	29.23 1	**33.64** 1	**22.99** 1	**17.50** 1
MadLad-400-7B	35.73 2	43.20 3	22.48 2	24.85 2	**29.37** 1	33.02 2	21.26 3	13.64 3
NLLB-3.3B	31.17 4	41.52 4	19.54 4	22.68 4	25.17 4	29.28 4	18.83 4	7.58 4

Table: Results for machine translation (BLEU) aggregated by language pair using Flores+200 devtest. We highlight the best ranked models in bold.

Results \Comparison with Tower-Instruct

English→XX

	DE	ES	FR	IT	NL	PT	RU
SalamandraTA 7B Base	40 2	28.8 2	51.3 2	31.0 2	27.4 2	51.0 1	32.0 2
SalamandraTA 7B-Instruct	41 1	31.6 1	53.2 1	33.9 1	28.3 1	49.6 2	32.6 1
TowerInstruct 7B v2.0	39.3 3	28.6 3	49.9 3	31.4 3	28.2 3	46.2 3	31.5 3

XX→English

	DE	ES	FR	IT	NL	PT	RU
SalamandraTA 7B Base	40 2	32.4 2	47.7 2	36.2 2	34.0 2	53.0 1	37.3 2
SalamandraTA 7B-Instruct	41 1	37.2 1	48.5 1	38.6 1	35.3 1	52.6 2	39.7 1
TowerInstruct 7B v2.0	39.3 3	31.7 3	47.4 3	35.2 3	33.3 3	51.5 3	37.3 2

Table: Results for machine translation (BLEU) using Flores+200 devtest. We highlight the best ranked models in bold.

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

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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.

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

SalamandraTA-7B instruct is an effective post-editor!

We evaluate automatic post-editing (APE) by measuring final translation quality **after post-editing EuroLLM-1.7B translations for English <-> Catalan**

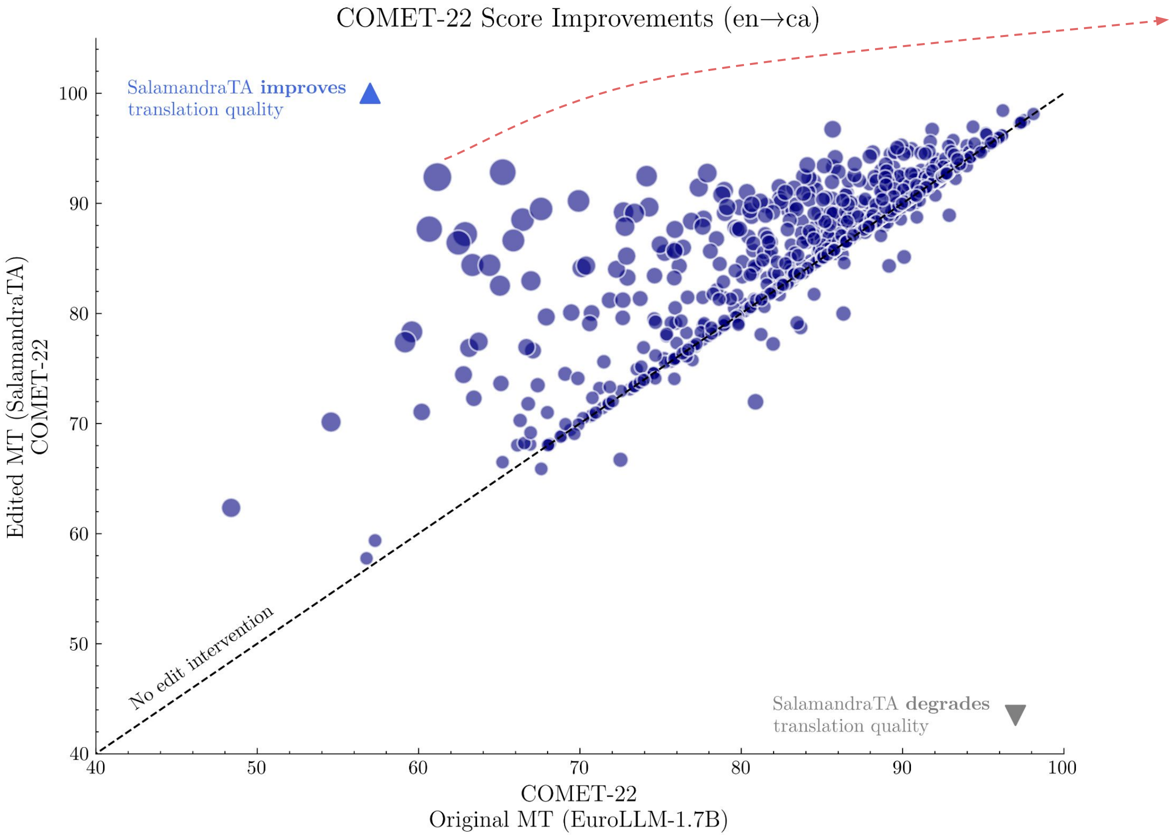
Catalan <-> English is not present in APE instruction dataset

