

BSC Submission for WMT26 General Machine Translation Shared Task

Audrey Mash¹ Jonathan Orama Ayebakuro¹
Ella Bohman¹ Xixian Liao¹ Francesca De Luca Fornaciari¹
Maite Melero¹

¹Barcelona Supercomputing Center

Abstract

We present the BSC submission to the WMT26 General Machine Translation shared task: a 7B decoder-only model covering 23 directions, adapted from the SALAMANDRA (Gonzalez-Agirre et al., 2025) family through vocabulary replacement, two rounds of continued pre-training (monolingual, followed by instruction-formatted parallel data), supervised fine-tuning on a translation-only instruction mixture, and preference optimisation applied selectively to six weak directions via CPO-SimPO (Xu et al., 2024; Meng et al., 2024) with LoRA adapters (Hu et al., 2021). Submissions are decoded with Minimum Bayes Risk selection over a candidate pool filtered for truncation, degeneracy and off-target script. As official human evaluation was unavailable at the time of writing, we report a stage-wise internal evaluation on FLORES+ and WMT24++, together with automatic scores for the submitted system on the 12 test-set directions for which references have been released. Instruction tuning accounts for almost all measurable gain, but its size differs by an order of magnitude between the two evaluation sets; preference optimisation moves automatic metrics by less than the drift among untreated directions. The model is released under a Research-Only Licence and is available on request.¹

1 Introduction

We describe the Barcelona Supercomputing Center submission to the WMT 2026 General Machine Translation shared task: a 7B-parameter decoder-only translation model covering 23 directions, adapted from the SALAMANDRA family.

Not all shared-task languages are represented in SALAMANDRA’s original vocabulary, and the vocabulary replacement required to cover them is

incompatible with the released SALAMANDRA (Garcia Gilabert et al., 2025) checkpoints. Our pipeline therefore restarts from the SALAMANDRA-7B base model and proceeds in four stages: vocabulary adaptation with monolingual continued pre-training (CPT1); continued pre-training on instruction-formatted parallel data (CPT2); supervised fine-tuning on a translation-only instruction mixture (SFT); and preference optimisation applied selectively to six weak directions (PO). CPT2 was planned as two shards but was stopped at the end of the first, at which point it had seen half of the planned data, when our compute allocation was exhausted. We report what was consumed rather than what was planned.

Because the 2026 task is assessed by human evaluation alone, and those results are unavailable at submission time, we report a stage-by-stage internal comparison of the pipeline. We evaluate three checkpoints (post-CPT2, post-SFT, post-PO) on FLORES+ devtest (22 directions) and WMT24++ (16 directions) with BLEU, chrF and COMET. English–Northern Sámi is excluded throughout due to the lack of evaluation data. We also score the submitted system against the human references released for 12 of the 23 directions (§5.2).

Our contributions are: (i) a stage-wise internal evaluation isolating the contribution of each adaptation stage across 23 directions (§5); (ii) a null result on selective preference optimisation, with the conditions under which it is uninformative rather than negative (§6); and (iii) an inference pipeline for the task’s structured and instruction-conditioned inputs, in which an ordered eligibility filter removes truncated, degenerate and off-script candidates before Minimum Bayes Risk decoding (Eikema and Aziz, 2020), addressing the two dominant selection failures we observed (§4).

¹Earlier MT models in the family are public: SALAMANDRA7B-v1, SALAMANDRA2B-v1, SALAMANDRA7B-v2.

2 Data

Training proceeds in two rounds of continued pre-training on top of the vocabulary-adapted SALAMANDRA-7B base model. All token counts are measured with our extended tokeniser.

2.1 CPT1: monolingual

For CPT1 we assembled a monolingual corpus covering all languages already supported by SALAMANDRA together with those newly introduced by WMT 2026 (the full split of existing versus new languages is in Appendix A). Each language was sampled up to a 1B-token cap, and languages with less available data contributed everything we had (Table 1). In total, 52 language datasets amounting to 44.91B tokens were globally shuffled into a single shard, with 0.05% held out for validation and 0.05% for testing.

Most data came from FineWeb (Penedo et al., 2024) for English and FineWeb-2 (Penedo et al., 2025) for the remaining languages. Where FineWeb-2 could not fill the budget we drew on CC-100 (Wenzek et al., 2020) (Kazakh, Traditional Chinese, and partially Belarusian and Armenian), Leipzig Corpora (Eckart and Quasthoff, 2013) and the Northern Sámi side of OPUS (Tiedemann, 2012), and regional corpora for Aranese (PILAR, DOGC, WMT24 low-resource Occitan), Valencian (Les Corts Valencianes) and Asturian. Simplified Chinese was obtained by script conversion of the FineWeb-2 Chinese sample. Data was used as released, without additional filtering.

Four coverage gaps should be noted. Thai was unintentionally omitted, as its sample was not ready when the shard was assembled; we compensate in CPT2. Ligurian and Ladin were excluded because no usable monolingual data was available. Northern Sámi was included with only ~ 30 M tokens, as it was all we could source. Lastly, Turkish (1B tokens) was included despite not being a WMT 2026 language, to provide transfer support for Kazakh.

2.2 CPT2: parallel and monolingual

CPT2 uses instruction-formatted parallel data ($\sim 72\%$ of the mix by source–target tokens) and monolingual data ($\sim 28\%$), restricted to the WMT 2026 languages plus Italian and Turkish. The mix contains 85.1B tokens including template overhead, globally shuffled, with the same 0.05%/0.05% held-out split, trained for a single epoch.

Languages	Tokens
ar, be, bg, ca, cs, da, de, el, en, es, et, eu, fi, fr, gl, hi, hr, hu, hy, id, is, it, ja, kk, ko, lt, lv, nb, nl, pl, pt, ro, ru, sk, sl, sr, sv, tr, uk, vi, zh, zh_TW	~ 1 B each
nn	768.0M
cy	723.9M
ga	591.1M
mt	425.5M
ast	202.7M
oc	88.8M
val (Valencian)	39.3M
se	30.0M
oc-aranes (Aranese)	28.0M
arg (Aragonese)	16.9M
Total	44.91B

Table 1: Monolingual data used in CPT1. Languages above the line were capped at 1B tokens; for those below, all available data was used.

Parallel data. We downloaded all corpora permitted for the constrained task with mtdata (Gowda et al., 2021) using the official WMT 2026 recipe, and audited licences, discarding corpora that do not permit derived work even for research. The data was cleaned with ParaCLEAN (Mash et al., 2026), which scores every pair with LaBSE (Feng et al., 2022) alignment and GlotLID (Kargaran et al., 2023) language identification, filters pairs below 0.75 alignment or 0.5 language-identification probability on either side, and deduplicates and normalises the remaining data. GlotLID was selected after comparing four identifiers (Lingua-py, fastText, Idiomata Cognitor, GlotLID), achieving the highest accuracy on FLORES+ (91.07%) and on the noisier WMT24++ (87.92%) (Mash et al., 2026). The result was merged with the existing SALAMANDRA parallel data for the same pairs and cross-deduplicated.

Each direction was converted to plain-text records by embedding sentence pairs in natural-language translation instructions, sampling uniformly from a pool of 35 templates, with and without an explicit source-language name; the pool is listed in Appendix B. Each direction was capped at 5B source-plus-target tokens, giving 52.4B tokens over 937M sentence pairs across 21 directions (Appendix C, Table 6).

