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 SALAMANDRA**TA** 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. SALAMANDRA**TA** 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 SALAMANDRA**TA**. 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 SALAMANDRA**TA**, along with the newer SALAMANDRA**TA**-V2 model, on Hugging Face[1].

1 Introduction

Traditionally, Massively Multilingual Neural Machine Translation (MMNMT) relied on the encoder-decoder architecture to translate across multiple languages (Fan et al., 2021; NLLB Team et al., 2022). More recently, however, Large Language Models (LLMs) have demonstrated strong MMNMT capabilities (Zhu et al., 2024) and thus some works have proposed several strategies to improve 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 initialized to the average of all existing embeddings and then rapidly optimized through continual pre-training (CPT). This method has not only proven to be viable but also demonstrably improves the model's overall performance in the target languages, even if the original model was never exposed to data from these languages during its initial training (Da Dalt et al., 2024).

Throughout this paper, we present the SALAMANDRA**TA** family of models, which serve as the backbone models of the BSC team's submission to the WMT25 General Machine Translation Shared Task. Our participation covers 15 out of the 16

[*]Core Contributor.
[1]SALAMANDRA**TA**7B-V1 , SALAMANDRA**TA**2B-V1 and SALAMANDRA**TA**7B-V2 .

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 614–637
November 8–9, 2025.

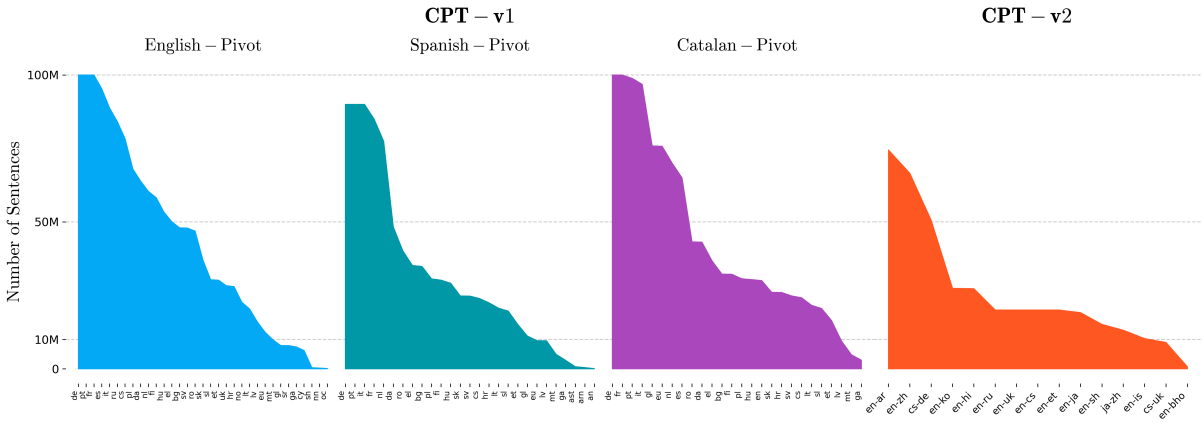

Figure 1: Distribution of sentence pairs for continual pre-training. The first three plots (■ **CPT-v1**) show the number of sentence pairs pivoting in English, Spanish and Catalan, respectively. The fourth plot (■ **CPT-v2**) corresponds to the second continual pre-training phase with direct language pairs.

translation directions in the general MT task under the constrained track. Additionally, we took part in the multilingual subtask for 7 out of the 16 directions. Contributions of this work are listed as follows:

- While most previous work have relied on English-centric parallel corpora for building translation-focused LLMs, we build SALAMANDRA**TA** pivoting in three languages for continual pre-training; English, Spanish and Catalan across 172 supervised directions.
- We show that instruction tuning improves both translation quality and robustness to character-level noise.
- We release all model checkpoints to facilitate reproducibility and future research on massively multilingual machine translation.

2 Data

Our base models are SALAMANDRA-2B and SALAMANDRA-7B (Gonzalez-Agirre et al., 2025), which were trained from scratch on highly multilingual data. However, SALAMANDRA models were not exposed to parallel data during pre-training. To address this, and following Alves et al. (2024), we improve their multilingual machine translation capabilities by performing continual pre-training on parallel data covering 38 European languages (35 of which were already present in the original pre-training corpus). This step is followed by supervised fine-tuning using high-quality instruction data. In this section, we detail the datasets used for both continual pre-training and supervised fine-tuning.

2.1 Continual pre-training

To train the SALAMANDRA**TA** models, we first compile a parallel corpus from publicly available data sources. A comprehensive list of these sources and the corresponding language pairs can be found in Table 5. We build two separate training sets: **CPT-v1** and **CPT-v2**. All data undergo initial filtering using LABSE (Feng et al., 2022), and off-target translations are excluded using the Lingua[2] library. After filtering, the data is de-duplicated and punctuation is normalized with the `Bifixer` library (Ramírez-Sánchez et al., 2020). The final corpora are formatted using the prompt template provided in Appendix Figure 3. Additional dataset details are available in Appendix C.

Using **CPT-v1** we continue pre-training SALAMANDRA 2B and 7B with the causal language modeling objective resulting in SALAMANDRA**TA**2B-BASE and SALAMANDRA**TA**7B-BASE models. Then, we use **CPT-v2** to continue pre-training SALAMANDRA**TA**7B-BASE.

■ **CPT-v1**: The first corpus, is employed during the initial round of continual pre-training (CPT), with the objective of enhancing the machine translation capabilities of SALAMANDRA across European languages. The final dataset has 38 languages across 6.57B sentence pairs and 172 machine translation directions in total pivoting in English, Spanish and Catalan, totaling in 424B tokens. We show in Figure 1 the data distribution of the **CPT-v1** corpus.

[2] https://github.com/pemistahl/lingua-py

■ **CPT-v2**: The second corpus, is used in the subsequent CPT round, where the focus shifts toward expanding coverage to include the additional language pairs featured in the WMT 2025 shared task. It includes 0.39B sentences across 14 languages and 15 directions, amounting to 27B tokens. To avoid the risk of catastrophic forgetting, we subsample 20M sentences for directions already present in **CPT-v1** (EN→CS, EN→ET, EN→RU, EN→UK). For EN→SH (English-to-Serbian, Latin script), we combined two sources from **CPT-v1**: English–Serbian (Latin script) data and English–Serbian (Cyrillic script) data, the latter converted to Latin script using rule-based transliteration. The per-direction data distribution is also shown in Figure 1. Note that we include the English-to-Hindi direction, which is not part of this year's shared task, in order to support better transfer for related languages such as Bhojpuri.

2.2 Instruction tuning

For instruction tuning we build two separate corpora: **IT-v1** and **IT-v2**. The first, **IT-v1**, is used to fine-tune SALAMANDRA**TA**2B-BASE and SALAMANDRA**TA**7B-BASE models into instruction-following models. The second corpus, is used to instruct SALAMANDRA**TA**7B-BASE after continue pre-training with **CPT-v2** corpus. We format each instruction using the `chatml` template (Open AI, 2023).

■ **IT-v1**: Following prior work on supervised fine-tuning for machine translation (Alves et al., 2024; Rei et al., 2024, 2025), we organize the instruction examples into three categories: pre-translation, translation, and post-translation tasks. The selection of tasks is motivated by the ablation results discussed in Section 4. The final corpus consists of 135k instructions, with the majority sourced from the TOWERBLOCKS collection (Alves et al., 2024). For translation related-tasks we focus on sentence, paragraph and document level data, primarily sourced from EUROPARL (Koehn, 2005). A big part of the data is drawn from multi-parallel datasets such as FLORES-200 (NLLB Team et al., 2022) or NTREX (Federmann et al., 2022), where a single source sentence has multiple translations in different target languages. When building the instruction tuning dataset, a naive strategy is to pivot trough different bridge languages across all languages including the complete dataset (e.g. for a given Catalan sentence that aligns to parallel sentences in, Spanish, French, and German, we might generate CA→ES, CA→FR, CA→DE and ES→CA, FR→CA, DE→CA). In our dataset we pivoted in five bridge languages: English, Catalan, Spanish, Basque and Galician across all the supported languages. However, this increases the number of duplicate training examples that share identical content on the target or source side. We found that doing this encourages target-side collapse, where the model produces off-target translations because many-to-one alignments blur the mapping between specific source inputs and their intended target languages. To mitigate this, we randomly sampled approximately equal numbers of translation instructions for each language pair. Further details on **IT-v1** are provided in Appendix C.