By post-editing using SalamandraTA-7B-instruct we are able to improve the translation quality

	Ca→En		En→Ca	
	BLEU	COMET	BLEU	COMET
EUROLLM-1.7B	44.83	0.87	37.99	0.86
Post-edited by SALAMANDRATA 7B-INSTRUCT	45.30 (+0.47)	0.88 (+0.01)	41.10 (+3.11)	0.88 (+0.02)
SALAMANDRATA 7B-INSTRUCT	**45.40**	**0.89**	**50.27**	**0.89**

Table: Translation performance (BLEU) of EuroLLM-1.7B, post-edited translations by SalamandraTA-7B-instruct and SalamandraTA-7B-instruct on Flores+200-devtest.

Results \Post-Editing en->ca



Source: The ministry responded by calling Apple's postponement of the report "truly regrettable."

EuroLLM MT: El ministeri va respondre que el posposament de l'Apple del **reportatge** "realment lamentable".

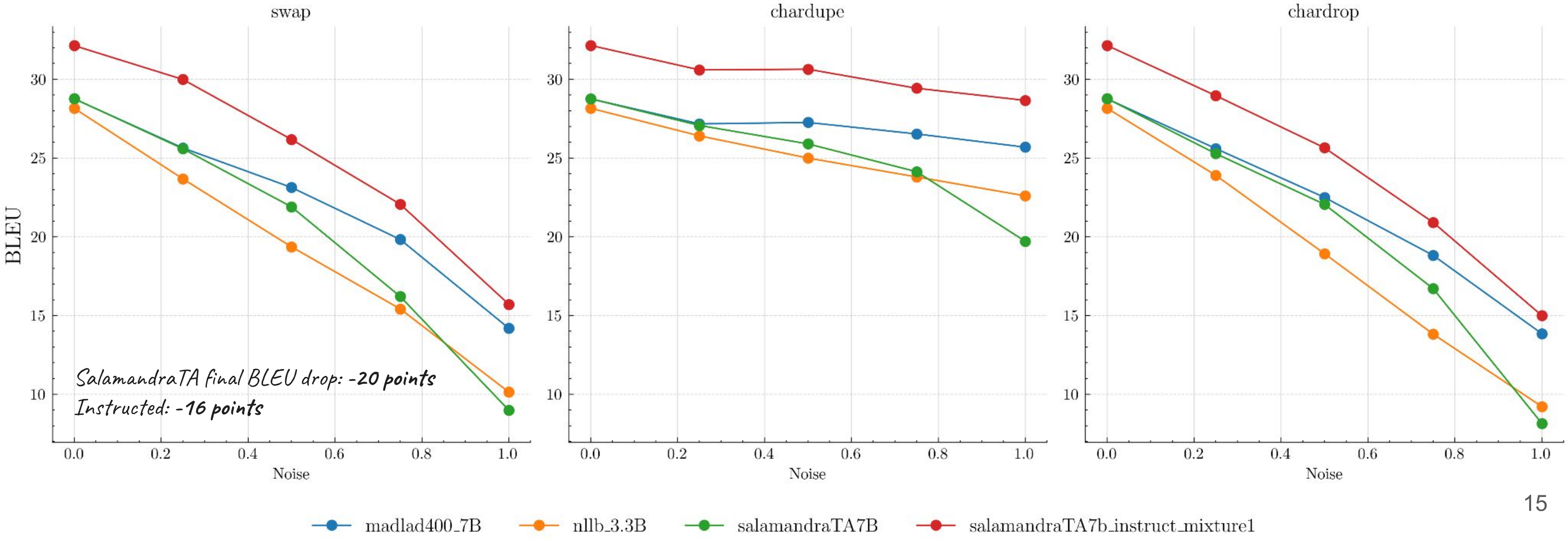
Edited: El ministeri va respondre que el posposament d'Apple de **l'informe** era "realment lamentable".

