SPECIALIZING WITHOUT FORGETTING: ANALYZING KNOWLEDGE PRESERVATION IN MULTILINGUAL MODEL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While continual pretraining (CPT) is a practical way to extend large language models to new languages, naïve finetuning often erodes existing capabilities through catastrophic forgetting. We investigate which model layers drive this trade-off, and whether interventions at these layers can guide knowledge preservation during adaptation. We interpolate `gemma-3-4b` model states before and after CPT on five language families to localize forgetting on reading comprehension and translation, finding that middle-layer reversion yields the largest comprehension recovery, while translation effects vary by language family and direction. Guided by these findings, we evaluate CPT strategies that leverage this layer information to mitigate forgetting: layer freezing, layer-range L2 regularization, post-hoc layer reversion, and model souping, comparing all strategies against joint multilingual and family-specific vanilla CPT baselines. We find that preserving the layer weights identified via model interpolation substantially reduces comprehension loss relative to joint CPT, with layer freezing exceeding base model performance on average. However, these strategies yield mixed translation results: dense training or post-hoc reversion often outperforms both training-time constraints and family-specific specialization, complicating prior assumptions about *how* models should be aligned when extended to new tasks. Instead, we argue that multilingual adaptation strategy should be informed by target language, base model knowledge, and downstream task, and propose interpolation-based localization as a diagnostic for identifying candidate layers before committing to a training-time intervention in a new setting.

1 INTRODUCTION

Adapting Large Language Models (LLMs) through continual pretraining (CPT) is a practical method of expanding model coverage to new languages while avoiding the prohibitive compute costs of pretraining from scratch (Zhao et al., 2025; Dou et al., 2024). However, naïve CPT on target languages yields strong language acquisition but often causes catastrophic forgetting (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017) of the model's original knowledge; this is especially true in multilingual settings, where the curse of multilinguality (Conneau et al., 2020) forces parametric trade-offs between language coverage and preserving existing capabilities. Furthermore, this balance cannot be assessed through language modeling alone, as lower target-language perplexity does not imply better downstream task performance (Tay et al., 2022; Li et al., 2022a).

Mechanistic analysis of LLMs has found that middle layers of transformer LMs are associated with language-agnostic processing, while early and late layers handle language-specific inputs and outputs (Wendler et al., 2024; Zhao et al., 2024; Dumas et al., 2025). Recent work on cross-lingual reasoning exploits this separation during adaptation, showing that language and mathematical capabilities can be effectively combined by composing the outer layers from a language-adapted model with the middle layers from a reasoning-adapted model (Bandarkar et al., 2025), and that similar benefits can be obtained simply by freezing or reverting the middle layers during adaptation (Bandarkar & Peng, 2025). However, these findings come from supervised fine-tuning on a single task with limited CPT experiments, leaving open whether they extend to multilingual continual pretraining, cross-lingual evaluation, or generation tasks.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This paper generalizes this layer-aware preservation approach by first identifying *which* layers specifically drive task knowledge change during joint multilingual (Dense) CPT with a new post-hoc interpolation diagnostic method (§3), rather than relying on general insights from prior work. We then test whether preserving the layer parameters identified by this method improves knowledge acquisition and retention during CPT (§4). Through these experiments, we test: (1) which layer updates drive comprehension loss and translation gains; (2) whether these effects are consistent across language families and tasks; and (3) whether layers identified through post-hoc intervention make effective training-time constraints. Specifically, we experiment with layer-aware preservation techniques (using the most impactful layers identified by our post-hoc analyses) for multilingual CPT of `gemma-3-4b` (Team et al., 2025) on 32 languages from five language families, comparing joint multilingual (Dense) CPT with independently trained family experts. We measure language acquisition with perplexity and downstream capabilities with reading comprehension and translation, on both CPT training languages (held-in) and related held-out languages.

Our results show that the benefits of layer reversion are **strongly** task-dependent (Figure 1). In our post-hoc analyses, middle-layer reversion produces the largest comprehension recovery, corroborating prior work, but harms translation: restoring layers $[10, 16)$ of the Dense model to Base improves reading comprehension accuracy by 5.9 points but reduces translation by 3.5 ChrF.[1] We also find that layer-wise improvements identified by our interpolation analysis transfer to CPT for comprehension, as freezing, L2 regularization, and post-hoc reversion of layers $[10, 16)$ all exceed unregularized family experts, Dense CPT, and the Base model on held-in reading comprehension, while also obtaining lower in-language perplexity than Base.

However, unlike prior work, the trends we observe for translation and held-out languages are less consistent: for translation, layer-preserving training paradigms with the identified layers [5,11) do not cleanly transfer to better results. Instead, post-hoc reversion of these layers improves over the respective baselines for both the Dense model and most family experts, whereas freezing or regularizing the same window fails to improve over baseline in three of five families. Moreover, dense training on multiple languages often outperform specializing to the target languages, particularly for the lower-resourced Indic and Austronesian families. For held-out languages, we similarly see that while targeted layer preservation improves comprehension over vanilla experts, no intervention approach surpasses the Base model; our translation gains also only hold when translating held-out languages to English (XX→EN: $31.8 \rightarrow 51.3$ ChrF under Dense CPT, versus $34.4 \rightarrow 34.6$ for EN→XX).

These results indicate that the layer regions that store downstream knowledge vary across tasks and that transferring layers to preserve across tasks can significantly harm performance. Furthermore, while the regions identified by our post-hoc interpolation analysis can inform better, layer-aware CPT strategies, the benefits only seem to hold if downstream task knowledge is otherwise lost during targeted CPT (e.g., on comprehension). We also note that language family specialization is not always the best method for improving target language task performance, contrary to prior assumptions. Thus, rather than a single preservation strategy for improving CPT, we argue that multilingual adaptation should be informed by the target task, base model knowledge, and languages, with interpolation-based localization used to identify candidate layers *before* committing to a training-time intervention.

The primary contributions of this work are:

- A new diagnostic method that uses layer interpolation to systematically identify which layer ranges drive performance gains and losses during model adaptation, along with a corresponding analysis demonstrating that the layers affecting performance changes after multilingual CPT can vary significantly across task types and target languages (§3).
- A controlled comparison of training-time parameter preservation strategies (layer freezing, layer-range L2-SP) against post-hoc reversion and souping, applied at the identified windows by our interpolation analysis across five language families, showing that localization can transfer training-time preservation for comprehension but not for translation (§4).
- An examination of the acquisition–forgetting trade-off after CPT on both held-in and held-out languages, identifying the limits of improving language-specific behavior while preserving downstream performance from the base model (§4.5).

[1] In contrast, restoring late layers $[27, 33)$ reduces translation by 18.5 ChrF while barely affecting comprehension accuracy (0.4).

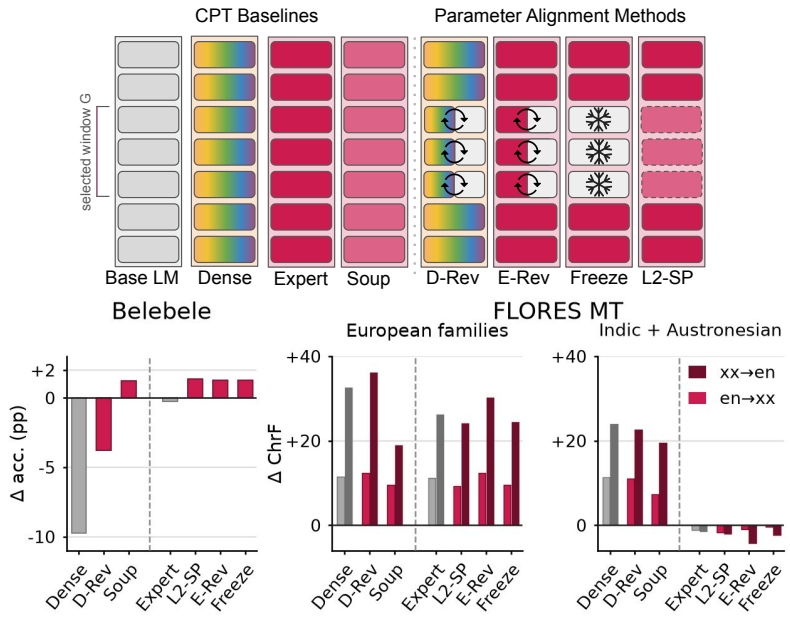

Figure 1: (a) CPT baselines and targeted parameter alignment methods. (b) Held-in changes relative to Base. D-Rev and E-Rev restore Base weights after CPT; Freeze holds the selected window G fixed during CPT, and L2-SP regularizes it. The targeted methods use $G = [10, 16)$ for Belebele and $G = [5, 11)$ for FLORES. The FLORES panels show translation directions separately and average the named families within each group.