■ **IT-v2**: The second corpus, consisting of approximately 51k instructions, is constructed to focus on paragraph-level translation, context-aware machine translation, and sentence-level translation for the language directions included in the WMT 2025 shared task. To construct paragraph level data we source from FLORES-200-dev, NTREX and NEWSCOMMENTARY datasets. Similar to **IT-v1**, we applied random sampling when using multi-parallel datasets. In addition, we included data from TOWERBLOCKS that we considered relevant to our tasks. More details about **IT-v2** can be found in Appendix C.

3 SalamandraTA Models

The SALAMANDRA**TA** family is composed of two base models, 2B and 7B parameters, which were continually pre-trained on the **CPT-v1** corpus and subsequently instruction-tuned on **IT-v1**. For our submission to the WMT25 General Translation Shared Task, we further adapted the 7B model, resulting in SALAMANDRA**TA**-V2.

3.1 Adding WMT languages: SALAMANDRATA-V2

To expand the language coverage of SALAMANDRA**TA** and accommodate the additional languages required by the WMT25 General Translation Shared Task, we implemented vocabulary adaptation. We trained a new tokenizer on a corpus comprising the original languages augmented with monolingual text for the new languages not included in the original SALAMANDRA tokenizer: Chinese, Korean, Japanese, Arabic, and Bhojpuri.

The old tokenizer was replaced with the new one, which required re-initializing the embedding and unembedding layers. To address this, we modified these layers to ensure that tokens common to both the old and new tokenizers retained their original embeddings. The embeddings for the remaining, newly introduced tokens were initialized as the average of all existing embeddings. We expected this strategy to be particularly successful given that the two tokenizers share over 58% of their vocabulary. Figure 7 shows the fertility per language pair, comparing our new SALAMANDRA tokenizer against previous tokenizer, MADLAD400 and NLLB. On average, SALAMANDRA achieves a fertility of 1.88, outperforming both NLLB (2.00) and MADLAD400 (2.33) on WMT25 language pairs.

The subsequent section details the continual pre-training stage of our model. This stage aims not only to enhance the model's translation capabilities but also to recover the embeddings of these newly initialized tokens. More details can be found in Appendix D.

3.2 Model training

3.2.1 Continual pre-training

For this phase, we chose SALAMANDRA-2B and SALAMANDRA-7B as base models, using checkpoints preceding the annealing phase described in Gonzalez-Agirre et al. (2025). This choice was intentional: the annealing phase narrows the data sources to shape the model into a general-purpose downstream performer, which we considered misaligned with (or even counterproductive to) our goal of improving translation capabilities. The training strategy followed a schedule similar to that of the annealing phase. The learning rate was linearly warmed up over the first 2,000 steps, reaching a peak of 3e-5, and then decayed using a cosine schedule down to 3e-6. To mitigate the risk of exploding gradients, we applied gradient clipping with a maximum norm of 1.0 after the warm-up stage. We used NVIDIA NeMo as the training framework, and all other training hyperparameters were kept consistent with those used in the original SALAMANDRA pre-training (see Appendix E for more details). We trained the 7B model for 105k steps and the 2B model for 50k steps on the **CPT-v1** corpus tokenized with the original SALAMANDRA tokenizer (see Appendix Figure 10).

After vocabulary adaptation, we continually pre-train the resulting SALAMANDRA**TA**-7B model using **CPT-v2**. The training strategy followed the same training configuration as previously described.

3.2.2 Supervised Fine-tuning

We fine-tune SALAMANDRA**TA** base models using `FastChat` framework (Zheng et al., 2023). Hyperparameter details are provided in Appendix Table 10.

3.3 Evaluation

Metrics We assess translation quality using several metrics. For reference-based evaluation, we report scores from the learned metrics COMET (Rei et al., 2022a), BLEURT (Sellam et al., 2020), and METRICX (Juraska et al., 2023). For reference-free quality estimation (QE), we use COMET-KIWI (Rei et al., 2022b), and METRICX-QE. We also report two lexical-based metrics: CHRF (Popović, 2015) and BLEU (Papineni et al., 2002).

Datasets We used the FLORES-200-devtest dataset for ablation studies on the SALAMANDRA**TA** models. For evaluating translation quality on the WMT 2025 directions, we primarily relied on the WMT24++ dataset (Deutsch et al., 2025). An exception is the English to Bhojpuri direction, which is not included in WMT24++; for this case, we used FLORES-200-devtest for evaluation.

Baselines We compare the different SALAMANDRA**TA** variants against the translation LLM TOWER-V2 7B (Rei et al., 2024), as well as dedicated MMNMT models such as MADLAD400 7B (Kudugunta et al., 2023) and NLLB 3.3B (NLLB Team et al., 2022).

Decoding strategies For inference with the baseline, base, and instruction-tuned models, we employ beam search with a beam size of 5. Additionally, we experiment with two alternative decoding approaches: we use diverse beam search (Vijayakumar et al., 2018), which promotes output diversity by penalizing similar beams, and two post-decoding strategies applied to the generated candidates: Tuned Re-ranking Decoding (TRR) and Minimum Bayes Risk Decoding (MBR) (Eikema and Aziz, 2020) using the `mbrs` library (Deguchi et al., 2024). For diverse beam search we set a beam size of 20 and 5 beam groups. For post-decoding methods, we use COMET-22 as the quality metric for MBR and COMET-KIWI for TRR.

	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	-
+ 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**

Table 1: COMET scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models, and finally with the application of quality-aware decoding strategies (TRR and MBR). Using Minimum Bayes Risk (MBR) decoding consistently yields the best results.

4 Results

Table 1 presents the main translation quality results on the WMT24++ test set, measured in COMET scores for the language directions in the general MT task. We report extra metrics in Appendix F. We additionally evaluate SALAMANDRA**TA**-2B and SALAMANDRA**TA**-7B using COMET and METRICX for the language directions present in the multilingual subtask and report them in Appendix Table 17.

As shown in Table 1, instruction tuning yields significant gains over the CPT baselines, improving the SALAMANDRA**TA**-7B, SALAMANDRA**TA**-2B, and SALAMANDRA**TA**-V2 models by an average of 3.51, 4.40, and 3.60 COMET points, respectively.

Although further adapting the SALAMANDRA**TA**-7B model to WMT-2025 language pairs initially causes an average performance drop of 1.09 COMET points on the language directions shared between SALAMANDRA**TA**-7B and SALAMANDRA**TA**-V2, this gap is largely mitigated when employing quality-aware decoding strategies. Applying Minimum Bayes Risk (MBR) and Tuned Re-ranking (TRR) decoding strategies reduces this drop to 0.16 and 0.20 COMET points, respectively.

On the impact of adding non-MT-Tasks To better understand the impact of different instruction types on translation quality, we conduct an ablation study of instruction fine-tuning across four main task categories: machine translation (MT), pre-translation tasks (Pre-MT) (e.g., Named Entity Recognition), post-translation tasks (Post-MT) (e.g., Gender Bias Mitigation), and chat/code-related tasks[3]. Table 2 presents the model's performance after fine-tuning on each of these categories.

Instruction fine-tuning using MT tasks consistently yields the best overall performance across most evaluation metrics, with the exception of METRICX. For METRICX, a combination of MT, Pre-MT, and Post-MT instructions results in slightly improved performance. In contrast, adding only Pre-MT or Post-MT instructions shows no significant difference compared to the MT-only baseline. Incorporating Chat and Code instructions,

[3]This last group includes TOWERBLOCKS synthetic chat data and code instruction data.

	en→xx			xx→en		
	COMET	METRICX	BLEU	COMET	METRICX	BLEU
SALAMANDRA**TA**7B BASE + CPT-V1	0.85	1.73	34.60	**0.88**	1.15	44.22
Supervised Finetuning						
MT	**0.87**	1.33	**36.71**	**0.88**	1.17	**45.02**
+ Pre-MT + Post-MT	**0.87**	**1.14**	36.42	**0.88**	**1.09**	45.00
+ Chat + Code	**0.87**	1.36	35.58	**0.88**	1.16	44.81
MT + Post-MT	**0.87**	1.33	36.57	**0.88**	1.15	44.88
MT + Pre-MT	**0.87**	1.33	36.34	**0.88**	1.16	44.67

Table 2: Ablation study on the impact of different supervised fine-tuning tasks for the SALAMANDRA**TA**7B-BASE model. We report COMET, METRICX, and BLEU scores for English-to-Other (en→xx) and Other-to-English (xx→en) directions.

however, leads to a consistent drop in BLEU scores without measurable gains in other metrics.

Based on these findings, we concluded that for SALAMANDRA**TA**-2B and 7B, incorporating both Pre-MT and Post-MT tasks alongside MT tasks provided a slight benefit or at least no degradation in performance, leading to their inclusion in the **IT-v1** dataset. However, for SALAMANDRA**TA**-V2 which was specifically tailored for the WMT25 General Translation Shared Task, we made a deliberate choice to focus exclusively on MT instructions. While Pre-MT and Post-MT tasks might offer benefits, gathering high-quality, task-specific instruction data for the unique language pairs and domains present in WMT25 would have required significant additional effort beyond the scope of this work.