Monolingual data. We added 1B tokens for each of the 19 CPT2 languages (fresh samples from the CPT1 sources, non-overlapping with the CPT1 shard), for 20B tokens in total. Thai received a double share to compensate for its omission from

CPT1. English monolingual data was not included, as English is abundantly represented on the source side of the parallel data.

Coverage. Italian (1B monolingual, 5B en–it parallel) was added to support the closely related Ligurian and Ladin, and Turkish (1B monolingual) to support Kazakh. The shared task’s Arabic direction is into Egyptian Arabic, but no variety filtering was applied at either continued pre-training stage: the 5B en–ar tokens in the CPT2 mix are generic Arabic, predominantly Modern Standard, and serve as the base on which the Egyptian variety is acquired at the SFT stage. Ligurian, Ladin and Northern Sámi could not be meaningfully represented: after cleaning, only 1.7M, 9K and 118K tokens respectively remained for en–lij, en–lld and en–se. These were left out of CPT2 and first appear in the SFT mixture.

2.3 Supervised fine-tuning

The SFT mixture contains 111,106 examples across eight translation-related tasks: sentence-, paragraph- and document-level MT, HTML tag transfer, terminology-constrained MT, formality-conditioned translation, multi-reference MT, and formality register transfer. Source datasets and language coverage are given in Table 9, and the task distribution in Figure 3. All instructions use the ChatML template (OpenAI, 2023).

The composition is motivated on two fronts. First, we previously found that instruction tuning on MT-related tasks either improves translation quality or leaves it unchanged, whereas unrelated chat and code instructions consistently hurt it (García Gilabert et al., 2025); we therefore restrict the mixture to translation-centric tasks. Second, the auxiliary tasks target the instruction-following phenomena introduced this year: formality control, glossary adherence and structured translation. Small datasets were used in full, and larger ones uniformly subsampled to 250–2,000 examples per language pair with a fixed seed. For the lowest-resource pairs we upsampled existing data (en–arz) or reused parallel sentences from the CPT stage (en–be, en–lld, en–lij, en–se). Egyptian Arabic is introduced for the first time at this stage, the continued pre-training data having been generic (mostly Modern Standard) Arabic. The Egyptian variant was used for most tasks in the mixture, with one exception: HTML tag transfer, for which no Egyptian Arabic data was available, so generic Arabic

was used instead.

2.4 Preference optimisation

We built preference pairs from a subset of the sentence-level SFT data for the six directions with the weakest post-SFT output: English into Egyptian Arabic, Thai, Belarusian, Northern Sámi, Ladin and Ligurian. For five of these we used the SFT reference as the chosen example and the post-SFT model output as the rejected one. For Ligurian we instead used Diverse Beam Search (Vijayakumar et al., 2018) to generate 32 candidates (32 beams, 8 groups) per source, re-ranked by MBR with ChrF++ (Popović, 2017) as utility, and formed two pairs per source with the best candidate as chosen and the 16th and 32nd as rejected, following the BMW strategy (Yang et al., 2024). MBR was limited to Ligurian because MBR-derived pairs are drawn from the model’s own output, and applying this to lower-quality directions risks amplifying existing errors. The set totals 20,090 sentences; the breakdown is in Table 11.

3 Model and Training

3.1 Tokeniser adaptation

The SALAMANDRA models have a 256K-token vocabulary. Rather than appending new tokens for the WMT26 languages, we retrain the tokeniser on a corpus spanning both the original languages and the newly added ones, allowing the merge budget to reallocate jointly across all of them. This changes a substantial fraction of the vocabulary, making the new tokeniser incompatible with released SALAMANDRA checkpoints; subsequent training therefore resumes from the original SALAMANDRA-7B base model.

We trained two candidate tokenisers, each on a corpus covering both the original SALAMANDRA languages and the new ones, using the SentencePiece (Kudo and Richardson, 2018) implementation of BPE (Sennrich et al., 2016). They differ only in language distribution: one uniform, one down-weighting Traditional Chinese to mitigate Chinese script bias. Fertility across all measurable FLORES+ languages showed no significant bias under the uniform distribution, so we adopted it. Figure 4 compares fertility against multilingual baselines; both candidates achieve consistently lower tokens-per-word and tokens-per-byte. Tokens shared with the original vocabulary retained their embeddings; the remainder were initialised

from a Gaussian derived from the mean of existing embeddings. Replacing the vocabulary at constant size leaves the embedding and unembedding matrices unchanged in shape, so no architectural modification was required.

3.2 Continued pre-training

We continually pre-trained the vocabulary-adapted 7B model with NVIDIA NeMo and Megatron-Core, retaining the original decoder architecture: 32 layers, hidden size 4096, SwiGLU intermediate size 11008, 32 attention heads with eight query groups, RMSNorm, rotary position embeddings, an 8192-token context window, and untied input and output embeddings. All parameters were updated in both CPT stages; no layers were frozen.

CPT1 ran for 10,708 updates on 44.91B tokens of multilingual monolingual text. CPT2 combines fresh monolingual data with sentence-aligned parallel data over 21 directions, rendered as translation instructions in both directions, split into two shards of 42.54B tokens each. We trained 10,143 updates, one full pass over the first shard. The second shard was configured for a further 10,145 updates but our compute allocation was exhausted before it began, so the checkpoint carried into SFT has seen half of the planned CPT2 data.

All stages use 8192-token sequences, a micro-batch of one and a global batch of 512 sequences (4,194,304 tokens per update). We optimise the causal LM objective in bfloat16 with distributed fused Adam ($\beta_1=0.9$, $\beta_2=0.95$), weight decay 0.1 and gradient clipping 1.0. CPT1 uses 1,024 warm-up updates and a cosine schedule from 3×10^{-4} to 3×10^{-5} . CPT2 resets the optimiser at 3×10^{-5} , with 200 warm-up updates. Training uses tensor parallelism of four, sequence parallelism, selective activation recomputation and bfloat16 Transformer Engine kernels without FP8.

CPT1 trained on 32 GPUs to a late validation loss of 2.24; the CPT2 shard trained on 128 GPUs to a final validation loss of 1.64. In total CPT comprised 20,851 updates over approximately 87.46B tokens. Robustness to queue limits and pre-emption came from periodic distributed checkpointing and resumption across Slurm jobs.

3.3 Instruction tuning and preference optimisation

The final CPT checkpoint was fully fine-tuned with FastChat (Zheng et al., 2023) and DeepSpeed ZeRO-3 on MareNostrum 5, for one epoch over

the full mixture with a cosine schedule decaying over the corresponding number of steps; the final checkpoint was carried forward. Preference optimisation used TRL (von Werra et al., 2020) with a combination of CPO (Xu et al., 2024) and SimPO (Meng et al., 2024), which adds length normalisation and a target reward margin, applied through LoRA (Hu et al., 2021) adapters merged into the final model. Hyperparameters for both stages are in Tables 10 and 12.

4 Inference Pipeline

The submission is produced by a single checkpoint wrapped in a deterministic, resumable pipeline that converts each blindset document into translation units, decodes and selects a hypothesis per unit, and rebuilds document-level outputs from a manifest of reconstruction metadata.