SalamandraTA MT: El ministeri va respondre dient que l'ajornament de l'informe per part d'Apple era "veritablement lamentable".

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

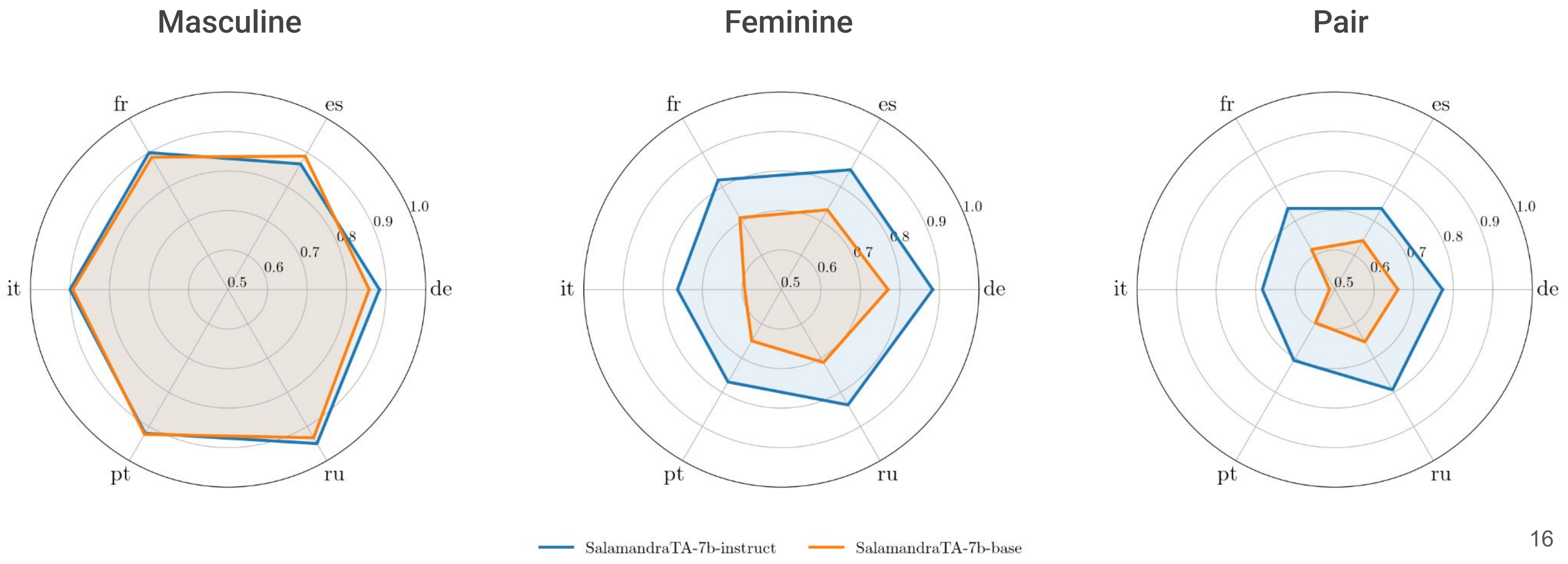
Instruction tuning data improves robustness to misspellings