On the robustness to character noise Following Peters and Martins (2025), we investigate model robustness by injecting character-level noise into the source sentences of FLORES-200-devtest for the English to Spanish direction using adjacent swaps, duplications, and deletions at different noise levels. Figure 2 shows the relative degradation in BLEU score compared to zero-noise baseline. The SALAMANDRA**TA** 7B instruction-tuned model consistently shows greater resilience than the base model across all perturbation types. At the maximum noise level (1.0), the performance degradation of the instruction-tuned model is smaller by 17.63 p.p. for swaps, 20.61 p.p. for duplications, and 18.33 p.p. for deletions. These results demonstrate that instruction tuning effectively improves a model's robustness to character-level input corruptions.

Adding a low-resource language: The case of Bhojpuri Table 3 presents our ablation experiments for English to Bhojpuri translation direction. 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. Finally, supervised fine-tuning (IT-V2) improves performance, improving the scores to 11.67 BLEU and 37.75 CHRF. This result shows the effectiveness of fine-tuning on high-quality data in the final stage, even for low-resource language pairs.

	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

Table 3: Ablation results for English→Bhojpuri translation in terms of BLEU and CHRF on FLORES-200-devtest. The table compares the impact of removing the EN→HI direction from the CPT data and the effect of supervised fine-tuning (**IT-V2**).

5 Submission

For our WMT25 general and multilingual MT tasks submissions, we apply a chunking strategy, splitting each input instance at \n\n delimiter prior to translation. We made two submissions using

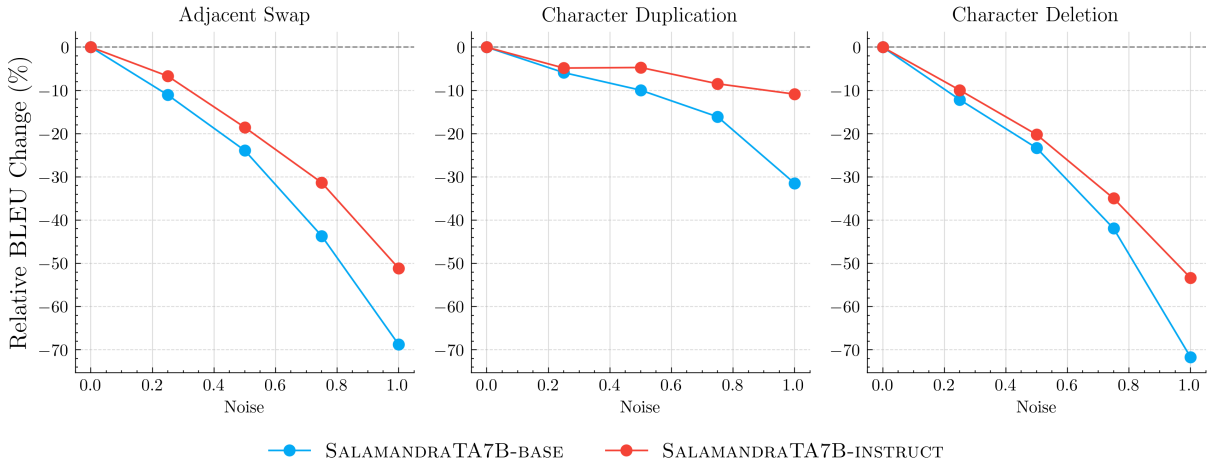

Figure 2: Relative change in BLEU scores (%) under increasing levels of input noise for three types of character-level perturbations: Adjacent Swap, Character Duplication, and Character Deletion.

two quality-aware decoding strategies: Minimum Bayes Risk Decoding employing COMET and Tuned Re-ranking relying on COMET-Kiwi.

6 Conclusion

In this paper, we introduced the SALAMANDRA**TA** family of models, a series of powerful, translation LLMs in 2B and 7B scales. Our approach combines a multi-stage training recipe, beginning with continual pre-training on parallel data that pivots through three languages: English, Spanish, and Catalan. This is followed by an instruction tuning stage to align the models with human translation outputs. For our WMT25 submission, we adapted our 7B model to new, non-European languages through vocabulary adaptation and a further round of continual pre-training and supervised fine-tuning.

Our experimental results show that instruction tuning is a critical step which not only improves translation quality but also the model's robustness against character-level noise. Furthermore, our analysis of the English-to-Bhojpuri direction validates the importance of including related languages during pre-training to enable cross-lingual transfer to low-resource pairs.

While our work successfully specializes models for translation and translation-related tasks, we observed that incorporating Chat and Code instructions during the supervised fine-tuning stage leads to a significant drop in translation quality as measured by BLEU. Future work could explore methods to mitigate this trade-off to train machine translation models that can follow general instructions without compromising their specialized translation capabilities.

7 Acknowledgements

This work has been promoted and financed by the Generalitat de Catalunya through the Aina Project.

This work has been supported by the Spanish project PID2021-123988OB-C33 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE.

This work is partially supported by MLLM4TRA (PID2024-158157OB-C32) funded by MCIN/AEI/10.13039/501100011033/FEDER, UE.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU – NextGenerationEU within the framework of ILENIA Project with reference 2022/TL22/00215337.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, and 1 others. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

CASMACAT. 2018. Global Voices Parallel Corpus 2018Q4. Accessed: July, 2025.

Severino Da Dalt, Joan Llop, Irene Baucells, Marc Pamies, Yishi Xu, Aitor Gonzalez-Agirre, and Marta Villegas. 2024. FLOR: On the effectiveness of language adaptation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7377–7388, Torino, Italia. ELRA and ICCL.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mbrs: A library for minimum Bayes risk decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–362, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

ELRC-Share. 2020. Bilingual corpus made out of pdf documents from the european medicines agency (emea). .

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

878–891, Dublin, Ireland. Association for Computational Linguistics.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024. PILAR: A collection of low-resource language corpora from the iberian peninsula. https://github.com/transducens/PILAR.

Pablo Gamallo, Marcos Garcia, Iria de-Dios-Flores, José Ramom Pichel Campos, Sandra Rodríguez Rey, and Daniel Bardanca. 2023a. NÓS corpus: Authentic English–Galician Parallel Corpus. Zenodo, https://doi.org/10.5281/zenodo.7675110. Accessed: July 2025.

Pablo Gamallo, Marcos García, Iria de Dios-Flores, José Ramom Pichel Campos, Sandra Rodríguez Rey, and Daniel Bardanca. 2023b. Nós corpus: Synthetic English–Galician Parallel Corpus. Zenodo, https://doi.org/10.5281/zenodo.7685180. Accessed: July 2025.

Mercedes García-Martínez, Laurent Bié, Aleix Cerdà, Amando Estela, Manuel Herranz, Rihards Krišlauks, Maite Melero, Tony O'Dowd, Sinead O'Gorman, Marcis Pinnis, Artūrs Stafanovič, Riccardo Superbo, and Artūrs Vasiļevskis. 2021. Neural translation for European Union (NTEU). In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 316–334, Virtual. Association for Machine Translation in the Americas.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 5 others. 2025. Salamandra technical report. *Preprint*, arXiv:2502.08489.

Kenneth Heafield, Elaine Farrow, Jelmer van der Linde, Gema Ramírez-Sánchez, and Dion Wiggins. 2022. The EuroPat corpus: A parallel corpus of European patent data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 732–740, Marseille, France. European Language Resources Association.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, Berlin, Germany. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Open AI. 2023. [link].

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.

Ben Peters and Andre Martins. 2025. Did translation models get more robust without anyone Even noticing? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2445–2458, Vienna, Austria. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the*

Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Project Ilenia. 2024. GAITU Corpus: Catalan–Basque Synthetic Parallel Sentences. Hugging Face dataset, published approx. Dec 2024. Accessed: 2025-07-20; license: CC BY-NC-SA 4.0.

Projecte Aina-Language Technologies Unit, BSC. 2024. CA–EN Parallel Corpus: Catalan–English Synthetic Parallel Sentences. Hugging Face dataset, DOI:10.57967/hf/1913. Accessed: 2025-07-20; license: CC BY 4.0.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *Preprint*, arXiv:2506.17080.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 454–459, Istanbul, Turkey. European Language Resources Association (ELRA).

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

The GNOME Project. n.d. GNOME. Accessed: July, 2025.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models. *Preprint*, arXiv:1610.02424.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. 2025. Exploring translation mechanism of large language models. *Preprint*, arXiv:2502.11806.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *Preprint*, arXiv:2408.10811.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A CPT Template

This section presents the template used to prepare parallel data for continued pre-training. We used only one single template. Placeholders:

- { source }: source sentence
- { target }: target sentence
- { source_lang }: source language name
- { target_lang }: target language name

Template used for CPT

```
{ source_lang }: { source }
{ target_lang }: { target }
```

Figure 3: Template used to format parallel data for CPT.