Prompt construction. Each blindset row carries a domain-specific instruction, which we classify from lexical cues into one of eleven prompt types covering plain translation, legal terminology, inclusive gender, ASR transcripts, social media with and without screenshot context, Korean ellipsis, news and educational HTML, game dialogue, and software-localisation JSON. Documents are segmented by type: plain and transcript inputs on non-empty lines, HTML on top-level `<p>` paragraphs, JSON on top-level string values. Structured genres receive a template instructing the model to preserve markup, structural tokens and placeholders verbatim while translating only readable text; the legal template injects the required-terminology list, and the game-dialogue template supplies the speaker’s in-game role for register control. For multi-unit documents the preceding source segment is prepended as context. Templates are given in Appendix E.

Decoding. We use Minimum Bayes Risk decoding (Eikema and Aziz, 2020), drawing $N=8$ candidates per unit by nucleus sampling (temperature 0.6, top- p 0.95, top- k 50, up to 2048 tokens). Before scoring, candidates pass an ordered eligibility filter that removes, whenever a compliant alternative exists, candidates that reached the generation cap, are degenerate (repeated 4-gram fraction ≥ 0.35), or are incompatible with the target script; it then retains those with the fewest structural violations against the required HTML tags, JSON placeholders and ellipses. MBR runs over

the surviving pool with a COMET utility metric (Rei et al., 2022a) (wmt22-comet-da). The filter targets the two dominant selection failures we observed: preferring a truncated candidate over a complete one, and preferring an off-target-language candidate when an on-target one was available.

Reconstruction and validation. Hypotheses are cleaned of code fences and boilerplate labels, then reassembled using stored prefixes and separators so that whitespace-sensitive structure is reproduced verbatim, with a small set of deterministic content-preserving repairs. Reference-free checks then flag, without altering, the submission: per-segment language identification (Kargaran et al., 2023), numeric-entity preservation, HTML and JSON structural integrity, placeholder and ellipsis retention, inclusive-gender completeness, and the official structural alignment check. Every selected hypothesis is additionally scored with COMET-Kiwi (Rei et al., 2022b) (wmt22-cometkiwi-da). Appendix E gives JSON parsing and reconstruction details.

5 Results

Metrics. We report BLEU and chrF (sacreBLEU 2.4.0 (Post, 2018)), and additionally the neural metric COMET-22 (Unbabel/wmt22-comet-da) (Rei et al., 2022a).² §5.1 compares training checkpoints using BLEU/chrF/COMET, while §5.2 evaluates the submitted system against the released human references using BLEU and chrF.

5.1 Effect of the training stages

Setup. We evaluate the post-CPT2, post-SFT and post-PO (Final) checkpoints with MT-Lens (García Gilabert et al., 2025). Decoding is identical across checkpoints (beam search, five beams, batch size five), so differences are attributable to the intervening training stage and not to generation settings. We report BLEU, chrF and COMET on FLORES+ devtest (22 directions) and WMT24++ (16).

This configuration is deliberately simpler than the submission-time pipeline of §4: it measures the checkpoints under fixed decoding, with no contribution from MBR or eligibility filtering. Since no

²The BLEU signature is nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0, except that the tokeniser is selected by target language (tok:zh for Chinese, ja-mecab and ko-mecab for Japanese and Korean), so BLEU for those targets is not comparable with the rest of any table. The chrF signature is nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.0.

	BLEU	chrF	COMET
<i>(a) Mean over directions, by stage</i>			
FLORES+ ($n = 22$)			
CPT2	26.40	50.04	0.838
SFT	28.79	51.77	0.856
Final	28.83	51.83	0.856
Δ SFT	+2.39	+1.73	+0.018
Δ PO	+0.04	+0.06	+0.000
WMT24++ ($n = 16$)			
CPT2	22.98	43.99	0.781
SFT	33.65	52.62	0.857
Final	33.64	52.61	0.857
Δ SFT	+10.66	+8.62	+0.076
Δ PO	−0.01	−0.01	0.000
<i>(b) Directions WMT24++ does not cover, FLORES+</i>			
en-lij			
CPT2	4.42	37.41	0.565
SFT	27.53	55.07	0.619
Final	28.14	55.53	0.620
en-lll			
CPT2	2.90	29.93	0.501
SFT	2.79	27.52	0.499
Final	2.58	27.54	0.498
en-arz			
CPT2	13.89	47.01	0.816
SFT	16.22	49.16	0.827
Final	16.43	49.56	0.828
en-be			
CPT2	14.08	45.12	0.826
SFT	15.92	46.91	0.863
Final	16.64	47.72	0.870
en-hy			
CPT2	20.98	56.11	0.870
SFT	24.87	59.54	0.900
Final	24.39	59.20	0.898
en-kk			
CPT2	23.32	58.20	0.896
SFT	23.90	58.83	0.900
Final	23.80	58.75	0.898

Table 2: Pipeline-stage results. Panel (a) gives means over all directions per stage and evaluation set; panel (b) gives the six directions WMT24++ does not cover, so every claim about them rests on FLORES+ alone. Ligurian and Ladin had no parallel data in CPT2, and Egyptian Arabic none of its own variety. Best stage per metric in bold. Full per-direction tables: Appendix D. English–Northern Sámi is excluded throughout.

degeneracy filtering is applied, repetition loops and prompt echoing are visible in these scores where they occur. The numbers in this subsection therefore characterise the training recipe, not the submitted system.

Findings. Mean BLEU rises from 26.40 to 28.79 across the SFT stage on FLORES+ and from 22.98 to 33.65 on WMT24++. The PO stage changes mean BLEU by +0.04 on FLORES+ and −0.01 on WMT24++. Ligurian is the largest single movement, rising from 4.42 to 27.53 BLEU at SFT;

System	BLEU	chrF
Lumen	41.94	60.38
Qingqiu-MT-9B	39.27	56.94
SalamandraTA (ours)	33.01	52.00
Gemma 4-4B	29.73	48.33
GPT-OSS-20B	25.01	44.54
Qwen 3.5-9B	24.32	44.82
Tower 9B	23.46	44.64
GLM 4-9B	23.15	42.75
Ministral 3-14B	18.15	42.82

Table 3: Constrained-track systems covering all reference directions, ranked by BLEU. Scores are averaged over the 12 WMT26 directions with human references (cs-de, cs-vi, en-cs, en-de, en-et, en-is, en-arz, en-ja, zh-ja, en-lij, en-lll, en-se).

Ladin is the only English-source direction that declines at every stage.

5.2 Released test set and human references

We report automatic scores on the WMT26 General Translation Shared Task test set against the organisers’ human references,³ which cover 12 of the 23 directions. We use the post-edited reference (refPE) where available and the from-scratch human reference (refA) otherwise. We report only surface metrics: many test documents exceed the input length of the neural metrics, which would require re-segmenting to score reliably. For Czech-German and Czech-Vietnamese, which provide both, we also report scores against refA in Appendix G. These are our own automatic-metric scores against references released after the evaluation period; official human evaluation and system rankings were not available at the time of writing, and will be reported in the findings paper (Kocmi et al., 2026).

We compare against the nine constrained-track systems that cover all 12 reference directions, out of 19 constrained submissions, using the system outputs published in the task repository.⁴ The comparison uses BLEU and chrF, averaged over the 12 directions; per-direction scores are in Appendix G.