2 Experimental Setting

In this paper, we investigate how to leverage layer-specific parametric knowledge to mitigate catastrophic forgetting while retaining the benefits of language-specific adaptation during multilingual continual pretraining. We first use layer-wise interventions to characterize the downstream effects of CPT updates (§3), then test whether these findings can guide targeted preservation during training (§4). In this section, we detail our overall experimental setup, which performs joint multilingual and family-specific CPT from a shared base (`gemma-3-4b`) checkpoint with various parameter-alignment strategies(Figure 1), and evaluates the resulting models on reading comprehension, translation, and in-domain perplexity.

2.1 Model, Data and Language Coverage

All our experiments build on the text-only pre-trained version of `gemma-3-4b` (Team et al., 2025), which we refer to as the base checkpoint. We construct a dataset for continued pretraining from MADLAD-400 (Kudugunta et al., 2023), covering 32 languages across the Slavic, Germanic, Indic, Austronesian, and Romance language families; we additionally evaluate model generalization on 22 held-out languages across these families (Table 1).[2] For each family, we sample 5B token corpus split uniformly across languages; Dense CPT trains on the union of the five training corpora (25B tokens). Documents are tokenized with the Gemma-3 tokenizer at a maximum sequence length of 2,048 tokens, using a 95%/5% train/validation split.

2.2 Dense, Family-Specific CPT and Souping

We consider three unregularized CPT settings initialized from the same Base checkpoint :

[2]We note that, since the Gemma 3 training language distribution is not publicly reported, we are unable to verify whether the held-out languages were seen during the original training process.

Family	Training Languages	Held-Out Languages
Slavic	Macedonian (mk), Croatian (hr), Russian (ru), Slovak (sk), Serbian (sr), Ukrainian (uk)	Bulgarian (bg), Czech (cs), Polish (pl), Slovenian (sl), Lithuanian (lt), Latvian (lv)
Germanic	Afrikaans (af), Frisian (fy), Luxembourgish (lb), Danish (da), Dutch (nl), English (en)	Swedish (sv), Norwegian (no), Icelandic (is)
Indic	Bengali (bn), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), Nepali (ne), Tamil (ta), Telugu (te)	Assamese (as), Gujarati (gu), Odia (or), Punjabi (pa), Sinhala (si), Urdu (ur), Sindhi (sd)
Austronesian	Samoan (sm), Javanese (jv), Cebuano (ceb), Filipino (fil), Indonesian (id), Malay (ms)	Ilocano (ilo), Waray (war), Malagasy (mg), Māori (mi), Sundanese (su)
Romance	Spanish (es), Portuguese (pt), French (fr), Galician (gl), Italian (it), Romanian (ro)	Catalan (ca)

Table 1: Language families, their training languages, and held-out languages for evaluating within-family generalization.

Dense CPT trains a single model jointly on the combined corpus spanning all 32 training languages.

Family Expert trains a separate model on each language family initialized from the Base model.

Expert Soup After training five family experts, we merge them into a single model by **uniformly averaging** their weights: $\theta_{\text{soup}} = \frac{1}{5}\sum_{f=1}^{5}\theta_f$, where θ_f are the weights of family f's expert.

The Dense checkpoint provides the starting point for the layer-interpolation analysis in §3. Family-specific checkpoints allow us to examine whether the identified intervention effects transfer across adaptation settings. Souping tests whether uniformly averaging independently trained family experts can recover base capabilities without any training-time constraint.

2.3 Training and Evaluation

Training CPT uses bfloat16 precision and gradient checkpointing. The Dense and family-specific runs use different training schedules and stopping criteria, detailed in Appendix A.3. The schedules and comparison controls for the targeted preservation experiments are specified in §4.

Evaluation We measure reading comprehension using Belebele (Bandarkar et al., 2024), translation using FLORES-200 ChrF in both translation directions (Team et al., 2022), and language acquisition using perplexity on held-out MADLAD text. Belebele and FLORES evaluations use `lm-eval-harness` in a 2-shot setting (Gao et al., 2024). FLORES scores follow a shared first-line post-truncation protocol across checkpoints (Appendix A.6).

3 Understanding Layer-Aware Adaptation

An open question for model adaptation is which regions mediate knowledge retention, under the hypothesis that selectively preserving these parameter values during CPT could mitigate catastrophic forgetting. Prior work suggests that the middle layers of transformer LMs support semantic and reasoning computations (Bandarkar & Peng, 2025; Bandarkar et al., 2025; Wendler et al., 2024), while the initial and final layers encode language-specific information (Tang et al., 2024; Kojima et al., 2024; Zhao et al., 2024). In this section, we test whether these layer intuitions hold for downstream reading comprehension and translation across language families. Specifically, we use interpolation analysis to localize which layer regions lead to the largest task performance changes after continual pretraining by interpolating between the original base and updated parameters within the target layers (Figures 2 and 3).

3.1 Causal Analysis of Layer Drift

We perform a targeted interpolation analysis on the unregularized Dense CPT model. For a layer group G, we replace that group's parameters with $\theta_G(\alpha) = \theta_G^0 + \alpha(\theta_G^{\text{Dense}} - \theta_G^0)$, while keeping all other layers fixed to the Dense CPT checkpoint. α=0 reverts the selected layer group to the base model and

α=1 recovers the original Dense CPT parameters for that group. We sweep $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ over the first (9), middle (19), and last (6) layer groups [3] and evaluate held-in language Belebele accuracy and ChrF for FLORES. These interventions measure the effect of replacing a selected group's CPT updates while holding the rest of the remaining weights fixed.

Figure 2 shows that retaining more of Dense CPT updates in the middle layer produces a monotonic 7.69 point drop in Belebele accuracy. Middle-layer reversion therefore produces the largest recovery among these interventions. (Appendix A.4 has family-wise results)

FLORES exhibits a different layer-sensitivity profile, with effects varying across language families and translation direction (Figure 3). The numbers are reported relative to the corresponding Dense endpoint i.e. $\Delta\text{ChrF} = \text{ChrF}(\alpha = 0) - \text{ChrF}(\alpha = 1)$. Reverting the last layer group reduces EN→XX performance in every family, with losses ranging from 3.57 to 15.77 ChrF. For XX→EN, the same intervention harms four of the five families; Indic is the exception, improving by 1.65 ChrF. These translation effects contrast with the small Belebele change under last-group reversion.

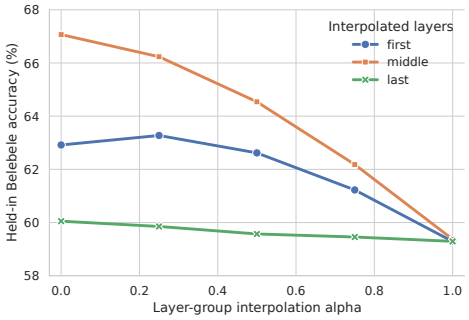

Figure 2: Layer interpolation between the base model and Dense CPT for held-in Belebele accuracy. All non-interpolated layers are kept at their Dense CPT values.