B Prompt templates used to construct translation instructions

All templates used to construct instructions were adapted from TowerBlocks (Alves et al., 2024). Figure 4 shows an example of a template used for translation instructions in our **IT-v1** and **IT-v2** datasets.

C Dataset

C.1 Continual pre-training v1

The pre-training corpus for **CPT-v1** consists of 424 billion tokens of Catalan-centric, Spanish-centric, and English-centric parallel data, including all of the official European languages plus Catalan, Basque, Galician, Asturian, Aragonese and Aranese. It amounts to 6,574,251,526 parallel sentence pairs.

This highly multilingual corpus is predominantly composed of data sourced from OPUS (Tiedemann, 2012), with additional data taken from the NTEU Project (García-Martínez et al., 2021), Aina Project,[4] and other sources (see Table 5, and Table 4 shows the mapping between the BCP-47 language code and the language name). Where little parallel Catalan ↔ xx data could be found, synthetic Catalan data was generated from the Spanish

[4] https://projecteaina.cat/

Template used for IT

```
Translate the following text from { source_lang } to { target_lang }:
{ source_lang }: { source }
{ target_lang }: { target }
```

Figure 4: Example of a prompt template used to construct translation instructions for **IT-v1** and **IT-v2**.

side of the collected Spanish ↔ xx corpora using Projecte Aina's Spanish-Catalan model.[5] The final distribution of languages is shown in Figure 1.

Datasets with "-BSC" in their names (e.g., BOUA-SYNTH-BSC, DOGV-SYNTH-BSC) are synthetic datasets obtained by machine translating pre-existing monolingual corpora with our own seq-to-seq models. These datasets were generated internally for model training and are not published.

C.2 Continual pre-training v2

In **CPT-v2** we focused on the language pairs featured in the WMT 2025 shared task. For pairs involving European languages, we reused part of the data from **CPT-v1**. Specifically, we sampled 20M sentence pairs each for English–Czech, English–Estonian, and English–Russian from the **CPT-v1** data. For English–Serbian (Latin), we included the authentic English–Serbian (Latin) parallel dataset from **CPT-v1**. Additionally, we transliterated the Serbian side of the English–Serbian (Cyrillic) dataset into Latin script, taking advantage of the one-to-one correspondence between the two scripts. For English–Icelandic, Czech–Ukrainian, and Czech–German, we used the WMT 2025 Translation Task Training Data.[6]

For language pairs involving non-European languages, we used sentence-level data from the WMT 2025 Translation Task Training Data. The Chinese side of all datasets were first processed using the Hanzi Identifier to detect Traditional Chinese,[7] which was subsequently converted to Simplified Chinese using OpenCC.[8] We also included paragraph-level English–Arabic data by concatenating sentences from NEWSCOMMENTARY.

We created two versions of **CPT-v2**. The first included only the language pairs featured in the WMT25 shared task. In the second, we additionally included English–Hindi data from the OPUS

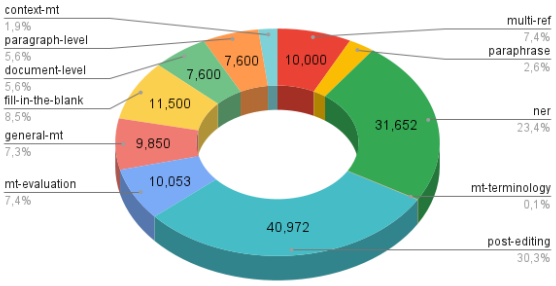

Figure 5: Distribution of tasks in **IT-v1** .

corpora CCMatrix (Schwenk et al., 2021b), MultiHPLT (de Gibert et al., 2024), NLLB (NLLB Team et al., 2022), and Samanantar (Ramesh et al., 2022), to support the model's performance on Bhojpuri (which uses the Devanagari script).

The pre-training corpus for **CPT-v2** without English-Hindi consists of 24 billion tokens, amounting to 366,179,935 parallel sentence pairs. For **CPT-v2** with English-Hindi, the corpus contains 26 billion tokens and 393,507,678 parallel sentence pairs. The data distribution is shown in Figure 1, and the corresponding sources are listed in Table 6.

As shown in Section 4, continual pre-training with Hindi data led to better performance, particularly for Bhojpuri.

C.3 Instruction tuning v1

During **IT-v1** the model was fine-tuned on ~135k instructions, primarily targeting machine translation performance for Catalan, English, and Spanish. Additional instruction data for other European and closely related Iberian languages was also included.

A portion of our fine-tuning data comes directly from, or is sampled from TOWERBLOCKS. While tasks related to machine translation are included, it is important to note that no chat data was used in the fine-tuning process. The final distribution of tasks is shown in Figure 5. The full list of tasks included in **IT-v1** is shown in Table 7.

[5] https://huggingface.co/projecte-aina/aina-translator-es-ca

[6] https://www2.statmt.org/wmt25/mtdata/

[7] https://github.com/tsroten/hanzidentifier

[8] https://github.com/BYVoid/OpenCC

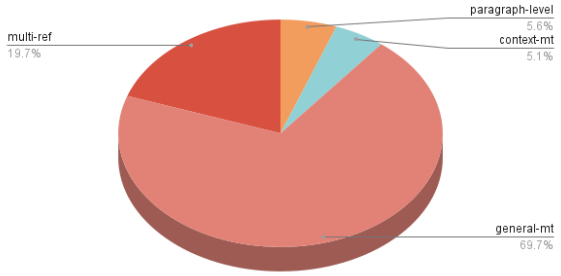

Figure 6: Distribution of tasks in **IT-v2** .

C.4 Instruction tuning v2

In **IT-v2** we focused on the languages pairs featured in the WMT 2025 shared task. We included paragraph-level data during instruction tuning to support paragraph-level translation. We constructed this data by concatenating adjacent sentences (randomly grouping 2, 3, or 4) from the same article or document in FLORES-200-dev, NTREX, and NEWSCOMMENTARY. To prevent over-representation of these sources, we sampled approximately equal amounts of paragraph-level data for each language pair. Serbian Cyrillic data from FLORES-200-dev was transliterated into Serbian Latin. In addition, we included data from TOWERBLOCKS that we considered relevant to our tasks. The instruction tuning dataset is summarized in Table 8 and the distribution of tasks is shown in Figure 6.

D Tokenizer

We evaluated the trained tokenizer using fertility metric on the FLORES-200 dataset (see Figure 7). For a given tokenizer T and a set of sentences S, fertility is defined as the ratio of the total number of tokens produced by T to the total number of words in S. Formally:

$$\text{Fertility}(T, S) = \frac{\#\text{tokens in } T(S)}{\#\text{words in } S} \quad (1)$$

The results in Figure 7 indicate that SALAMANDRA**TA**7B-V2 consistently achieves the lowest fertility scores on average among WMT25 languages.

E Training

F Results

Language Code	Language
ar	Arabic
arn	Aranese
ast	Asturian
arg	Aragonese
bho	Bhojpuri
bg	Bulgarian
ca	Catalan
cs	Czech
cy	Welsh
da	Danish
de	German
el	Greek
es	Spanish
en	English
et	Estonian
eu	Basque
fi	Finnish
fr	French
ga	Irish
gl	Galician
hi	Hindi
hr	Croatian
hu	Hungarian
is	Icelandic
it	Italian
ja	Japanese
ko	Korean
lt	Lithuanian
lv	Latvian
mt	Maltese
nl	Dutch
nn	Norwegian Nynorsk
no	Norwegian
oc	Occitan
pl	Polish
pt	Portuguese
ro	Romanian
ru	Russian
sh	Serbian (Latin)
sk	Slovak
sl	Slovenian
sr	Serbian (Cyrillic)
sv	Swedish
uk	Ukrainian
val	Catalan-Valencian
zh	Chinese

Table 4: Mapping from BCP-47 language codes to full language names.