Our 7B submission ranks third of the nine on both chrF (52.0) and BLEU (33.0), behind only the larger Qwen-based Lumen (a 14B model with a 4B post-editor) and Qingqiu-MT-9B, making it the

³<https://github.com/wmt-conference/wmt26-general-mt>

⁴The constrained track admits open-weight models of at most 20B parameters; unconstrained systems, such as proprietary API models, are omitted as not directly comparable. The remaining ten constrained systems are language-specific, covering one to four directions each.

smallest of the top three. As refPE references are post-edits of machine output, surface metrics may favour systems closer to that output; the ranking should be read with this in mind and is superseded by the official human evaluation.

6 Analysis

The evaluation set determines the size of the instruction-tuning effect. Averaged over directions, SFT gains 2.39 BLEU on FLORES+ against 10.66 on WMT24++, and 0.018 against 0.076 COMET. This is not driven by a few directions: of the sixteen both sets cover, every one shows a larger gain on WMT24++, and four (cs-de, cs-uk, cs-vi, zh-ja) move from flat or slightly negative on FLORES+ to gains of 0.4–6.3 BLEU. On FLORES+ alone these four would suggest that non-English-source directions regress under instruction tuning; the second set does not support that reading. Nor does the gain track CPT2 data volume: Spearman ρ between the FLORES+ SFT gain and per-direction token counts is -0.08 , and en-zh_TW, with the smallest allocation at 0.23B, gains 5.1 BLEU while the capped directions gain 1.0–2.2. We suggest, as a hypothesis rather than a result, that the sets place different demands on the model: FLORES+ is uniform and sentence-level, and a continued-pre-trained checkpoint prompted with a translation instruction already handles it, whereas WMT24++ is naturalistic post-edited text where instruction compliance, avoiding over-generation and staying on-target matter more. The practical observation does not depend on the explanation: FLORES+ alone would have valued the SFT stage at roughly a fifth of what the second set reports.

Preference optimisation did not change automatic metrics. The five treated directions with evaluation data move between -0.21 and $+0.98$ BLEU on FLORES+, against -0.49 to $+0.31$ among the seventeen untreated directions; on WMT24++ every direction moves by less than 0.75 BLEU. In Tables 7 and 8 the final checkpoint holds the best score in roughly half of all cells, the distribution expected from an intervention with no systematic effect. Preference-set size does not predict the outcome either: Ligurian received the largest set at 7,252 pairs and moved $+0.61$ BLEU, en-th received 4,000 and moved $+0.98$, and en-arz received 5,000 and moved $+0.21$. Only Ladin, at 93 pairs, is small enough that no effect would be expected. Two readings remain open and our

evidence does not separate them: the stage may have had little effect, or reference-based metrics may be insensitive to the failures it was built to remove, namely repetition loops and prompt echoing. These failures would be present in the outputs being scored if they remained, since no degeneracy filtering is applied during evaluation; the question is whether the metrics register them. A repetition-rate measure or targeted human evaluation would distinguish the two. We ran neither, nor the matched-cost uniform-application ablation that would establish whether concentrating a small preference set on six directions was preferable to distributing it.

Low-resource outcomes track data volume.

Ligurian and Ladin passed through the same stages: no data in either CPT round, Italian as a high-resource relative, first representation at SFT, and preference optimisation on top. Their outcomes diverge. Ligurian rises from 4.42 to 27.53 BLEU at SFT and ends at 28.14, within range of our mid-resource directions. Ladin declines at every stage, 2.90 to 2.79 to 2.58, with chrF falling from 29.93 to 27.52; it is the only English-source direction that ends below its starting score. The difference follows the data: after cleaning, en-lij retained 1.7M parallel tokens against 9K for en-lll, and the Ladin corpus was additionally judged of poor quality on inspection. The same gap holds at the preference stage (7,252 pairs against 93). The Ligurian result is the more informative: a direction with 1.7M cleaned tokens, none seen during continued pre-training, reached mid-resource quality through SFT alone (+23.12 BLEU, against +0.61 for PO). A qualitative example in our logs suggested that preference optimisation had resolved Ligurian’s repetition problem; the stage-wise results do not support attributing the recovery to that stage. We cannot separate data quantity from the other factors that differ between the two directions, since they also differ in typological distance from Italian and in how their preference sets were built. What the pair suggests is a working threshold on this pipeline: roughly 1.7M cleaned parallel tokens introduced at SFT sufficed for a direction absent from continued pre-training, and 9K did not. Both directions fall outside WMT24++ coverage, so this rests on FLORES+ alone. On the WMT26 test set, however, en-lll reaches 17.03 BLEU (Table 3); the FLORES+ result may therefore reflect the evaluation set as much as the model.

7 Conclusion

We submitted a 7B decoder-only model covering 23 directions, adapted from SALAMANDRA-7B through vocabulary replacement, two rounds of continued pre-training, instruction tuning on a translation-only mixture, and preference optimisation on six weak directions, decoded through a structure-aware inference pipeline with candidate filtering ahead of MBR selection.

We draw two conclusions. Stage-wise internal evaluation is inexpensive once the checkpoints exist and locates where the measurable value of a pipeline sits. And a single automatic evaluation set is not sufficient evidence for a claim about a training stage: our two sets differed by an order of magnitude on the largest effect we measured, agreed on the null result, and one covers none of our six lowest-resource directions. For directions at the low end, our two Italian-supported cases suggest the binding constraint is the quantity and quality of usable parallel text rather than the choice of adaptation strategy, and we would direct future effort at data collection before pipeline design.

Limitations

We hold no internal evaluation data for English–Northern Sámi, a shared-task direction and one of the six targeted with preference optimisation; its only automatic assessment is the released test-set reference (Table 3). CPT2 stopped at the end of its first shard, half of its planned data, when our compute allocation was exhausted, so the CPT2 column throughout is an under-trained checkpoint and the gains attributed to later stages are measured against it. The negative result for preference optimisation is a null result on automatic metrics rather than on translation quality; we did not run the degeneracy-specific or human evaluation that would test the failure modes the stage was built to remove. Ligurian and Ladin were not treated identically at the preference stage: Ligurian was the one direction for which we used MBR-derived rather than reference-based pairs. The model is released under a Research-Only Licence by the Language Technologies Unit, Barcelona Supercomputing Center, and may not be used commercially.

Acknowledgments

This work has been promoted and financed by the Government of Catalonia through the Aina project. It is supported by MLLM4TRA

References

- Thomas Eckart and Uwe Quasthoff. 2013. Statistical corpus and language comparison on comparable corpora. In *Building and using comparable corpora*, pages 151–165. Springer.
- 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.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic bert sentence embedding](#). *Preprint*, arXiv:2007.01852.
- Javier García Gilabert, Carlos Escolano, Audrey Mash, Xixian Liao, and Maite Melero. 2025. [MT-LENS: An all-in-one toolkit for better machine translation evaluation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 51–60, Albuquerque, New Mexico. Association for Computational Linguistics.
- Javier García Gilabert, Xixian Liao, Severino Da Dalt, Ella Bohman, Audrey Mash, Francesca De Luca Fornaciari, Irene Baucells, Joan Llop, Miguel Claramunt, Carlos Escolano, and Maite Melero. 2025. [From SALAMANDRA to SALAMANDRATA: BSC submission for WMT25 general machine translation shared task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 614–637, Suzhou, China. Association for Computational Linguistics.
- 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.
- Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. [Many-to-English machine translation tools, data, and pretrained models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. [GlotLID: Language identification for low-resource languages](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Sergey Dukanov, Anton Dvorkovich, Mark Fishel, Markus Freitag, Samuel Frontull, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Stan Kharevich, Philipp Koehn, Zheng Li, Jean Maillard, Christof Monz, Alexander Murauski, and 12 others. 2026. Findings of the WMT26 general machine translation shared task: Contrastive dynamic human evaluation at scale. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Audrey Mash, Ella Paulina Bohman, and Maite Melero. 2026. ParaCLEAN: Improving Translation Quality through Systematic Parallel Data Cleaning. *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6630–6640.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- OpenAI. 2023. ChatML: Chat markup language. <https://github.com/openai/openai-python/blob/release-v0.28.1/chatml.md>. GitHub documentation, openai-python v0.28.1.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The fineweb datasets: Decanting the web for the finest text data at scale](#). *Preprint*, arXiv:2406.17557.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language](#). *Preprint*, arXiv:2506.20920.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on*

- Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 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, 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.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural Machine Translation of Rare Words with Subword Units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, 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.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Galouédec. 2020. [TRL: Transformers Reinforcement Learning](#).
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, and 1 others. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *LREC*.
- Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. [Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation](#). *Preprint*, arXiv:2401.08417.
- Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2024. [Direct preference optimization for neural machine translation with minimum Bayes risk decoding](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.
- 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.

A Language coverage

Table 4 lists the languages introduced for WMT 2026 and Table 5 those carried over from SALAMANDRATA. Counts are the CPT1 monolingual totals measured with the extended tokeniser; † marks a language appearing in at least one shared-task direction, and * a transfer anchor included for cross-lingual support rather than because it is a task language. Three of the ten new languages contributed no CPT1 data at all: Thai was not ready when the shard was assembled, and no usable monolingual Ligurian or Ladin could be sourced. All three are addressed at later stages (§2.2, §2.3).

Egyptian Arabic is a shared-task target but has no row of its own: CPT1 applied no variety filtering, so the Arabic data is generic and appears as ar (§2.2).

Language	Code	CPT1 tokens
Armenian [†]	hy	1.00B
Belarusian [†]	be	1.00B
Chinese (Traditional) [†]	zh_TW	1.00B
Icelandic [†]	is	1.00B
Indonesian [†]	id	1.00B
Kazakh [†]	kk	1.00B
Northern Sámi [†]	se	30.0M
Thai [†]	th	—
Ligurian [†]	lij	—
Ladin [†]	lld	—
Total		6.03B

Table 4: Languages newly introduced for WMT 2026, none of which were supported by SALAMANDRATA. All ten are shared-task languages. Dashes mark languages that contributed no CPT1 data.

Lang.	Code	Tokens	Lang.	Code	Tokens
Arabic [†]	ar	1.00B	Italian*	it	1.00B
Aragonese	arg	16.9M	Japanese [†]	ja	1.00B
Aranese	oc-aranes	28.0M	Korean [†]	ko	1.00B
Asturian	ast	202.7M	Latvian	lv	1.00B
Basque	eu	1.00B	Lithuanian	lt	1.00B
Bulgarian	bg	1.00B	Maltese	mt	425.5M
Catalan*	ca	1.00B	Norw. Bokmål	nb	1.00B
Croatian	hr	1.00B	Norw. Nynorsk	nn	768.0M
Czech [†]	cs	1.00B	Occitan	oc	88.8M
Danish	da	1.00B	Polish	pl	1.00B
Dutch	nl	1.00B	Portuguese	pt	1.00B
English [†]	en	1.00B	Romanian	ro	1.00B
Estonian [†]	et	1.00B	Russian [†]	ru	1.00B
Finnish	fi	1.00B	Serbian	sr	1.00B
French	fr	1.00B	Slovak	sk	1.00B
Galician	gl	1.00B	Slovenian	sl	1.00B
German [†]	de	1.00B	Spanish*	es	1.00B
Greek	el	1.00B	Swedish	sv	1.00B
Hindi	hi	1.00B	Turkish*	tr	1.00B
Hungarian	hu	1.00B	Ukrainian [†]	uk	1.00B
Irish	ga	591.1M	Valencian	val	39.3M
Chinese (Simp.) [†]	zh	1.00B	Vietnamese [†]	vi	1.00B
Welsh	cy	723.9M			
Total				38.88B	

Table 5: Languages carried over from SALAMANDRATA. † appears in a shared-task direction; * included as a transfer anchor (Italian for Ligurian and Ladin, Turkish for Kazakh, Catalan and Spanish as SALAMANDRA anchors). Counts capped at 1B are exact to three significant figures.

B Instruction templates

Figure 1 gives the pool from which CPT2 training records were built. The pairing of templates with and without a source-language name is systematic except for {src_lang} -> {tgt_lang} translation request, which has no source-language-free counterpart. Placeholders are filled with the language names and the cleaned, normalised sentence pair, the target text following the rendered prompt.

Direct

Translate the following {src_lang} text to {tgt_lang}. Source text: {src_text}
Translate the following text to {tgt_lang}. Source text: {src_text}
Please translate this text from {src_lang} into {tgt_lang}: {src_text}
Please translate this text into {tgt_lang}: {src_text}
Translate from {src_lang} to {tgt_lang}: {src_text}
Translate to {tgt_lang}: {src_text}
Convert this {src_lang} text into {tgt_lang}: {src_text}
Convert this text into {tgt_lang}: {src_text}
I need the following {src_lang} text translated into {tgt_lang}: {src_text}
I need the following text translated into {tgt_lang}: {src_text}

Role-based

Provide a high-quality translation from {src_lang} to {tgt_lang} for: {src_text}
Provide a high-quality translation to {tgt_lang} for: {src_text}
As a translator, rewrite this {src_lang} passage in {tgt_lang}: {src_text}
As a translator, rewrite this passage in {tgt_lang}: {src_text}
Perform a fluent translation of this {src_lang} text into {tgt_lang}: {src_text}
Perform a fluent translation of this text into {tgt_lang}: {src_text}

Structured

Task: Translate {src_lang} to {tgt_lang}. \nInput: {src_text}
Task: Translate to {tgt_lang}. \nInput: {src_text}
Source Language: {src_lang}\nTarget Language: {tgt_lang}\nText: {src_text}
Target Language: {tgt_lang}\nText: {src_text}
{src_lang} -> {tgt_lang} translation request:\n{src_text}
Input ({src_lang}): {src_text}\nOutput ({tgt_lang}):
Input {src_text}\nOutput ({tgt_lang}):
Translate the text below.\nFrom: {src_lang}\nTo: {tgt_lang}\nContent: {src_text}
Translate the text below.\nTo: {tgt_lang}\nContent: {src_text}

Action-oriented

Transform the following sentences from {src_lang} to {tgt_lang}: {src_text}
Transform the following sentences to {tgt_lang}: {src_text}
Express the meaning of this {src_lang} text in {tgt_lang}: {src_text}
Express the meaning of this text in {tgt_lang}: {src_text}
Generate a {tgt_lang} version of this {src_lang} text: {src_text}
Generate a {tgt_lang} version of this text: {src_text}
Rewrite the following {src_lang} text using {tgt_lang}: {src_text}
Rewrite the following text using {tgt_lang}: {src_text}
Give me the {tgt_lang} translation for the following {src_lang} string: {src_text}
Give me the {tgt_lang} translation for the following string: {src_text}

Figure 1: The CPT2 instruction-template pool, reordered so that each template sits next to its source-language-free counterpart.