The effects of first-group reversion are more mixed. For XX→EN, restoring base model weights improve Slavic, Germanic, and Romance by 3.49, 0.89, and 4.44 ChrF respectively, but reduces Indic and Austronesian performance by 3.27 and 1.72 ChrF. The corresponding EN→XX gains for Slavic, Germanic, and Romance are 0.95, 1.08, and 0.89 ChrF. Thus, the benefit is more pronounced into English for Slavic and Romance, while Germanic shows small gains in both directions (Figure 3).

Middle-group reversion also produces these mixed results. While the gains are positive across all families, for XX→EN direction ranging from 0.75 to 2.60 ChrF, the same is not true for EN→XX.

3.2 Localizing Task-Dependent Reversion Effects

The three partitions that we studied above differ in width and may have wide variation within each group. We therefore refine this analysis using a fixed-width sliding window reversion, which holds the number of restored layers constant while varying their position. For each six-layer window $G_s = [s, s+6)$, we restore the selected parameters to their Base-model values, corresponding to $\alpha = 0$, while retaining CPT weights everywhere else. This experiment measures the effect of the same-sized intervention at different depths and tests whether the coarse layer-group findings reflect localized sensitivity. This analysis helped us in identifying the highest-scoring window among those evaluated for both Belebele and FLORES (Table 2).

Layers reverted	**Belebele** Δ accuracy (pp)	**FLORES** Δ ChrF
$[0, 6)$	+2.2	+0.26
$[5, 11)$	+4.1	**+0.75**
$[10, 16)$	**+5.9**	−3.47
$[24, 30)$	−0.3	−4.28
$[27, 33)$	+0.4	−18.50
$[28, 34)$	+0.7	−8.00

Table 2: Selected six-layer reversions to Base weights: changes relative to Dense CPT on Belebele (28 languages) and FLORES (20-direction screen). Positive values indicate improvement. Family-wise results in Appendix A.4.

Reverting $[10, 16)$ produces the largest Belebele recovery, improving accuracy by 5.9 pp over Dense CPT, but reduces FLORES performance by 3.47 ChrF. In contrast, reverting the earlier window $[5, 11)$ improves Belebele by 4.1 points and yields a 0.75 ChrF gain on the FLORES screening set (Appendix A.4 has language family wise gains). Late-layer reversion shows another contrast: restoring $[27, 33)$ reduces FLORES performance by 18.50 ChrF, while changing Belebele accuracy by only +0.4 points. A region whose reversion benefits

[3] Motivated by prior outer/middle layer decompositions (Bandarkar & Peng, 2025), we adopt a coarse 9/19/6 partition.

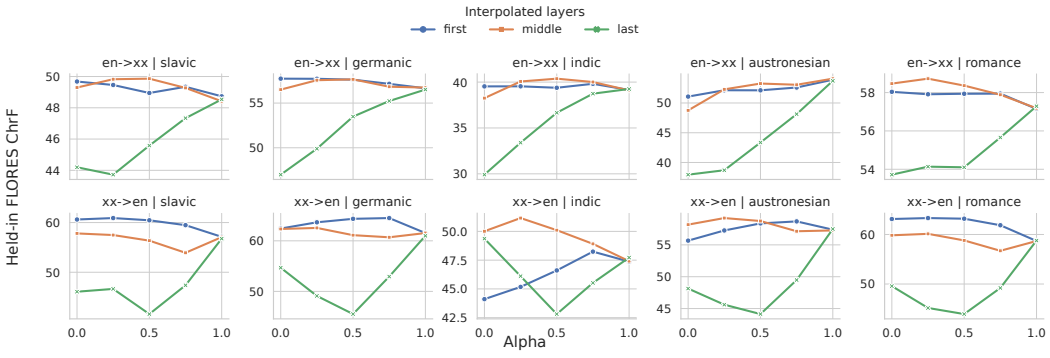

Figure 3: Held-in FLORES ChrF under first-, middle-, and last-layer interpolation, broken down by language family (columns) and translation direction (rows). Scores use the same first-line post-truncation re-scoring protocol as the main FLORES tables.

comprehension need not benefit translation, and translation can be highly sensitive to interventions with little effect on comprehension.

We select $[10, 16)$ for Belebele and $[5, 11)$ for FLORES to perform CPT with freezing and Layer Range L2-SP applied to these layers on the base checkpoint to check how our findings transfer to these layer-specific training methodologies.

4 Testing Targeted Interventions

4.1 Targeted Preservation Strategies

For a selected window G, we compare two training-time preservation strategies, both initialized from Base parameter θ^0.

Figure 1 summarizes our alignment strategies and baselines (§2.2):

Layer Freezing We hold all parameters within G fixed at their base values throughout CPT, while allowing parameters outside G to update. This imposes a hard preservation constraint on the selected region.

Layer-Range L2-SP We allow all parameters to update but penalize deviations from Base within G using L2 starting-point regularization (Li et al., 2018) as adapted by Kumar et al. (2024), with layer-dependent penalty strengths, offering a soft alternative to layer freezing during training. :$\mathcal{L}(\theta) = \mathcal{L}_{\text{CPT}}(\theta) + \lambda \sum_{l \in G} \|\theta_l - \theta_l^0\|_2^2$. For the targeted experiments, we use $\lambda = 0.05$ within the selected window.

Post-hoc Reversion We restore the same window to Base weights after unregularized CPT, retaining the adapted parameters elsewhere. This allows us to compare preservation during training with correction after training.

4.2 Language Acquisition

Table 3 shows the held-in validation perplexity for the two targeted windows across all families. The Dense CPT checkpoint had the best language acquisition compared to all our baselines (7.20) followed by the family experts (7.30). The targeted intervention experiments achieve similar average perplexities (7.56 - 7.62) while still remaining below the base. This indicates that choosing either targeted window preserves broadly comparable in-family acquisition quality. Per-language breakdowns are in Appendix A.8.

	Shared baselines				$[5, 11)$				$[10, 16)$			
Family	**Base**	**Dense**	**Expert**	**Soup**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**
Slavic	7.66	6.41	**6.40**	7.18	6.50	6.50	6.48	6.54	6.57	6.57	6.48	6.55
Germanic	11.41	**7.18**	7.23	9.90	7.50	7.53	7.50	7.54	7.49	7.53	7.49	7.49
Indic	10.07	**7.66**	7.73	8.83	8.02	8.04	8.04	8.09	7.98	8.04	8.00	7.98
Austronesian	11.78	**6.22**	6.74	10.10	7.31	7.31	7.29	7.28	6.91	7.27	7.20	7.20
Romance	10.01	8.55	**8.40**	9.28	8.63	8.67	8.65	8.64	8.82	8.67	8.63	8.68
Average	10.19	**7.20**	7.30	9.06	7.59	7.61	7.59	7.62	7.55	7.62	7.56	7.58

Table 3: Held-in validation perplexity ↓, averaged over training languages within each family. D.-Rev. restores Dense layers in the indicated window to Base; Freeze and L2-SP preserve the indicated layers during family CPT, while E.-Rev. restores the indicated layers to Base after Expert CPT. **Bold** = lowest value within a row.