Dataset	Ca-xx Languages	Es-xx Languages	En-xx Languages
AINA (Projecte Aina-Language Technologies Unit, BSC, 2024)	en		
ARANESE-SYNTH-CORPUS-BSC	arn		
BOUA-SYNTH-BSC		val	
BOUMH (Galiano-Jiménez et al., 2024)		val	
BOUA-PILAR (Galiano-Jiménez et al., 2024)		val	
CCMatrix (Schwenk et al., 2021b)	eu		ga
DGT (Steinberger et al., 2012)	bg, cs, da, de, el, et, fi, fr, ga, hr, hu, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	da, et, ga, hr, hu, lt, lv, mt, sh, sl	
DOGV-SYNTH-BSC		val	
DOGV-PILAR (Galiano-Jiménez et al., 2024)		val	
ELRC-EMEA (ELRC-Share, 2020)	bg, cs, da, hu, lt, lv, mt, pl, ro, sk, sl	et, hr, lv, ro, sk, sl	
EMEA (Tiedemann, 2012)	bg, cs, da, el, fi, hu, lt, mt, nl, pl, ro, sk, sl, sv	et, mt	
EUBookshop (Skadiņš et al., 2014)	lt, pl, pt	cs, da, de, el, fi, fr, ga, it, lv, mt, nl, pl, pt, ro, sk, sl, sv	cy, ga
Europarl (Koehn, 2005)		bg, cs, da, el, en, fi, fr, hu, lt, lv, nl, pl, pt, ro, sk, sl, sv	
Europat (Heafield et al., 2022)		en, hr	no
GAITU Corpus (Project Ilenia, 2024)			eu
KDE4 (Tiedemann, 2012)	bg, cs, da, de, el, et, eu, fi, fr, ga, gl, hr, it, lt, lv, nl, pl, pt, ro, sk, sl, sv	bg, ga, hr	cy, ga, nn, oc
GlobalVoices (CASMACAT, 2018; Tiedemann, 2012)	bg, de, fr, it, nl, pl, pt	bg, de, fr, pt	
GNOME (The GNOME Project, n.d.; Tiedemann, 2012)	eu, fr, ga, gl, pt	ga	cy, ga, nn
JRC-Arquis (Steinberger et al., 2006)	cs, da, et, fr, lt, lv, mt, nl, pl, ro, sv		et
LES-CORTS-VALENCIANES-SYNTH-BSC		val	
MaCoCu (Bañón et al., 2022)	en		hr, mt, uk
MultiCCAligned (El-Kishky et al., 2020)	bg, cs, de, el, et, fi, fr, hr, hu, it, lt, lv, nl, pl, ro, sk, sv	bg, fi, fr, hr, it, lv, nl, pt	bg, cy, da, et, fi, hr, hu, lt, lv, no, sl, sr, uk
MultiHPLT (de Gibert et al., 2024)	en, et, fi, ga, hr, mt	fi, ga, gl, hr, mt, nn, sr	
MultiParaCrawl (Bañón et al., 2020)	bg, da	de, en, fr, ga, hr, hu, it, mt, pt	bg, cs, da, de, el, et, fi, fr, ga, hr, hu, lt, lv, mt, nn, pl, ro, sk, sl, uk
MultiUN (Eisele and Chen, 2010)		fr	
News-Commentary (Tiedemann, 2012)		fr	
NLLB (NLLB Team et al., 2022)	bg, da, el, en, et, fi, fr, gl, hu, it, lt, lv, pt, ro, sk, sl	bg, cs, da, de, el, et, fi, fr, hu, it, lt, lv, nl, pl, pt, ro, sk, sl, sv	bg, cs, cy, da, de, el, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, no, oc, pl, pt, ro, ru, sk, sl, sr, sv, uk
NÓS Authentic Corpus (Gamallo et al., 2023a)			gl
NÓS Synthetic Corpus (Gamallo et al., 2023b)			gl
NTEU (García-Martínez et al., 2021)	bg, cs, da, de, el, en, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	da, et, ga, hr, lt, lv, mt, ro, sk, sl, sv	
OpenSubtitles (Lison and Tiedemann, 2016)	bg, cs, da, de, el, et, eu, fi, gl, hr, hu, lt, lv, nl, pl, pt, ro, sk, sl, sv	da, de, fi, fr, hr, hu, it, lv, nl	bg, cs, de, el, et, hr, fi, fr, hr, hu, no, sl, sr
OPUS-100 (Zhang et al., 2020; Tiedemann, 2012)	en		gl
StanfordNLP-NMT (Luong and Manning, 2016; Luong et al., 2015; Luong and Manning, 2015)			cs
Tatoeba (Tiedemann, 2012)	de, pt	pt	
TildeModel (Rozis and Skadiņš, 2017)	bg	et, hr, lt, lv, mt	
UNPC (Ziemski et al., 2016)		en, fr	ru
PILAR-VALENCIAN-AUTH (Galiano-Jiménez et al., 2024)		val	
PILAR-VALENCIAN-SYNTH (Galiano-Jiménez et al., 2024)		val	
WikiMatrix (Schwenk et al., 2021a)	bg, cs, da, de, el, et, eu, fi, fr, gl, hr, hu, it, lt, nl, pl, pt, ro, sk, sl, sv	bg, en, fr, hr, it, pt	oc, sh
Wikimedia			cy, nn
XLENT (El-Kishky et al., 2021)	eu, ga, gl	ga	cy, et, ga, gl, hr, oc, sh

Table 5: Data sources of **CPT-v1**.

Source	Language Pair
WMT 2025 Translation Task Training Data	en-ar en-zh cs-de en-ko en-ja ja-zh en-is cs-uk en-bho
NEWSCOMMENTARY (paragraph-level)	en-ar
CCMATRIX (Schwenk et al., 2021b)	en-hi
MULTIHPLT (de Gibert et al., 2024)	en-hi
NLLB (NLLB Team et al., 2022)	en-hi
SAMANANTAR (Ramesh et al., 2022)	en-hi
CPT-v1	en-cs en-et en-ru en-uk en-sh

Table 6: Data sources of **CPT-v2**.

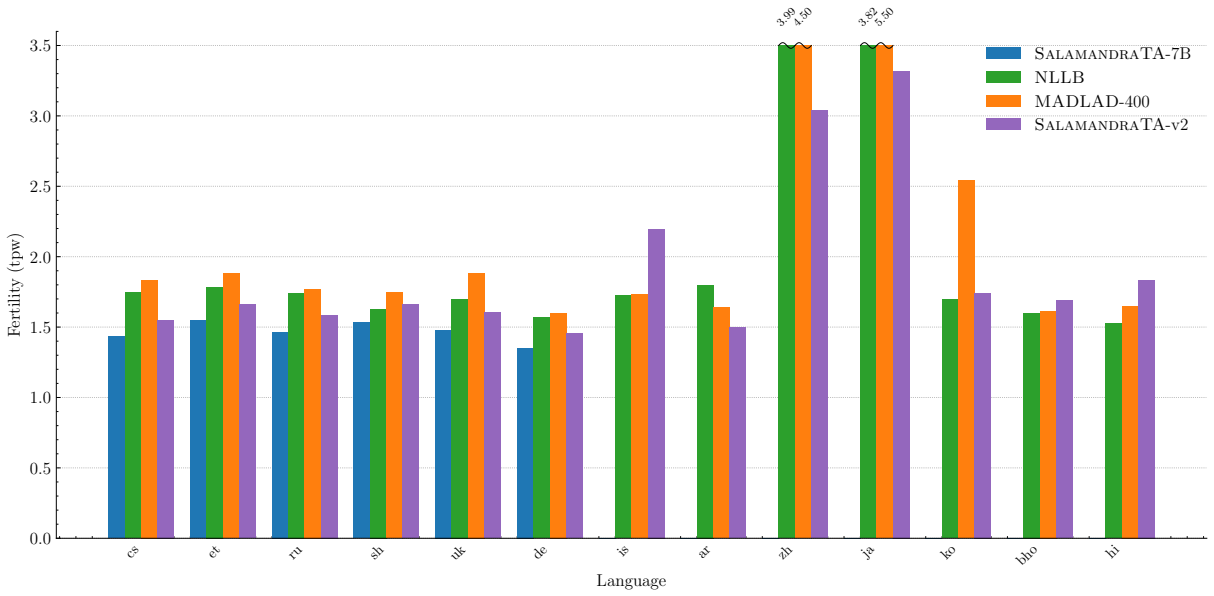

Figure 7: Tokenization fertility comparison across 13 languages from the FLORES-200 dataset. Fertility is shown on the vertical axis for each language on the horizontal axis. Results are presented for four multilingual models: SALAMANDRATA-7B, NLLB, MADLAD-400, and SALAMANDRATA-V2.