C Data volume by direction

Table 6 collects the parallel data available to each shared-task direction in CPT2 together with the size of its preference set. The instruction mixture (Table 9) is organised by task and source rather than by direction, and several sources cover many directions without a per-direction breakdown, so SFT volume cannot be decomposed per direction and is not given here.

Three directions entered SFT with no parallel data from either CPT round; a fourth, en-arz, entered with generic rather than Egyptian Arabic. For Ligurian, Ladin and Northern Sámi this was a coverage decision: after cleaning, 1.7M, 9K and 118K tokens respectively remained, below the threshold at which inclusion in the CPT2 mix would have contributed usefully. Ladin’s figures are the smallest in the table at

Direction	CPT2 parallel		PO
	Pairs	Tokens	Pairs
en-cs	104.1M	5.00B [†]	
en-de	85.3M	5.00B [†]	
en-id	88.6M	5.00B [†]	
en-ru	73.0M	5.00B [†]	
en-zh	85.6M	4.34B	
cs-de	62.9M	3.08B	
en-uk	28.3M	2.76B	
en-ja	53.4M	2.59B	
en-ko	43.3M	2.12B	
en-et	30.2M	1.70B	
en-th	25.4M	1.28B	4,000
zh-ja	15.5M	0.79B	
en-is	15.7M	0.77B	
cs-uk	14.7M	0.70B	
en-be	16.4M	0.67B	2,997
cs-vi	15.0M	0.52B	
en-kk	11.9M	0.50B	
en-hy	5.6M	0.31B	
en-zh_TW	9.4M	0.23B	
<i>Unfiltered for variety; base for en-arz</i>			
en-ar	73.7M	5.00B [†]	5,000 [‡]
<i>Transfer support, not a shared-task direction</i>			
en-it	79.5M	5.00B [†]	
<i>Excluded from CPT2: insufficient data after cleaning</i>			
en-lij		1.7M	7,252
en-se		118K	748
en-lld		9K	93

Table 6: Parallel data by direction. CPT2 token counts are source-plus-target, excluding instruction-template overhead; [†] marks directions truncated at the 5B cap. The three excluded directions were reserved for SFT. [‡] The shared-task direction is en-arz. No variety filtering was applied in either CPT round, so the 5.00B en-ar tokens are generic Arabic; the Egyptian variety is introduced only at SFT, and the 5,000 preference pairs are en-arz. Preference pairs are final counts from Table 11; blank entries received no preference optimisation.

every stage, and it also received the smallest realistic share of the 2,841 resampled examples distributed across four directions in the instruction mixture.

D Full per-direction results

Tables 7 and 8 give per-direction scores for all three checkpoints. English–Northern Sámi is absent from both: no internal evaluation exists for that direction. English–Arabic and English–Polish were evaluated in error and are not shared-task directions; they are excluded throughout. Note that en-ar here is the generic-Arabic evaluation run in place of en-arz, distinct from the en-ar training data discussed in §2.2.

E Inference pipeline details

Prompt templates. The eleven prompt types of §4 are realised by four template families, listed in Figure 2. Four of the types (plain translation, inclusive gender, ASR transcripts and Korean ellipsis) share the generic template, differing only in the segmentation applied upstream; the five structured genres (news HTML, educational HTML, social media with and without screenshot context, and software-localisation JSON) share the markup-preserving template; legal terminology and game dialogue have one each. Each of the generic and structured families has a context-carrying variant, applied to every unit after the first when a document splits into more than one unit, giving six concrete strings. Line breaks marked \n are literal; wrapping in the figure is ours.

JSON parsing. Software-localisation inputs are parsed tolerantly. After stripping Markdown code fences we fall back to a string-aware scanner that removes trailing commas before] and }, leaving

Direction	BLEU			chrF			COMET		
	CPT2	SFT	Final	CPT2	SFT	Final	CPT2	SFT	Final
cs-de	30.74	30.34	30.54	58.97	59.18	59.24	0.858	0.863	0.863
cs-uk	25.26	24.80	24.54	54.14	53.39	53.30	0.901	0.894	0.895
cs-vi	33.57	33.53	33.81	52.90	53.07	53.34	0.874	0.875	0.877
en-arz	13.89	16.22	16.43	47.01	49.16	49.56	0.816	0.827	0.828
en-be	14.08	15.92	16.64	45.12	46.91	47.72	0.826	0.863	0.870
en-cs	33.83	35.28	34.88	60.26	61.17	61.00	0.891	0.915	0.911
en-de	39.64	41.38	41.57	65.87	67.03	67.06	0.862	0.874	0.871
en-et	26.59	30.22	30.18	59.23	61.96	62.03	0.889	0.914	0.914
en-hy	20.98	24.87	24.39	56.11	59.54	59.20	0.870	0.900	0.898
en-id	46.44	47.44	47.20	70.40	71.37	71.14	0.901	0.911	0.910
en-is	25.69	29.15	28.97	53.04	55.80	55.83	0.812	0.865	0.864
en-ja	33.65	33.61	33.93	40.05	41.20	41.48	0.884	0.904	0.906
en-kk	23.32	23.90	23.80	58.20	58.83	58.75	0.896	0.900	0.898
en-ko	29.89	30.09	30.00	37.26	36.93	36.81	0.871	0.889	0.889
en-lij	4.42	27.53	28.14	37.41	55.07	55.53	0.565	0.619	0.620
en-lld	2.90	2.79	2.58	29.93	27.52	27.54	0.501	0.499	0.498
en-ru	31.75	33.90	33.61	58.98	60.25	60.20	0.879	0.896	0.894
en-th	11.64	11.68	12.66	54.79	54.03	54.54	0.851	0.874	0.877
en-uk	31.12	32.26	32.01	59.55	60.20	60.02	0.887	0.899	0.899
en-zh	42.34	45.10	45.16	38.37	39.67	39.49	0.855	0.873	0.874
en-zh_TW	31.23	36.30	36.07	28.20	32.22	31.92	0.855	0.878	0.878
zh-ja	27.83	27.17	27.23	35.04	34.48	34.54	0.894	0.892	0.895

Table 7: Pipeline-stage results on FLORES+ devtest. Best stage per direction and metric in bold.

Direction	BLEU			chrF			COMET		
	CPT2	SFT	Final	CPT2	SFT	Final	CPT2	SFT	Final
cs-de	24.33	28.01	28.15	53.66	56.73	56.79	0.805	0.824	0.822
cs-uk	20.70	26.07	25.52	47.18	52.34	51.91	0.824	0.850	0.845
cs-vi	26.02	26.39	26.56	46.67	47.17	47.28	0.792	0.814	0.814
en-cs	27.32	39.59	39.68	53.95	63.56	63.79	0.803	0.880	0.882
en-de	28.81	39.49	39.65	57.48	65.03	65.17	0.785	0.848	0.849
en-et	25.76	38.62	38.71	55.34	66.14	66.25	0.795	0.896	0.896
en-id	30.92	33.14	33.17	59.37	61.33	61.38	0.804	0.847	0.845
en-is	18.65	30.95	31.20	43.92	55.22	55.65	0.706	0.847	0.849
en-ja	19.35	38.61	38.45	28.73	45.13	45.12	0.778	0.891	0.889
en-ko	25.62	40.68	40.67	32.14	45.36	45.36	0.790	0.890	0.889
en-ru	20.41	32.44	32.09	49.20	56.85	56.60	0.759	0.837	0.832
en-th	7.13	10.00	9.57	46.06	49.04	48.48	0.745	0.833	0.833
en-uk	25.52	39.96	39.86	52.75	63.61	63.44	0.783	0.875	0.875
en-zh	33.93	56.22	56.93	31.02	52.13	52.63	0.741	0.873	0.876
en-zh_TW	14.86	33.42	33.31	18.98	30.45	30.25	0.747	0.845	0.848
zh-ja	18.43	24.76	24.70	27.48	31.80	31.66	0.838	0.862	0.861

Table 8: Pipeline-stage results on WMT24++, for the 16 directions it covers. Best stage per direction and metric in bold. Instruction-tuning gains are an order of magnitude larger here than on FLORES+ for the same checkpoints and directions; see §6.

commas inside quoted strings untouched, so that these documents split into short value-level units rather than being translated as a single oversized block. Only string values are extracted; keys, non-string values and placeholders are preserved as reconstruction metadata.