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



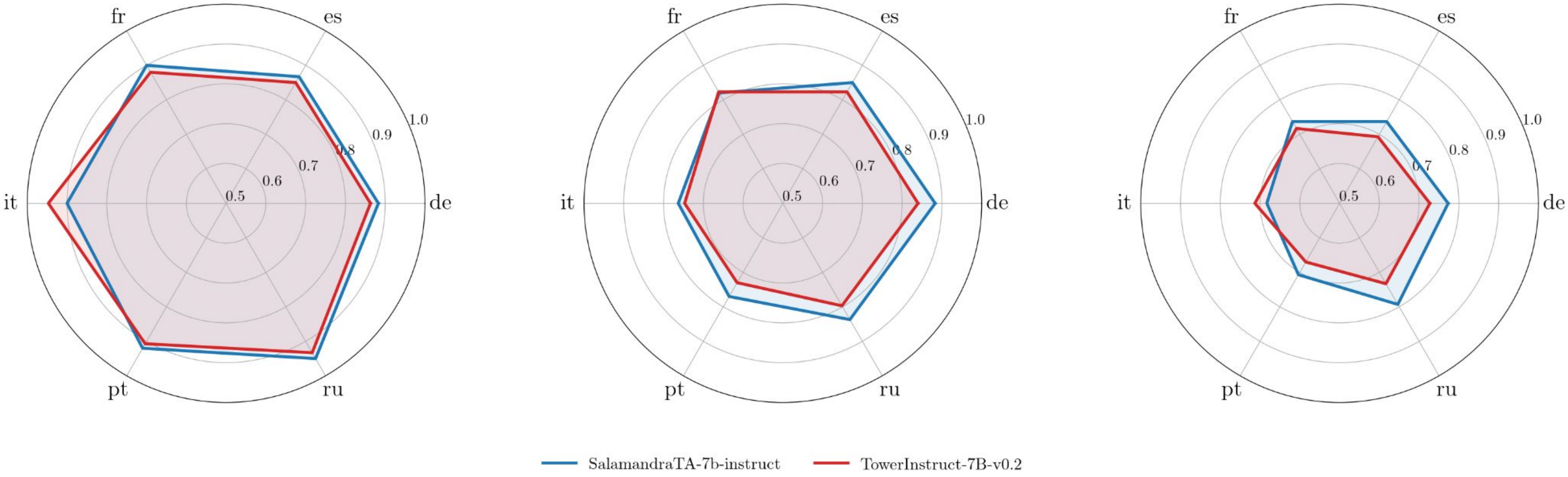
Instruction tuning data improves feminine and full-pair accuracy without sacrificing masculine accuracy

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



Results \Gender bias comparison with TowerInstruct v2.0

Masculine | Feminine | Pair



Knowledge distillation \SalamandraTA-2B

SalamandraTA-7B base is an effective teacher model

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

We run 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.

[Hinton, et al.] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

/ Takeaways /

1. We are better than Tower

/ Takeaways /

1. *How can we improve translation quality in zero-shot directions after continual pre-training?*
   Multi-Parallel data improves translation in zero-shot directions

2. *What is the impact of non-MT tasks in instruction tuning for MT quality?*
   We can add non-MT tasks with a minimal translation quality drop

3. *Does adding gender-bias instructions help improve gender accuracy?*
   Yes

4. *Does instruction tuning make the base model more robust to word-level synthetic errors?*
   Yes

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**