Category	Task	Source	Languages	Count
Pre-Translation	Named-entity Recognition	ANCORA-CA-NER	ca	12,059
		BASQUEGLUE, EUSIE	eu	4,304
		SLI NERC Galician Gold Corpus	gl	6,483
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	pt	854
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	nl	800
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	es	1,654
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	en	1,671
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	ru	800
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	it	858
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	fr	857
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	de	1,312
Translation	Multi-reference Translation	TOWERBLOCKS: TATOEBA Dev	mixed	10,000
	Terminology-aware Translation	TOWERBLOCKS: WMT21 TERMINOLOGY DEV	en-ru	50
		TOWERBLOCKS: WMT21 TERMINOLOGY DEV	en-fr	29
	Fill-in-the-Blank	Non-public	Five pivot languages (ca, es, eu, gl, en) paired with European languages (cs, da, de, el, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv)	11,500
	General Machine Translation	TOWERBLOCKS: WMT14 to WMT21, NTREX, FLORES DEV, FRMT, QT21, APEQUEST, OPUS (Quality Filtered), MT-GENEVAL	nl-en, en-ru, it-en, fr-en, es-en, en-fr, ru-en, fr-de, en-nl, de-fr	500
		FLORES DEV, NTREX	Four pivot languages (es, ca, eu, gl) paired with the rest of languages. We sample 50 instances for each pair.	9350
	Document-level Translation	Non-public	Two pivot languages (es, en) paired with European languages (bg, cs, da, de, el, et, fi, fr, hu, it, lt, lv, nl, pl, pt, ro, ru, sk, sv)	7,600
	Paragraph-level Translation	Non-public	Two pivot languages (es, en) paired with European languages (bg, cs, da, de, el, et, fi, fr, hu, it, lt, lv, nl, pl, pt, ro, ru, sk, sv)	7,600
	Context-Aware Translation	TOWERBLOCKS: MT-GENEVAL	en-it	348
			en-ru	454
			en-fr	369
			en-nl	417
			en-es	431
			en-de	558
Post-Translation	Paraphrase	TOWERBLOCKS: PAWS-X DEV	mixed	3,521
	Machine Translation Evaluation	TOWERBLOCKS (sample): WMT20 to WMT22 METRICS MQM, WMT17 to WMT22 METRICS DIRECT ASSESSMENTS	en-ru, en-pl, ru-en, en-de, en-ru, de-fr, de-en, en-de	353
		Non-public	Four pivot languages (eu, es, ca, gl) paired with European languages (bg, cs, da, de, el, en, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv)	9,700
	Automatic Post Editing	TOWERBLOCKS: QT21, APEQUEST	en-fr	6,133
		TOWERBLOCKS: QT21, APEQUEST	en-nl	9,077
		TOWERBLOCKS: QT21, APEQUEST	en-pt	5,762
		TOWERBLOCKS: QT21, APEQUEST	de-en	10,000
		TOWERBLOCKS: QT21, APEQUEST	en-de	10,000
Total				**135,404**

Table 7: Overview of tasks, data sources, language coverage, and counts in **IT-v1**.

Category	Task	Source	Languages	Count
Translation	Paragraph-level Translation	FLORES DEV	en-ar	30
			en-bho	30
			en-ja	30
			en-uk	30
			en-ru	21
			cs-uk	30
			ja-zh	30
			en-zh	30
			en-ko	30
			en-et	30
			en-is	30
			en-sh	30
			en-cs	30
			cs-de	30
		NTREX	en-ja	58
			en-uk	58
			en-ru	50
			cs-uk	58
			ja-zh	58
			en-zh	58
			en-ko	58
			en-et	58
			en-is	58
			en-sh	58
			en-cs	58
			cs-de	58
		NEWS COMMENTARY	en-zh	250
			cs-de	250
			en-cs	250
			en-de	250
			en-ja	250
			ja-zh	250
			en-ru	250
	Context-Aware Translation	TOWERBLOCKS: MT-GENEVAL	en-it	348
			en-fr	369
			en-nl	417
			en-es	431
			en-de	558
			en-ru	454
	Multi-reference Translation	TOWERBLOCKS: TATOEBA Dev	mixed	10,000
	General Machine Translation	TOWERBLOCKS: WMT14 to WMT21, NTREX, FLORES DEV, FRMT, QT21, APE-QUEST, OPUS (Quality Filtered), MT-GENEVAL	en-ru	22,112
			en-zh	10,521
			en-ko	2,782
Total				**50,841**

Table 8: Overview of tasks, data sources, language coverage, and counts in **IT-v2**.

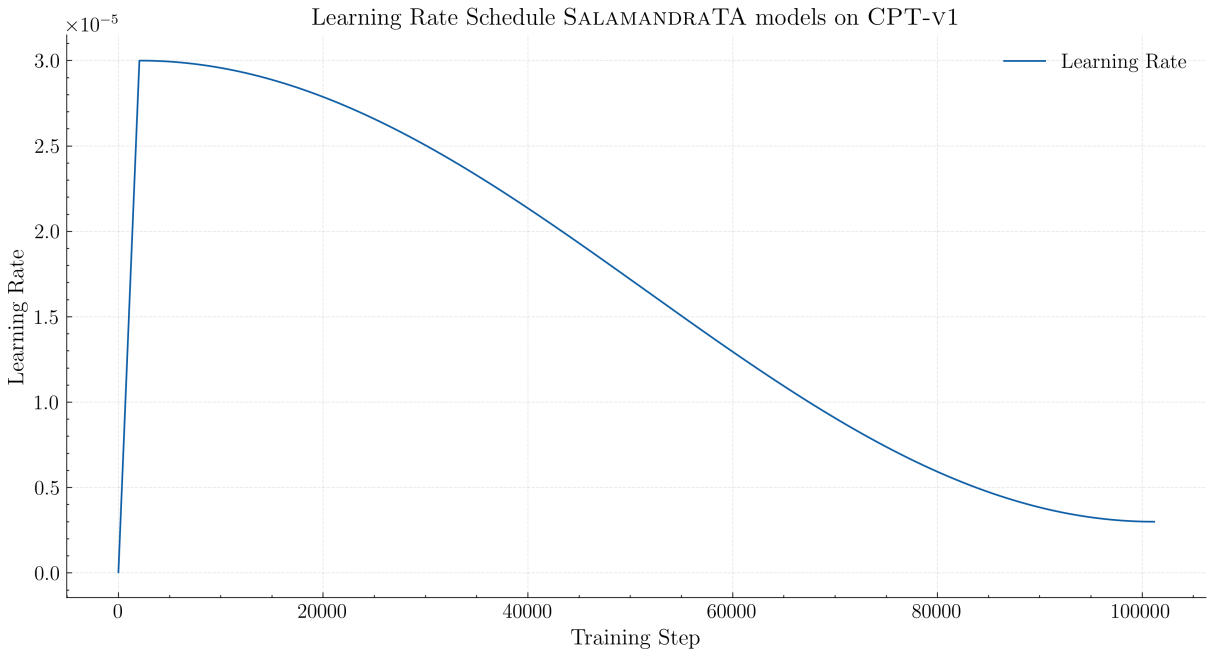

Figure 8: Learning Rate for the SalamandraTA-7B and SalamandraTA-2B on **CPT-v1**.

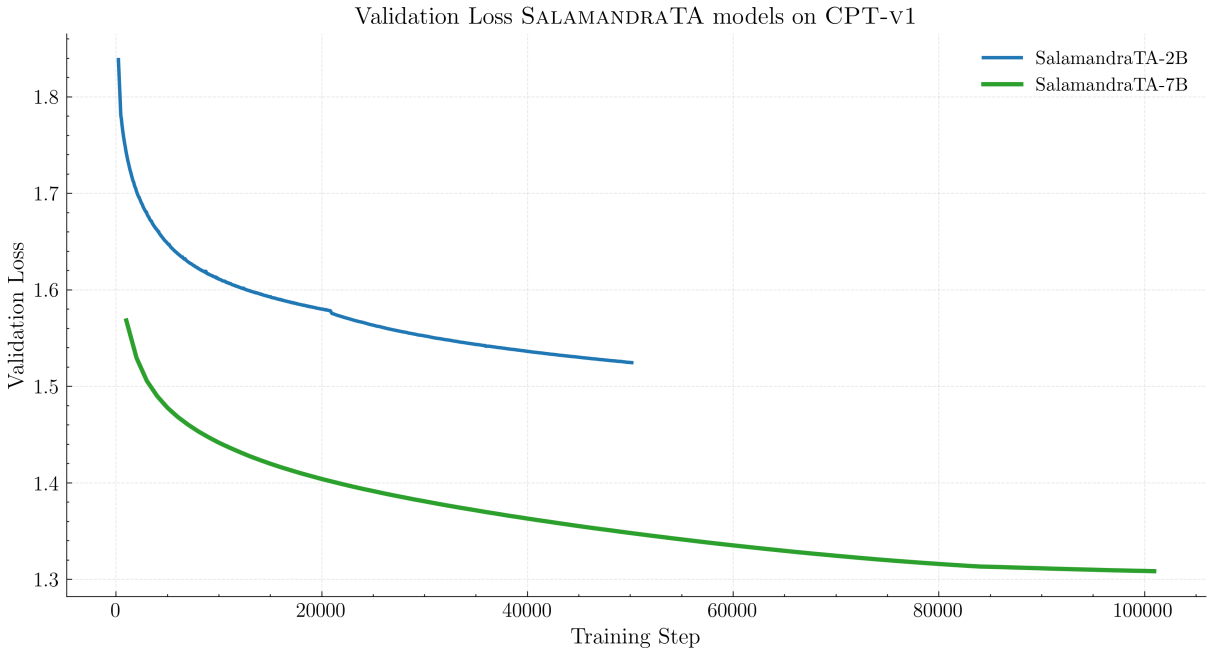

Figure 9: Validation loss for the SalamandraTA-7B and SalamandraTA-2B on **CPT-v1**.