Family	**Base**	**Dense**	**Expert**	**Soup**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**
Slavic	0.719	0.619	0.726	**0.734**	0.675	**0.734**	0.727	0.731
Germanic	0.758	0.642	0.756	0.761	0.708	0.759	0.762	**0.763**
Indic	0.601	0.535	0.593	**0.629**	0.584	0.624	0.623	0.627
Austronesian	0.668	0.587	0.661	0.682	0.642	0.688	**0.698**	0.692
Romance	0.748	0.625	0.746	0.749	0.696	**0.753**	0.748	0.750
Average	0.699 [0.693, 0.705]	0.602 [0.595, 0.608]	0.696 [0.691, 0.702]	0.711 [0.706, 0.717]	0.661 [0.655, 0.667]	0.712 [0.706, 0.717]	0.711 [0.706, 0.717]	**0.713** [0.707, 0.718]

Table 4: Held-in Belebele accuracy ↑ (2-shot), averaged over training languages within each family. D.-Rev., E.-Rev., Freeze, and L2-SP target $[10, 16)$; Base, Dense, Expert, and Soup are archival baselines. Average intervals are 95% bootstrap CIs, macro-averaged over languages within each family and then over families. **Bold** = best per row.

4.3 Preserving Reading Comprehension

We next evaluate whether targeted interventions preserve reading comprehension after CPT. Table 4 summarizes Belebele results across language families and strategies. Dense CPT causes substantial forgetting, with accuracy dropping 6.6–12.3 percentage points relative to Base across families, whereas vanilla family experts largely retain Base performance. All three targeted interventions outperform vanilla experts in every family. Together with the perplexity results, these comparisons illustrate the acquisition–forgetting trade-off: targeted interventions achieve higher reading-comprehension accuracy but slightly higher in-domain perplexity than vanilla family experts.

4.4 Translation Quality (FLORES-200)

Table 5 reports results for interventions targeting $[5, 11)$[4]. Dense reversion has family-dependent effects: restoring $[5, 11)$ layers improves the Dense family-macro average from 53.7 to 54.7 ChrF. Slavic, Germanic, Austronesian, and Romance improve, whereas Indic declines.

For family experts, post-hoc reversion improves translation in four families, but the training-time preservation yields mixed results. Both freezing and L2-SP outperform vanilla experts on Germanic and Indic, but score less on Slavic, Austronesian and Romance. The consistent gains that we observed earlier with Belebele, do not cleanly transfer here to FLORES. One possible explanation is that anchoring layers to Base constrains updates useful for translation. Also, freezing and L2-SP constrain the selected layers during training and that might change how the rest of the layers adapt, whereas post-hoc reversion retains their unconstrained updates.

We also noticed a huge discrepancy with Austronesian results. While the Dense, D.-Rev. and Expert Soup outperform Base, the vanilla family expert and its targeted variants score below Base.

[4]Appendix A.6 describes the scoring protocol, and Appendix A.7 gives direction-wise results and comparison limitations.

Family	Base	Dense	Expert	Soup	D.-Rev.	E.-Rev.	Freeze	L2-SP
Slavic	33.6	52.7	47.4	46.2	**54.7**	51.3	47.1	46.9
Germanic	34.8	59.0	57.0	50.4	**61.3**	59.7	57.7	58.1
Indic	29.9	**43.4**	36.1	42.6	41.3	36.4	36.3	36.6
Austronesian	33.8	55.5	25.0	48.1	**56.2**	22.2	24.7	23.4
Romance	34.4	57.9	54.9	49.2	**60.2**	56.4	49.5	48.5
Average	33.3 [33.1, 33.5]	53.7 [53.5, 54.0]	44.1 [43.8, 44.4]	47.3 [47.0, 47.6]	**54.7** [54.6, 54.9]	45.2 [44.9, 45.5]	43.1 [42.8, 43.4]	42.7 [42.4, 43.0]

Table 5: Held-in FLORES-200 ChrF ↑ (2-shot), averaged over languages and both directions within each family; Average weights families equally. D.-Rev., E.-Rev., Freeze, and L2-SP target $[5, 11)$. All Average cells include percentile 95% bootstrap CIs. **Bold** = best per row; See Appendix A.7 for scoring and comparison details.

Inspection of their outputs shows a high empty-rate (Appendix A.7) for these checkpoints. Motivated by the late-layer sensitivity observed in the Dense sliding-window analysis, we investigated whether restoring $[27, 33)$ to Base weights mitigate this failure in the Austronesian expert. The reverted checkpoint achieves 40.3 ChrF with a 7.0% empty-output rate, compared with 25.0 ChrF and 57.5% for the vanilla Austronesian expert in Table 5 (29.7 ChrF and 38.9% in a separate generation run of the same expert). This improvement also leads to an increase in held-in perplexity from 6.75 to 7.95, illustrating a trade-off between translation recovery and in-domain language modeling. These results suggest that a localized intervention can mitigate the generation failure, but do not establish its underlying cause (Appendix A.7).

4.5 Out-Of-Domain Generalization

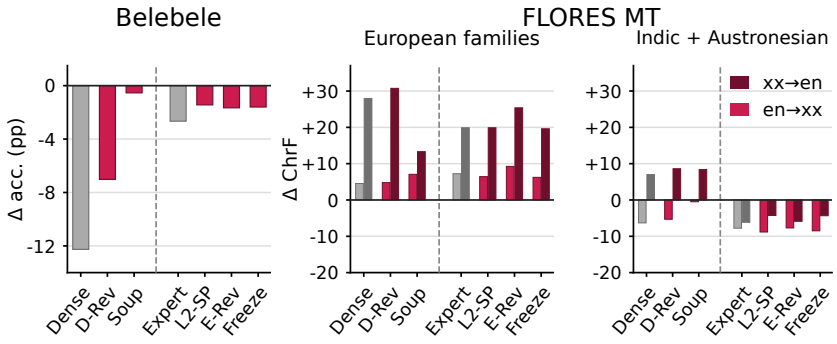

Figure 4: Held-out Belebele accuracy delta and FLORES MT ChrF delta relative to the base model.

We next ask whether the acquisition–forgetting patterns observed on CPT training languages extend to held-out languages within the same family groups. We evaluate 22 languages excluded from CPT (Appendix 1).[5]

Language modeling Dense CPT increases held-out perplexity relative to Base in every family, and family-expert variants generally remain above Base. Expert Soup lowers perplexity in four of five families (Table 11). Thus, the lower perplexity achieved on CPT training languages does not imply improved language modeling on their held-out relatives.

Reading Comprehension Freeze, L2-SP, and E.-Rev. improve held-out Belebele accuracy by approximately 1pp over vanilla experts, but remain below Base in every family, unlike their above-Base held-in averages. Expert Soup comes closest to Base overall (64.7% versus 65.2%. (Table 12, Figure 4).

[5]These languages are held out from our CPT data, not necessarily from Gemma's original pretraining.

Translation Across families and both directions, E.-Rev. improves the Expert reference from 38.4 to 40.8 ChrF, whereas Freeze and L2-SP both average 38.3 (Tables 20 and 21). Dense reversion likewise improves the average from 42.9 to 44.4 ChrF. Transfer is asymmetric: Dense improves xx→en from 31.8 to 51.3 ChrF, but barely changes en→xx (34.4 to 34.6). Vanilla experts and their targeted variants also remain below Base on Indic and Austronesian. Together, these results reinforce that comprehension preservation does not ensure translation improvements, and that beneficial post-hoc reversion need not translate into effective training-time constraints.

5 Related Work

Multilingual Catastrophic Forgetting Catastrophic forgetting (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017) is a central challenge when adapting pretrained models to new languages. Khelli et al. (2025) find that partial parameter sharing can mitigate forgetting in multilingual CPT, while Owodunni & Kumar (2025) study layer-selective fine-tuning but find no clear advantage of parameter-efficient methods over full fine-tuning. However, these methods typically constrain parameters uniformly, or in regions chosen apriori. Our work instead looks into which layers should be preserved and whether the answer depends on the downstream task.