Reconstruction. JSON documents are rebuilt by substituting only the translated string values and re-emitted as valid JSON with the original code fence restored. HTML paragraphs are normalised to exactly one <p> wrapper per unit, preserving inner markup, which keeps document-level paragraph structure aligned with the source. The content-preserving repairs applied are inclusive masculine and feminine slash forms for Czech, Croatian and Russian, and removal of copied legal candidate labels.

- (1) Generic
Translate from {src_lang} to {tgt_lang}: {source}
- (2) Generic, with context
Context: {previous_context}\n
Translate only this current {src_lang} text to {tgt_lang}; use the context only for disambiguation: {source}
- (3) Structured (news/educational HTML, social media, software-localisation JSON)
Source Language: {src_lang}\n
Target Language: {tgt_lang}\n
Instructions: Keep all markup, structural tokens, and formatting verbatim and in place. Translate only the readable text, leaving the layout identical.\n
Text: {source}
- (4) Structured, with context
Context: {previous_context}\n
Translate only the current text below. Use the context only for disambiguation.\n
Source Language: {src_lang}\n
Target Language: {tgt_lang}\n
Instructions: Keep all markup, structural tokens, and formatting verbatim and in place. Translate only the readable text, leaving the layout identical.\n
Text: {source}
- (5) Legal terminology
Translate from Italian to {tgt_lang}. Pay close attention to the required terms.\n\n
Legal context: standard German for {jurisdiction}.\n\n
Required Terminology:\n\n
- {source_term} -> choose exactly one listed candidate\n\n
Candidates:\n\n
- {candidate_1}
- {candidate_2}
... \n\n
Source Text: \n\n{source}
- (6) Game dialogue
Context: {speaker_role}\n
Translate this {src_lang} sentence to {tgt_lang}: {source}

Figure 2: Inference prompt templates. The jurisdiction field defaults to “the target legal system” when absent from the instruction. When no source term is extracted, the bullet under Required Terminology is empty but the heading is still emitted. The speaker role in template (6) is the speaker’s in-game role, parsed from the blindset instruction.

F Instruction mixture and hyperparameters

This appendix reports the full training configuration for the checkpoints in §5.1. Table 9 and Figure 3 give the SFT instruction mixture and its task distribution; Figure 4 compares the adapted tokeniser’s fertility against open baselines; Table 10 lists the SFT hyperparameters; and Tables 11 and 12 give the preference-optimisation data and hyperparameters.

G Full test-set results

This appendix expands Table 3: per-direction chrF for all nine full-coverage constrained systems (Table 13), and the refA comparison for the two directions with two references (Table 14).

Task	Source	Languages	Count
General Machine Translation	FLORES+	en paired with (arz, be, cs, de, et, hy, id, is, ja, kk, ko, lij, ru, th, uk, zh, zh_TW); cs-de, cs-uk, cs-vi, zh-ja	20,937
	ALT	en-id, en-ja, en-th, en-vi, en-zh, zh-ja	12,000
	TATOEBA	cs-de, cs-en, cs-uk, en-et, en-is, en-ja, en-ko, en-ru, en-uk, en-zh, ja-zh	11,000
	WMT24++	en-arz, en-cs, en-de, en-et, en-is, en-ja, en-ko, en-ru, en-uk, en-zh	9,600
	SMOL (SMOLSENT)	en-arz, en-id, en-ja, en-ko, en-lij, en-th	5,711
	ARZEN-MULTIGENRE	en-arz	5,000
	NTEU	en-cs, en-de, en-et	2,688
	Resampled WMT26 parallel data	en-be, en-ld, en-lij, en-se	2,841
	EU press releases	en-cs, en-de, en-et	1,500
	NUSAX-MT	en-id	500
Paragraph-level Translation	NEWS COMMENTARY	cs-de, en-ja, en-ru, en-zh, ja-zh	3,043
	NEWSPALM	en-de	1,000
	ACADTRAIN	en-de	500
Document-level Translation	EUROPARL	cs-de, cs-et, cs-it, en-et, en-it	5,000
	SMOL (SMOLDOC)	en-arz, en-id, en-ja, en-ko, en-lij, en-ru, en-th	1,304
	PROJECT GUTENBERG	de-en	485
Multi-reference Translation	TATOEBA	cs-de, cs-en, cs-uk, en-cs, en-de, en-et, en-is, en-ja, en-ko, en-ru, en-uk, en-zh, ja-zh	4,959
Terminology-aware Translation	MESPEN (synthetic glossaries)	en-ar, en-de, en-id, en-it, en-ja, en-ko, en-ru, en-th, en-tr, en-zh	2,033
	WMT25 TERMINOLOGY TASK (Track 2)	en-zh_TW	1,600
	WMT23 TERMINOLOGY TASK	en-cs, en-de, en-zh	1,500
	EMEA (document-level)	cs-de, en-et	1,014
	WMT25 TERMINOLOGY TASK (Track 1)	en-de, en-ru	1,000
Formality-controlled Translation	Synthetic (GEMMA 4)	en paired with (arz, be, cs, de, et, hy, id, is, ja, kk, ko, ru, th, uk, vi, zh, zh_TW); cs-de, cs-uk, cs-vi, zh-ja	5,250
Formality Register Transfer	Synthetic (GEMMA 4)	en paired with (arz, be, cs, de, et, hy, id, is, ja, kk, ko, ru, th, uk, vi, zh, zh_TW); cs-de, cs-uk, cs-vi, zh-ja	2,100
HTML Tag Transfer	FLORES+ (HTML-augmented)	en-ar, en-cs, en-de, en-et, en-is, en-ja, en-ko, en-ru, en-uk, en-zh	5,000
	NEWS COMMENTARY (HTML-augmented)	ar-en, cs-de, cs-en, de-en, en-ja, en-ru, en-zh, ja-zh	2,296
	EUROPARL (HTML-augmented)	cs-de, cs-en, cs-et, de-en, de-et, en-et	995
	ACADTRAIN (HTML-augmented)	en-de	250
Total			111,106

Table 9: Overview of tasks, data sources, language coverage, and counts in the instruction mixture.

Table 10: Hyperparameters for supervised-fine tuning.