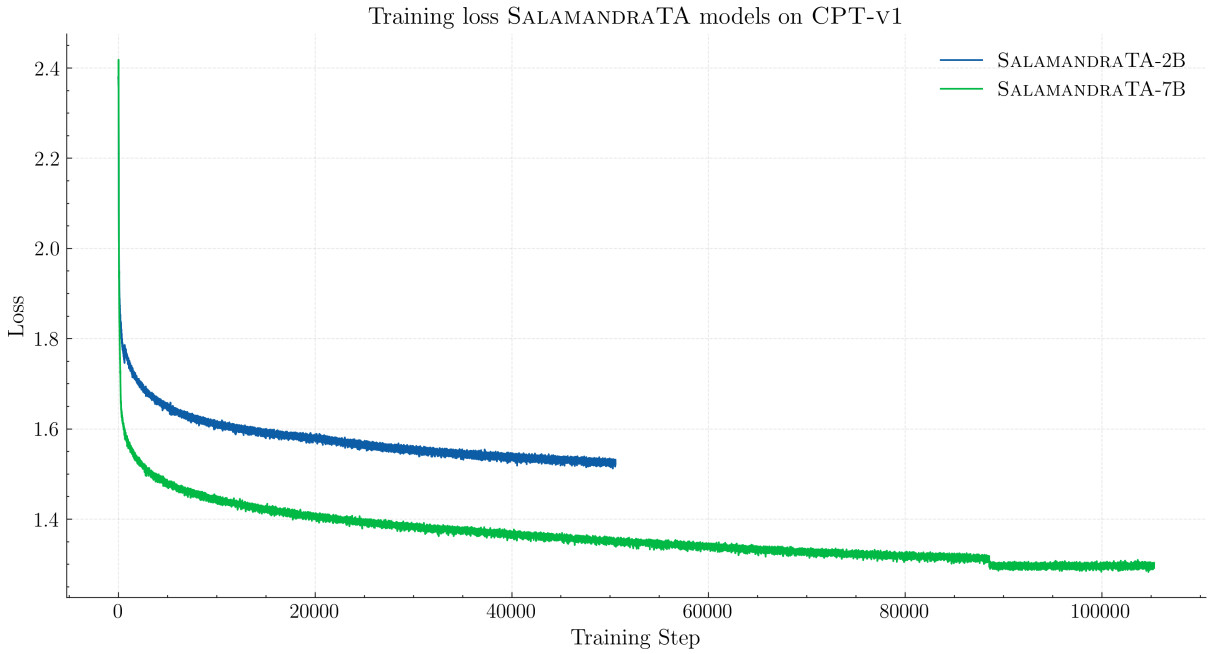

Figure 10: Training loss for the SALAMANDRA**TA**-7B and SALAMANDRA**TA**-2B on **CPT-v1**.

Table 9: Hyperparameters for SALAMANDRA**TA** continual pre-training.

Hyperparameter	**Value**
Micro Batch Size	2
Global Batch Size	512
Optimizer	Distributed Fused Adam
Learning Rate	3e-5
Minimum LR	3e-6
Weight Decay	0.1
Betas	(0.9, 0.95)
LR Scheduler	CosineAnnealing
Warmup Steps	2048
Mixed Precision	AMP O2
Sequence Length	8,192
Gradient Sync DType	bfloat16

Table 10: Hyperparameters for SALAMANDRA**TA** supervised-fine tuning.

Hyperparameter	**Value**
Train epochs	1
Train batch size per device	1
Gradient accumulation steps	16
Learning rate	1e-5
Weight decay	0
Warmup ratio	0.03
LR scheduler	Cosine
Model max length	8,192

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	11.1	-	21.2	-	-	-	-	35.6	-	24.7	18.4	-	-
MADLAD400 7B	28.4	27.2	22.5	-	26.8	17.5	6.9	30.2	19.8	25.3	25.2	20.9	20.8
NLLB 3.3B	23.0	21.8	20.7	-	23.4	16.2	6.9	23.9	13.6	22.5	19.4	16.4	15.5
SALAMANDRA**TA**2B													
BASE + CPT-V1	17.9	19.4	18.1	-	9.5	-	-	-	-	-	19.8	3.5	-
+ INSTRUCT-V1	17.1	12.1	13.7	-	14.6	-	-	-	-	-	10.2	10.7	-
+ TRR	24.7	23.5	19.4	-	24.9	-	-	-	-	-	20.5	17.5	-
+ MBR	25.1	22.1	19.4	-	24.6	-	-	-	-	-	21.1	17.4	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	25.9	25.1	20.2	-	25.6	-	-	-	-	-	24.9	20.1	-
+ INSTRUCT-V1	29.0	27.7	22.2	-	28.7	-	-	-	-	-	24.4	20.9	-
+ TRR	26.4	25.4	21.2	-	27.1	-	-	-	-	-	22.4	19.6	-
+ MBR	26.8	25.9	20.9	-	27.1	-	-	-	-	-	23.5	20.1	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	25.6	24.7	19.6	26.1	24.1	16.9	5.3	33.0	11.9	17.4	24.8	20.6	20.1
+ INSTRUCT-V2	27.3	25.7	19.5	27.8	29.2	17.6	6.0	36.6	14.4	18.8	20.0	19.1	22.3
+ TRR	26.5	25.1	21.0	26.6	26.7	17.4	6.1	35.8	17.7	20.9	22.6	20.3	22.3
+ MBR	26.1	25.4	20.4	27.0	27.5	17.5	6.3	36.2	16.7	20.9	22.7	20.6	22.1

Table 11: BLEU scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	39.6	-	49.7	-	-	-	-	32.5	-	32.1	49.2	-	-
MADLAD400 7B	55.0	57.8	49.7	-	53.2	43.4	36.2	27.7	28.0	31.5	54.7	47.8	20.6
NLLB 3.3B	49.7	51.7	46.6	-	48.6	40.9	35.9	22.4	23.6	29.6	47.7	42.8	15.9
SALAMANDRA**TA**2B													
BASE + CPT-V1	48.4	51.7	44.5	-	33.6	-	-	-	-	-	49.7	15.5	-
+ INSTRUCT-V1	49.3	47.6	44.9	-	45.9	-	-	-	-	-	44.7	40.4	-
+ TRR	52.7	55.7	48.6	-	52.3	-	-	-	-	-	51.4	45.5	-
+ MBR	52.5	55.0	48.4	-	51.9	-	-	-	-	-	51.8	46.0	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	52.8	55.6	48.7	-	52.3	-	-	-	-	-	54.0	47.9	-
+ INSTRUCT-V1	55.9	58.4	50.7	-	55.1	-	-	-	-	-	54.4	48.6	-
+ TRR	54.0	57.3	50.1	-	54.2	-	-	-	-	-	52.9	47.8	-
+ MBR	54.4	57.2	50.1	-	54.2	-	-	-	-	-	53.9	48.2	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	52.6	54.7	48.2	54.5	51.7	42.2	34.6	28.7	22.8	26.3	54.4	47.9	22.0
+ INSTRUCT-V2	53.9	56.8	48.7	56.8	54.9	43.8	35.5	32.7	26.9	28.5	52.2	47.2	21.1
+ TRR	54.3	57.2	50.0	56.0	54.2	44.6	36.2	32.5	28.2	28.8	53.2	48.4	21.7
+ MBR	54.0	57.3	49.6	56.5	54.4	44.4	36.3	32.8	27.9	28.9	53.7	48.4	21.6