Layer Specialization in Multilingual Models Wendler et al. (2024); Dumas et al. (2025) suggest that intermediate representations of a model are biased toward English, with language-specific processing concentrated in early and middle layers (Tang et al., 2024; Kojima et al., 2024; Zhao et al., 2024). Another approach that combines language and mathematical-reasoning experts through layer swapping (Bandarkar et al., 2025). Bandarkar & Peng (2025) also find that fine-tuning all parameters and then reverting selected updates can outperform freezing them during fine-tuning. These studies focus mostly on transferring reasoning through fine-tuning, with limited exploration of CPT. We extend this line of work to multilingual CPT and find that this phenomenon is task-dependent, as the middle-layer region whose reversion best recovers reading comprehension harms translation, and post-hoc reversion benefits do not reliably carry over to other preservation methods during training.

Expert Language Modeling Grouping languages by family can reduce cross-lingual interference and improve transfer to related low-resource languages (Chronopoulou et al., 2023; Ogueji et al., 2021; Downey et al., 2024). X-ELM (Blevins et al., 2024) extends Branch-Train-Merge (Li et al., 2022b) to multilingual pretraining by training experts on typologically clustered languages subsets that outperform compute-matched dense model. We adopt family experts as one of our CPT settings and find that specialization is task-dependent which is experts forget less than dense CPT on comprehension but under-perform on translation.

6 Conclusion

We investigate how layer-level parameter updates shape the acquisition–forgetting knowledge tradeoff during multilingual CPT. Specifically, we perform a layer-group interpolation analysis to uncover which layers most affect task performance change after jointly finetuning a model on five language families, resulting in decreased performance for reading comprehension but much stronger translation performance. Our analysis shows that the layers causing these performance changes are task-dependent: reverting the middle layers yields the largest gains in comprehension performance, while translation generally exhibits *worse* performance when reverting to the base model, though exact trends vary across language families and translation direction. Following these findings, targeted freezing and regularization interventions during CPT at these layers improve comprehension performance when training language family expert models over that base, but these gains do not consistently transfer to translation (where jointly fine-tuning with layer reversion performs the best).

These findings caution us against generalizing the insights from previous work with respect to layer-wise effects on reasoning tasks to other tasks. Instead, we observe that a layer range whose reversion benefits one task (comprehension) can impair another (translation), with the effect of a given layer range likely depending on underlying task features such as whether it is discriminative or generative, monolingual or cross-lingual, and the type of knowledge required. Thus, we propose our interpolation-based localization approach as a diagnostic framework to determine which layers (if any) should be preserved during task- and language-specific CPT, rather than positioning our results as a source of universal, task-agnostic guidelines.

AI USE STATEMENT

In this work, we used generative AI for assistance with writing code. We used AI-assisted literature review tools to supplement our manual literature reviewing process. We also used coding agents for assistance in running our experiments, with human in loop. AI was used for typesetting and generating LaTeX tables for our paper and minor writing refinements, with manual verification. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Lucas Bandarkar and Nanyun Peng. The unreasonable effectiveness of model merging for cross-lingual transfer in LLMs. In David Ifeoluwa Adelani, Catherine Arnett, Duygu Ataman, Tyler A. Chang, Hila Gonen, Rahul Raja, Fabian Schmidt, David Stap, and Jiayi Wang (eds.), *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pp. 131–148, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-345-6. doi: 10.18653/v1/2025.mrl-main.10. URL https://aclanthology.org/2025.mrl-main.10/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. Layer swapping for zero-shot cross-lingual transfer in large language models, 2025. URL https://arxiv.org/abs/2410.01335.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. Breaking the curse of multilinguality with cross-lingual expert language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10822–10837, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.604. URL https://aclanthology.org/2024.emnlp-main.604/.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. Language-family adapters for low-resource multilingual neural machine translation. In Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jade Abbott, Jonathan Washington, Nathaniel Oco, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao (eds.), *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pp. 59–72, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.loresmt-1.5. URL https://aclanthology.org/2023.loresmt-1.5/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. Sailor: Open language models for south-east asia, 2024. URL https://arxiv.org/abs/2404.03608.

C. M. Downey, Terra Blevins, Dhwani Serai, Dwija Parikh, and Shane Steinert-Threlkeld. Targeted multilingual adaptation for low-resource language families. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15647–15663, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.918. URL https://aclanthology.org/2024.findings-emnlp.918/.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31822–31841, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1536. URL https://aclanthology.org/2025.acl-long.1536/.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Maria Khelli, Samuel Cahyawijaya, Ayu Purwarianti, and Genta Indra Winata. What causes knowledge loss in multilingual language models? In Éric Le Ferrand, Elena Klyachko, Anna Postnikova, Tatiana Shavrina, Oleg Serikov, Ekaterina Voloshina, and Ekaterina Vylomova (eds.), *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pp. 15–25, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-282-4. URL https://aclanthology.org/2025.fieldmatters-1.2/.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons, 2024. URL https://arxiv.org/abs/2404.02431.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296, 2023.

Saurabh Kumar, Henrik Marklund, and Benjamin Van Roy. Maintaining plasticity in continual learning via regenerative regularization, 2024. URL https://arxiv.org/abs/2308.11958.

Dingcheng Li, Zheng Chen, Eunah Cho, Jie Hao, Xiaohu Liu, Fan Xing, Chenlei Guo, and Yang Liu. Overcoming catastrophic forgetting during domain adaptation of seq2seq language generation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5441–5454, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.398. URL https://aclanthology.org/2022.naacl-main.398/.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. Branch-train-merge: Embarrassingly parallel training of expert language models, 2022b. URL https://arxiv.org/abs/2208.03306.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2830–2839. PMLR, 2018.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pp. 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin (eds.), *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL https://aclanthology.org/2021.mrl-1.11/.

Abraham Toluwase Owodunni and Sachin Kumar. Continually adding new languages to multilingual language models, 2025. URL https://arxiv.org/abs/2509.11414.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models, 2024. URL https://arxiv.org/abs/2402.16438.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers, 2022. URL https://arxiv.org/abs/2109.10686.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 15296–15319. Curran Associates, Inc., 2024. doi: 10.52202/079017-0489. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1bd359b32ab8b2a6bbafa1ed2856cf40-Paper-Conference.pdf.

Yiran Zhao, Chaoqun Liu, Yue Deng, Jiahao Ying, Mahani Aljunied, Zhaodonghui Li, Lidong Bing, Hou Pong Chan, Yu Rong, Deli Zhao, and Wenxuan Zhang. Babel: Open multilingual large language models serving over 90% of global speakers, 2025. URL https://arxiv.org/abs/2503.00865.

A Appendix

A.1 Limitations

All experiments use a single 4B-parameter model (Gemma-3 4B) with a fixed budget of 5B tokens per family from one web corpus (MADLAD-400); we do not evaluate whether the localization results or intervention effects transfer to other model scales, architectures, or data regimes. The preservation techniques we evaluate (freezing, L2-SP, post-hoc reversion, and souping) are established methods; our contribution is using interpolation-based localization to select where to apply them, and testing whether the selected layers transfer across tasks and from post-hoc to training-time interventions. The localization analysis is performed only on the Dense CPT checkpoint, and the FLORES window is selected on a 20-direction screen of that checkpoint before being evaluated on the full held-in FLORES sets.

A.2 Held-Out Evaluation Languages

Table 6 lists all held-out languages evaluated in §4.5, grouped by family, together with their benchmark coverage. These languages are excluded from the CPT training set but are related to the training languages of the corresponding family group, allowing us to probe within-family generalization. The Slavic group also includes Lithuanian and Latvian, which are Baltic languages closely related to, but not members of, the Slavic family.

A.3 Training Hyperparameters and Fairness Controls

Tables 7 and 8 summarize the optimization settings used for the Gemma-3 4B experiments. Post-hoc reversion (D.-Rev. and E.-Rev.) and Expert Soup reuse the trained Dense and vanilla Expert checkpoints and introduce no additional optimization steps.