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

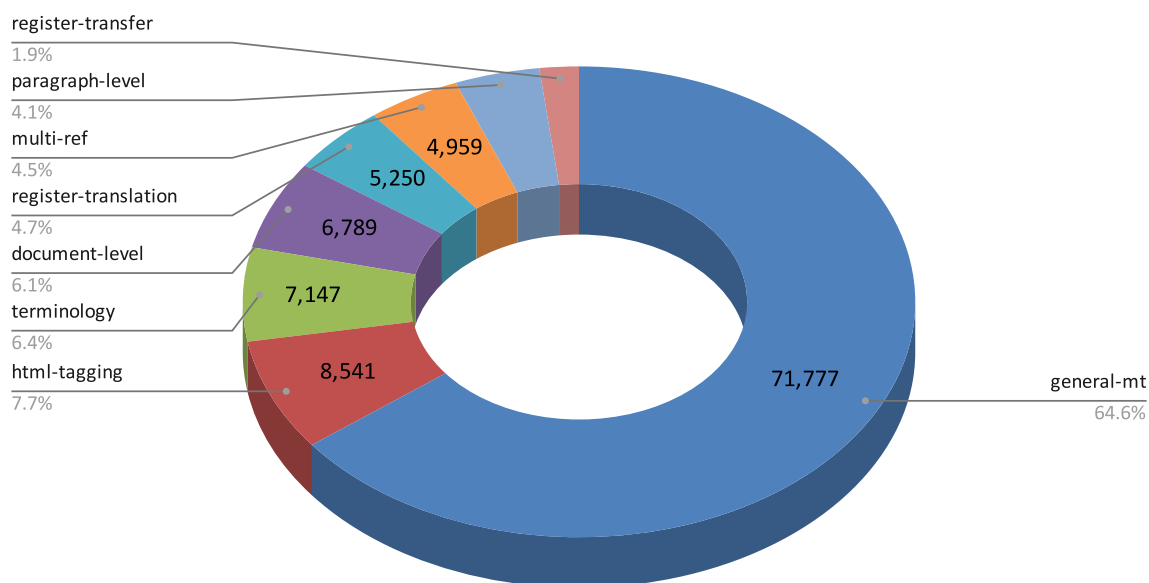


Figure 3: Distribution of tasks in the SFT instruction mixture.

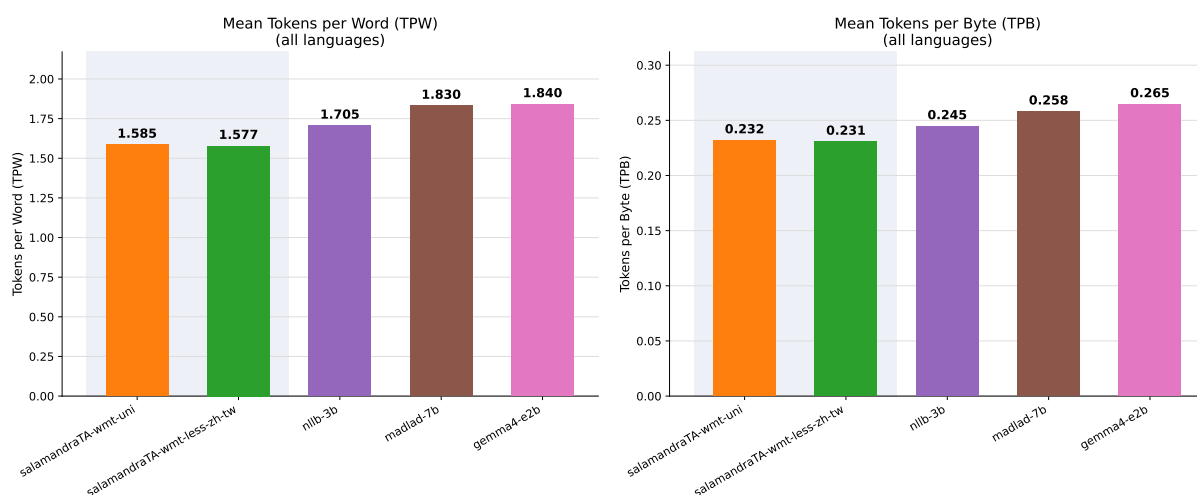


Figure 4: Tokeniser fertility comparison (mean tokens per word and tokens per byte across all evaluated languages). salamandraTA-wmt-uni and salamandraTA-wmt-less-zh-tw perform comparably and both outperform the nllb-3b, madlad-7b and gemma4-e2b baselines on both metrics.

Direction	Source	Initial Count	Final Count
en-arz	WMT24++	960	960
	FLORES+	997	997
	SMOL (SMOLSENT)	863	863
	ARZEN-MULTIGENRE	2,180	2,180
en-th	FLORES+	997	997
	SMOL (SMOLSENT)	753	753
	ALT	2,250	2,250
en-be	FLORES+	997	997
	Resampled WMT26 parallel data	2,000	2,000
en-se	Resampled WMT26 parallel data	748	748
en-lll	Resampled WMT26 parallel data	93	93
en-lij	FLORES+	997	1,994
	SMOL (SMOLSENT)	863	1,726
	Resampled WMT26 parallel data	1,766	3,532
Total		16,464	20,090

Table 11: Overview of translation directions, sources, and sample counts (initial vs. final).

Table 12: Hyperparameters for preference optimization fine-tuning.

Hyperparameter	Value
Train epochs	1
Train batch size per device	1
Gradient accumulation steps	16
Learning rate	5e-7
Beta	0.6
Loss Type	simpo
CPO alpha	1.0
SimPO gamma	1.0
LoRA alpha	32
LoRA rank	16
LoRA dropout	0.0

System	cs-de	cs-vi	en-cs	en-de	en-et	en-is	en-arz	en-ja	zh-ja	en-lij	en-lll	en-se	Mean
Lumen	75.7	70.0	61.0	66.8	66.0	65.9	60.9	51.1	44.8	51.9	51.8	58.8	60.4
Qingqiu-MT-9B	78.0	70.5	58.5	66.7	62.6	62.1	63.7	51.2	44.9	53.4	51.5	20.3	56.9
SalamandraTA (ours)	69.9	59.5	55.0	62.1	62.2	60.6	47.8	40.1	30.2	53.1	36.0	47.6	52.0
Gemma 4-4B	68.8	63.3	49.9	60.7	51.2	47.3	55.1	41.9	35.8	41.3	35.5	29.1	48.3
GPT-OSS-20B	63.8	52.4	49.2	58.0	52.6	46.6	43.2	35.8	28.0	39.2	34.1	31.7	44.5
Qwen 3.5-9B	64.8	54.3	49.6	56.8	52.4	47.7	41.4	34.7	29.6	38.4	34.9	33.2	44.8
Tower 9B	66.6	48.5	52.5	58.7	45.3	57.6	32.2	38.2	31.7	40.9	34.6	28.9	44.6
GLM 4-9B	63.5	53.7	49.0	55.3	47.0	39.9	38.7	34.4	27.1	36.9	34.9	32.6	42.8
Ministral 3-14B	63.1	54.4	47.6	55.0	47.9	45.3	39.3	33.5	25.5	39.6	32.7	30.0	42.8

Table 13: Per-direction chrF for the nine constrained-track systems covering all reference directions (breakdown of Table 3), in the order of Table 3.

		BLEU	chrF
cs-de	refPE	49.15	69.88
	refA	48.01	69.46
cs-vi	refPE	42.81	59.45
	refA	42.89	59.46

Table 14: Post-edited (refPE, used in the main comparison) versus from-scratch (refA) reference, for the two directions that provide both. Scores are essentially unchanged between the two reference types.