Table 12: CHRF scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	6.69	-	4.16	-	-	-	-	3.83	-	3.70	2.25	-	-
MADLAD400 7B	4.28	4.14	5.50	-	4.18	7.18	7.75	6.60	4.49	5.98	1.73	4.04	6.05
NLLB 3.3B	5.95	6.03	6.38	-	6.64	8.46	7.71	7.91	6.09	5.74	2.74	6.45	8.12
SALAMANDRA**TA**2B													
BASE + CPT-V1	5.03	5.08	5.58	-	6.53	-	-	-	-	-	2.02	6.24	-
+ INSTRUCT-V1	3.99	4.19	4.90	-	5.28	-	-	-	-	-	2.40	5.10	-
+ TRR	3.02	2.67	3.86	-	3.82	-	-	-	-	-	1.63	3.91	-
+ MBR	3.02	2.82	3.83	-	3.92	-	-	-	-	-	1.63	3.85	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	4.43	5.24	5.30	-	5.58	-	-	-	-	-	1.73	4.12	-
+ INSTRUCT-V1	2.87	2.30	3.76	-	3.60	-	-	-	-	-	1.52	3.46	-
+ TRR	2.51	2.00	3.21	-	3.11	-	-	-	-	-	1.48	3.23	-
+ MBR	2.48	1.96	3.19	-	3.12	-	-	-	-	-	1.43	3.22	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	4.91	5.26	5.52	6.54	5.97	8.24	6.87	5.79	5.89	6.40	1.73	4.03	5.08
+ INSTRUCT-V2	3.60	2.79	4.13	4.44	3.44	5.26	8.48	4.03	4.59	5.15	1.77	3.83	4.66
+ TRR	2.81	2.08	3.30	3.96	2.98	4.43	7.47	3.60	3.98	4.50	1.50	3.25	4.13
+ MBR	2.79	2.12	3.35	4.00	2.99	4.57	7.73	3.62	4.00	4.51	1.49	3.28	4.21

Table 13: METRICX scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	55.9	-	61.7	-	-	-	-	61.8	-	59.8	62.4	-	-
MADLAD400 7B	69.3	71.2	58.7	-	62.4	52.4	39.2	53.2	53.8	52.8	68.9	61.5	54.5
NLLB 3.3B	65.2	67.7	58.3	-	60.4	51.0	40.3	48.5	45.9	53.2	62.9	58.6	43.8
SALAMANDRA**TA**2B													
BASE + CPT-V1	66.2	67.3	57.9	-	49.8	-	-	-	-	-	67.1	29.1	-
+ INSTRUCT-V1	68.5	69.5	59.7	-	61.7	-	-	-	-	-	66.1	59.4	-
+ TRR	70.7	73.3	62.4	-	64.5	-	-	-	-	-	68.0	61.2	-
+ MBR	70.7	73.1	61.9	-	64.4	-	-	-	-	-	68.3	62.6	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	68.3	67.1	58.2	-	60.5	-	-	-	-	-	68.5	63.2	-
+ INSTRUCT-V1	72.5	75.4	62.6	-	66.7	-	-	-	-	-	69.5	64.4	-
+ TRR	72.6	75.7	64.2	-	67.4	-	-	-	-	-	69.2	65.0	-
+ MBR	73.1	75.9	63.9	-	67.6	-	-	-	-	-	70.0	64.9	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	67.0	66.8	58.0	68.5	59.3	52.0	41.5	54.2	48.9	50.3	68.5	63.3	55.5
+ INSTRUCT-V2	69.8	74.1	62.2	73.3	67.6	58.4	38.5	61.4	54.1	55.5	69.0	64.6	55.7
+ TRR	71.8	75.2	63.8	73.7	67.7	59.2	39.5	62.6	55.6	56.8	69.6	65.3	57.1
+ MBR	71.8	75.4	63.8	74.1	67.8	59.2	39.2	62.4	55.6	56.8	69.5	65.4	56.8

Table 14: BLEURT scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	4.87	-	2.28	-	-	-	-	2.46	-	1.74	3.50	-	-
MADLAD400 7B	3.38	3.38	3.89	-	2.94	4.95	5.31	6.00	3.50	3.66	3.28	3.31	8.32
NLLB 3.3B	4.83	4.92	4.80	-	4.91	6.11	4.61	7.83	4.48	3.21	6.35	5.16	9.65
SALAMANDRA**TA**2B													
BASE + CPT-V1	3.73	4.02	3.46	-	5.48	-	-	-	-	-	3.97	4.46	-
+ INSTRUCT-V1	2.89	3.46	3.21	-	3.67	-	-	-	-	-	4.12	3.38	-
+ TRR	1.78	1.74	1.96	-	2.02	-	-	-	-	-	2.59	1.94	-
+ MBR	1.86	1.92	2.01	-	2.21	-	-	-	-	-	2.68	2.03	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	3.40	4.15	3.27	-	3.71	-	-	-	-	-	3.17	2.73	-
+ INSTRUCT-V1	1.82	1.75	2.07	-	2.14	-	-	-	-	-	2.69	1.87	-
+ TRR	1.49	1.42	1.63	-	1.69	-	-	-	-	-	2.46	1.58	-
+ MBR	1.52	1.44	1.74	-	1.80	-	-	-	-	-	2.46	1.60	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	3.66	4.22	3.43	4.12	4.17	5.91	3.88	3.90	3.89	3.70	3.00	2.68	4.86
+ INSTRUCT-V2	2.44	2.04	2.39	2.70	2.07	3.04	5.11	2.52	2.65	2.54	3.06	2.27	4.30
+ TRR	1.67	1.40	1.67	2.26	1.66	2.35	3.95	2.14	2.07	1.93	2.50	1.57	3.75
+ MBR	1.83	1.50	1.78	2.24	1.73	2.55	4.20	2.22	2.22	2.04	2.51	1.75	3.86

Table 15: METRICX-QE scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	69.4	-	79.5	-	-	-	-	78.5	-	82.1	75.6	-	-
MADLAD400 7B	78.8	79.3	76.7	-	78.3	70.4	70.4	70.4	79.5	77.1	79.4	79.3	69.5
NLLB 3.3B	75.5	76.0	75.3	-	74.5	69.0	70.9	66.9	76.6	79.0	73.5	75.1	60.5
SALAMANDRA**TA**2B													
BASE + CPT-V1	77.2	77.3	76.6	-	67.0	-	-	-	-	-	77.4	76.0	-
+ INSTRUCT-V1	78.5	77.7	77.8	-	75.9	-	-	-	-	-	75.0	77.1	-
+ TRR	83.4	85.1	82.4	-	81.6	-	-	-	-	-	81.4	82.6	-
+ MBR	81.2	82.6	80.3	-	79.4	-	-	-	-	-	78.5	80.2	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	77.8	77.1	77.2	-	75.6	-	-	-	-	-	78.0	79.2	-
+ INSTRUCT-V1	81.3	82.6	80.4	-	79.9	-	-	-	-	-	78.5	80.0	-
+ TRR	84.2	86.0	83.1	-	82.6	-	-	-	-	-	81.7	83.2	-
+ MBR	82.5	83.8	81.3	-	80.8	-	-	-	-	-	79.3	81.0	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	77.4	76.9	76.9	78.2	74.6	69.2	72.8	73.7	76.8	75.9	78.5	78.7	71.8
+ INSTRUCT-V2	80.2	81.6	79.7	82.7	79.9	75.2	68.8	78.7	80.6	79.3	77.7	78.4	70.1
+ TRR	84.0	86.0	83.0	85.5	82.7	79.8	74.1	81.5	84.0	83.1	81.6	82.8	75.7
+ MBR	82.0	83.9	81.1	83.9	80.6	76.9	71.1	80.0	82.3	81.1	78.9	80.3	71.7

Table 16: COMET-KIWI scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	COMET							METRICX						
	DE	EL	IT	LT	RO	SR	SV	DE	EL	IT	LT	RO	SR	SV
SALAMANDRA**TA**2B														
BASE + CPT-V1														
+ INSTRUCT-V1	76.6	83.5	78.6	79.7	80.3	75.3	80.9	2.31	4.10	4.03	5.20	4.22	6.18	3.28
+ TRR	80.6	85.7	82.2	83.7	84.1	80.8	84.5	1.63	3.37	2.62	3.85	3.02	4.53	2.25
+ MBR	81.9	86.6	83.4	85.1	85.0	81.5	85.3	1.60	3.39	2.69	3.84	3.08	4.71	2.33
SALAMANDRA**TA**7B														
BASE + CPT-V1														
+ INSTRUCT-V1	80.6	86.0	82.2	83.1	82.8	79.8	84.4	1.75	3.35	2.78	3.81	3.47	4.32	2.47
+ TRR	82.0	86.5	83.2	85.5	85.4	82.4	85.7	1.40	2.91	2.26	3.02	2.46	3.53	1.81
+ MBR	83.3	87.6	84.5	86.6	86.6	83.6	86.6	1.37	2.85	2.30	2.84	2.50	3.60	1.91

Table 17: COMET and METRICX scores for the WMT-Multilingual Sub-Task (English to seven target languages) on the WMT24++ test set. Results are shown for the instruction-tuned SALAMANDRA**TA** 2B and 7B models, with and without post-decoding strategies (MBR and TRR).