Targeted window constraints. For a selected six-layer window G ($[10, 16)$ or $[5, 11)$), Layer Freezing holds all parameters in G fixed at their Base values and trains every other layer. Layer-Range L2-SP trains all parameters and applies the penalty $\lambda = 0.05$ only to parameters in G, with $\lambda = 0$ elsewhere. We keep the learning rate unchanged across all family-specific variants.

Layer-Range L2-SP λ selection. We use a single penalty strength, $\lambda = 0.05$, for both windows and all families, without per-window or per-family tuning. Other values may change the balance between preservation and acquisition, particularly for translation.

A.4 Granular Sliding Window Results

Tables 9 and 10 provide the corresponding family-wise breakdown for the fixed-width window reversions in Table 2.

A.5 Held-Out Family-Level Results

Table 11 reports targeted-window perplexity on held-out relatives.

Table 12 gives the corresponding held-out Belebele family averages for the targeted $[10, 16)$ interventions.

Family	Code	Language	PPL	Belebele	FLORES
Slavic	bg	Bulgarian	✓	✓	✓
	cs	Czech	✓	✓	✓
	lt	Lithuanian	✓	✓	✓
	pl	Polish	✓	✓	✓
	sl	Slovenian	✓	✓	✓
	lv	Latvian	✓	✓	✓
Germanic	is	Icelandic	✓	✓	✓
	no	Norwegian	✓	✓	✓
	sv	Swedish	✓	✓	✓
Indic	as	Assamese	✓	✓	✓
	gu	Gujarati	✓	✓	✓
	or	Odia	✓	✓	✓
	pa	Punjabi	✓	✓	✓
	sd	Sindhi	✓	✓	✓
	si	Sinhala	✓	✓	✓
	ur	Urdu	✓	✓	✓
Austronesian	ilo	Ilocano	✓	✓	✓
	mi	Māori	✓	✓	✓
	su	Sundanese	✓	✓	✓
	war	Waray	✓	✓	✓
	mg	Malagasy	✓	✓	✓
Romance	ca	Catalan	✓	✓	✓

Table 6: Held-out evaluation languages per benchmark. ✓ = evaluated. PPL = held-out perplexity; FLORES results cover both EN→XX and XX→EN directions.

Hyperparameter	Dense CPT
Training data	All five families
Corpus budget	25B tokens total
Maximum steps	50,000
Learning rate	5×10^{-5}
Warmup / scheduler	500 steps / cosine
Batching	Per-device batch size 4, gradient accumulation 12, sequence length 2048, packed sequences
Validation	Every 1000 steps
Early stopping	Patience 3, threshold 0.001
Optimizer defaults	Weight decay 0.01, max gradient norm 1.0, seed 42, bfloat16, gradient checkpointing
Layer treatment	All layers trainable

Table 7: Training configuration for Dense CPT on Gemma-3 4B.

Hyperparameter	Family Expert	Layer Freezing	Layer-Range L2-SP
Training data	One family	One family	One family
Corpus budget	5B tokens/family	5B tokens/family	5B tokens/family
Maximum steps	16,954	18,000 or 21,000 (fixed budget of ≈0.35 epoch)†	
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}
Warmup / scheduler	200 steps / cosine	200 steps / cosine	200 steps / cosine
Batching	Per-device batch size 4, gradient accumulation 12, sequence length 2048, packed sequences		
Validation	Every 500 steps	Every 500 steps	Every 500 steps
Early stopping	Patience 6, threshold 0.005	Patience 6, threshold 0.005	
Optimizer defaults	Weight decay 0.01, max gradient norm 1.0, seed 42, bfloat16, gradient checkpointing		
Layer treatment	All layers trainable	Freeze the six layers in G; train all other layers	L2-SP on layers in G only; λ=0.05 in G, λ=0 elsewhere

Table 8: Training configuration for family-specific expert variants on Gemma-3 4B. G is the targeted window, $[10, 16)$ or $[5, 11)$. †21,000 steps for Romance (both windows) and for Indic and Austronesian $[10, 16)$; 18,000 steps otherwise.

Layers reverted	Slavic	Germanic	Indic	Austronesian	Romance
$[0, 6)$	+1.5	+3.0	+1.6	+1.5	+3.1
$[5, 11)$	+3.4	+4.1	+3.1	+5.0	+4.7
$[10, 16)$	+5.6	+6.5	+4.9	+5.5	+7.1
$[24, 30)$	−0.7	−0.5	−0.5	+0.4	−0.5
$[27, 33)$	+0.4	+0.2	+0.1	+0.9	+0.5
$[28, 34)$	+0.3	+0.8	+0.5	+1.0	+1.0

Table 9: Family-wise Belebele accuracy changes (percentage points) relative to Dense CPT for the selected fixed-width layer reversions. Each family score averages its held-in evaluation languages.

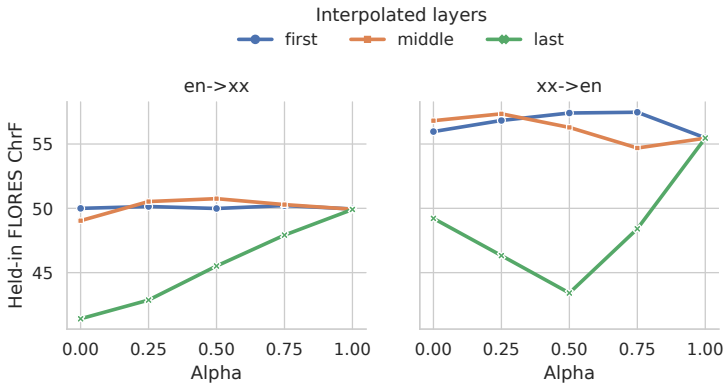

Figure 5: Held-in FLORES ChrF under layer interpolation between the base model and Dense CPT, reported separately for en→xx and xx→en directions. All non-interpolated layers are kept at their Dense CPT values.

Layers reverted	Family	xx→en Δ ChrF	en→xx Δ ChrF
$[0, 6)$	Slavic	$+3.14$	$+2.13$
	Germanic	-1.55	$+0.75$
	Indic	-0.19	-0.65
	Austronesian	-4.14	-0.48
	Romance	$+3.38$	$+0.18$
$[5, 11)$	Slavic	$+3.08$	$+1.92$
	Germanic	$+1.61$	$+0.68$
	Indic	-5.24	-0.75
	Austronesian	$+1.01$	$+0.32$
	Romance	$+3.86$	$+1.01$
$[10, 16)$	Slavic	-11.03	-0.51
	Germanic	-7.53	-0.10
	Indic	-2.89	-1.27
	Austronesian	-3.97	-0.04
	Romance	-6.62	-0.76
$[24, 30)$	Slavic	-8.50	-0.23
	Germanic	-4.89	-0.76
	Indic	-0.03	-1.38
	Austronesian	-3.41	-17.03
	Romance	-5.82	-0.70
$[27, 33)$	Slavic	-24.63	-10.35
	Germanic	-29.91	-9.42
	Indic	-12.19	-12.15
	Austronesian	-23.52	-28.82
	Romance	-27.38	-6.68
$[28, 34)$	Slavic	-11.98	-4.95
	Germanic	-8.11	-7.01
	Indic	$+1.79$	-7.44
	Austronesian	-7.95	-20.63
	Romance	-11.14	-2.57

Table 10: Family- and direction-wise FLORES ChrF changes relative to Dense CPT for the selected six-layer Dense-to-Base reversions. Each cell averages two held-in languages from the 20-direction screen. The Dense reference is the mean of the three archived Dense evaluations, rescored with the first-line post-truncation protocol.

	Shared baselines				$[5, 11)$				$[10, 16)$			
Family	**Base**	**Dense**	**Expert**	**Soup**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**
Slavic	7.36	9.16	7.64	**7.10**	9.26	7.67	7.63	7.65	9.12	7.59	7.63	7.68
Germanic	8.56	11.44	9.23	**8.30**	11.26	9.23	9.17	9.18	11.33	9.09	9.18	9.15
Indic	8.99	14.70	9.76	**8.94**	14.65	9.79	9.78	9.79	14.26	9.65	9.86	9.76
Austronesian	**14.04**	28.36	28.35	14.17	29.93	30.00	29.99	30.25	28.75	25.08	27.69	27.58
Romance	7.90	9.07	7.93	**7.57**	9.31	8.01	7.93	7.89	9.21	7.94	7.97	7.97

Table 11: Held-out validation perplexity ↓, averaged over held-out relatives within each family. D.-Rev. restores Dense layers in the indicated window to Base; Freeze and L2-SP preserve the indicated layers during family CPT; E.-Rev. restores them to Base after Expert CPT. Germanic excludes German. **Bold** = lowest value within a row.

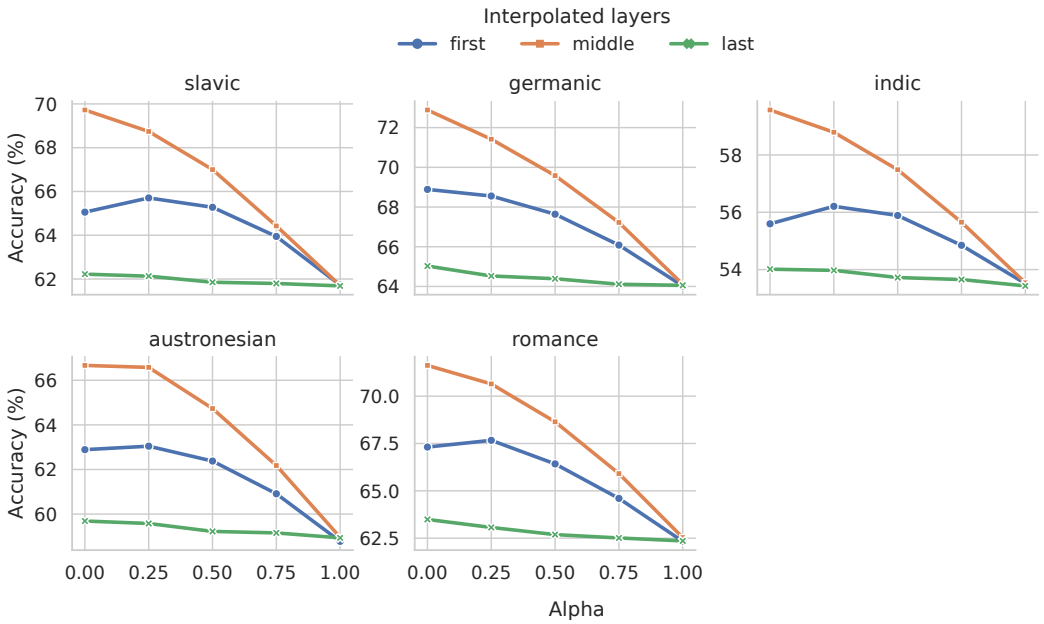

Figure 6: Held-in Belebele accuracy under first-, middle-, and last-layer interpolation, broken down by language family.

	Shared baselines				$[10, 16)$			
Family	**Base**	**Dense**	**Expert**	**Soup**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**
Slavic	**0.725**	0.598	0.710	0.719	0.657	0.721	0.708	0.716
Germanic	**0.722**	0.592	0.699	0.716	0.652	0.708	0.713	0.709
Indic	**0.576**	0.476	0.557	0.573	0.521	0.566	0.564	0.566
Austronesian	**0.487**	0.383	0.432	0.475	0.399	0.440	0.456	0.455
Romance	**0.752**	0.600	0.731	**0.752**	0.682	0.744	0.741	0.744
Average	**0.652**	0.530	0.626	0.647	0.582	0.636	0.636	0.638

Table 12: Held-out Belebele accuracy ↑ (2-shot), averaged over held-out relatives within each family; Average weights families equally. D.-Rev. restores Dense layers $[10, 16)$ to Base; Freeze and L2-SP preserve $[10, 16)$ during family CPT; E.-Rev. restores $[10, 16)$ to Base after Expert CPT. Germanic excludes German. **Bold** = best value within a row.

A.6 FLORES-200 Evaluation Protocol

We evaluate FLORES-200 with lm-eval-harness in a 2-shot setting and report corpus ChrF for both EN→XX and XX→EN directions. Our initial goal was to match the Gemma-3 technical report's FLORES setup as closely as possible for the shared base model (Team et al., 2025). However, exact replication was not possible with the default lm-eval-harness task configuration, since newline stopping could terminate some base-model generations before a translation was produced.

We therefore use a uniform post-truncation protocol for all checkpoints. After generation, we strip leading whitespace, keep only the text before the first generated newline or literal $\backslash\mathrm{n}$, and score the remaining span against the reference with ChrF. This keeps decoding and scoring comparable across the base model, Dense CPT, family experts, reverted checkpoints, Layer Freezing, Layer-Range L2-SP, and Expert Soup. The resulting FLORES numbers should therefore be read as a controlled comparison under a shared evaluation pipeline, rather than as a direct reproduction of the Gemma-3 technical report score.

A.7 Targeted FLORES Results and Comparison Details

Table 13 separates the $[5, 11)$ results in Table 5 by translation direction.

Empty outputs. Empty-output rates provide an important diagnostic alongside ChrF. All Austronesian family-expert variants have high rates: 57.5% for the vanilla Expert, 62.55% for E.-Rev., 58.97% for Freeze, and 61.37% for L2-SP, compared with at most 3.69% for any other family-specific checkpoint. Their translation scores therefore reflect substantial output failures as well as the quality of nonempty translations. The final block of Table 13 shows that reverting the Austronesian expert's $[27, 33)$ layers to Base reduces the empty-output rate to 7.0% (§4.4), from 57.5% for the archived Expert evaluation and 38.9% for the separate Expert generation run used as that block's reference; here β is the fraction of the way from Expert to Base weights, so $\beta = 1$ corresponds to $\alpha = 0$ in the notation of §3.1.

A.8 Per-Language Results Tables

Tables 14–17 report per-language results for perplexity, Belebele, and FLORES. For each language, the Expert, E.-Rev., Freeze, and L2-SP columns report the model trained on that language's family (e.g., the Slavic expert for Croatian and Russian). All other strategy columns (Dense, D.-Rev., Soup) are single models evaluated across all languages.

Group	Method	XX→EN	EN→XX	Empty (%)
Slavic	Dense	56.74	48.66	1.38
	D.-Rev.	59.80	49.64	0.02
	E.-Rev.	52.98	49.66	0.17
	Freeze	45.99	48.24	0.19
	L2-SP	45.34	48.38	0.33
Germanic	Dense	61.39	56.64	0.56
	D.-Rev.	64.73	57.80	0.00
	E.-Rev.	61.48	57.82	1.21
	Freeze	58.53	56.91	1.31
	L2-SP	58.91	57.23	1.82
Indic	Dense	47.68	39.16	0.15
	D.-Rev.	42.79	39.75	0.02
	E.-Rev.	34.28	38.48	3.16
	Freeze	34.61	38.09	2.16
	L2-SP	35.56	37.61	1.86
Austronesian	Dense	57.22	53.87	0.95
	D.-Rev.	59.58	52.90	0.08
	E.-Rev.	14.24	30.11	62.55
	Freeze	17.88	31.46	58.97
	L2-SP	17.50	29.28	61.37
Romance	Dense	58.61	57.12	0.68
	D.-Rev.	62.72	57.75	0.04
	E.-Rev.	55.05	57.71	3.69
	Freeze	47.48	51.42	3.47
	L2-SP	46.91	50.13	3.18
Austronesian: Expert-to-Base interpolation in $[27, 33)$				
	Expert (re-run)*	29.05	30.36	38.9
	$\beta = 0.5$	31.40	39.17	35.3
	$\beta = 1.0$	40.02	40.61	7.0

Table 13: Direction-wise held-in FLORES ChrF for the pipeline-matched Dense/D.-Rev. pair and the $[5, 11)$ family-specific interventions in Table 5. Dense is the pipeline-matched control ($\alpha = 1$) for D.-Rev. Scores use first-line post-truncation scoring and are macro-averaged within each group. Empty-output percentages aggregate both directions. The additional Austronesian block interpolates the vanilla expert's $[27, 33)$ layers toward Base, with $\beta = 1$ denoting full replacement (equivalent to $\alpha = 0$). These interventions are separate from the Freeze and L2-SP checkpoints above. *A separate generation run of the vanilla Austronesian expert, used as the paired reference for this block; the Expert evaluation in Table 5 scores 19.28 / 30.75 ChrF with a 57.5% empty-output rate.

Family	Language	Base	Dense	Expert	Soup	D.-Rev.$_{10}$	E.-Rev.$_{10}$	Freeze$_{10}$	L2-SP$_{10}$
Slavic	Macedonian (mk)	6.70	**4.96**	5.06	6.11	5.10	5.20	5.15	5.20
Slavic	Croatian (hr)	8.42	**6.90**	**6.90**	7.90	7.09	7.11	7.01	7.09
Slavic	Russian (ru)	10.37	9.41	**9.21**	9.91	9.58	9.41	9.29	9.38
Slavic	Slovak (sk)	7.25	6.14	**6.12**	6.72	6.30	6.29	6.19	6.26
Slavic	Serbian (sr)	7.03	**5.45**	5.58	6.57	5.62	5.73	5.69	5.73
Slavic	Ukrainian (uk)	6.20	5.59	**5.51**	5.89	5.71	5.65	5.55	5.61
Germanic	Afrikaans (af)	11.24	**7.41**	7.71	9.92	7.72	7.98	8.03	8.04
Germanic	Frisian (fy)	14.89	**4.86**	5.28	11.20	5.35	5.76	5.69	5.70
Germanic	Luxembourgish (lb)	14.32	**5.61**	6.08	11.77	6.10	6.50	6.48	6.49
Germanic	Danish (da)	7.42	6.14	**6.09**	6.92	6.36	6.30	6.22	6.22
Germanic	Dutch (nl)	8.69	7.53	**7.40**	8.18	7.77	7.66	7.55	7.55
Germanic	English (en)	11.91	11.51	**10.83**	11.42	11.63	10.99	10.94	10.94
Indic	Bengali (bn)	13.47	11.05	**10.99**	12.31	11.49	11.40	11.34	11.29
Indic	Hindi (hi)	10.17	8.47	**8.39**	9.52	8.81	8.70	8.62	8.59
Indic	Kannada (kn)	7.73	**5.35**	5.49	6.53	5.58	5.72	5.71	5.71
Indic	Malayalam (ml)	7.94	**5.69**	5.79	6.73	5.94	6.05	6.03	6.01
Indic	Marathi (mr)	10.83	**8.25**	8.28	9.44	8.62	8.65	8.57	8.55
Indic	Nepali (ne)	11.72	**8.11**	8.40	9.95	8.42	8.68	8.74	8.73
Indic	Tamil (ta)	11.49	**8.78**	8.88	9.91	9.23	9.30	9.24	9.21
Indic	Telugu (te)	7.16	**5.54**	5.58	6.20	5.76	5.81	5.77	5.76
Austronesian	Samoan (sm)	11.39	**4.00**	4.65	9.39	4.61	5.08	5.09	5.10
Austronesian	Javanese (jv)	17.58	**6.45**	7.69	14.36	7.64	8.49	8.62	8.64
Austronesian	Cebuano (ceb)	10.12	**4.63**	5.29	8.72	5.28	5.75	5.82	5.82
Austronesian	Filipino (fil)	10.10	**5.32**	5.61	8.54	6.18	6.41	5.94	5.93
Austronesian	Indonesian (id)	10.89	**8.74**	8.85	9.99	9.24	9.27	9.15	9.13
Austronesian	Malay (ms)	10.60	**8.20**	8.38	9.59	8.49	8.63	8.60	8.59
Romance	Spanish (es)	10.81	9.93	**9.63**	10.19	10.13	9.84	9.83	9.90
Romance	Portuguese (pt)	10.12	9.09	**8.81**	9.48	9.27	9.04	9.01	9.06
Romance	French (fr)	9.47	8.55	**8.33**	8.78	8.79	8.56	8.56	8.61
Romance	Galician (gl)	11.19	**8.21**	8.37	10.06	8.44	8.55	8.60	8.66
Romance	Italian (it)	10.77	8.94	**8.77**	10.03	9.54	9.31	9.09	9.14
Romance	Romanian (ro)	7.67	6.56	**6.50**	7.16	6.76	6.69	6.69	6.72

Table 14: Per-language perplexity ↓ on held-in validation text. D.-Rev.$_{10}$ and E.-Rev.$_{10}$ restore layers $[10, 16)$ of the Dense and Expert models to Base; Freeze$_{10}$ and L2-SP$_{10}$ use the same targeted layer window. Expert, E.-Rev.$_{10}$, and the targeted training columns report the model trained on the language's family. **Bold** = best per row; underline = within 0.2 of best.

Family	Language	Base	Dense	Expert	Soup	D.-Rev.$_{10}$	E.-Rev.$_{10}$	Freeze$_{10}$	L2-SP$_{10}$
Slavic	Macedonian (mk)	0.684	0.626	0.702	0.696	0.654	**0.719**	0.709	0.717
Slavic	Croatian (hr)	0.727	0.606	0.736	**0.751**	0.666	0.736	0.737	0.738
Slavic	Russian (ru)	0.730	0.606	0.720	**0.746**	0.664	0.737	0.722	0.727
Slavic	Slovak (sk)	0.744	0.638	0.737	**0.751**	0.681	0.744	**0.751**	0.749
Slavic	Serbian (sr)	0.708	0.628	0.736	0.739	0.699	**0.743**	0.729	0.731
Slavic	Ukrainian (uk)	0.722	0.610	0.726	0.724	0.687	**0.728**	0.714	0.722
Germanic	Afrikaans (af)	0.734	0.617	**0.747**	0.742	0.684	0.743	0.739	**0.747**
Germanic	Danish (da)	0.743	0.624	0.740	**0.756**	0.699	0.748	0.750	**0.756**
Germanic	Dutch (nl)	0.742	0.653	0.742	**0.748**	0.692	0.741	0.739	0.744
Germanic	English (en)	0.813	0.674	0.793	0.798	0.754	0.803	**0.820**	0.807
Indic	Bengali (bn)	0.654	0.550	0.616	0.653	0.603	0.658	0.657	**0.659**
Indic	Hindi (hi)	0.514	0.539	0.567	**0.616**	0.579	0.612	0.603	0.608
Indic	Kannada (kn)	0.601	0.518	0.569	0.602	0.563	0.592	0.600	**0.610**
Indic	Malayalam (ml)	0.632	0.532	0.616	**0.664**	0.603	0.654	0.659	0.651
Indic	Marathi (mr)	0.649	0.539	0.617	0.654	0.596	0.644	0.650	**0.661**
Indic	Nepali (ne)	0.574	0.547	0.564	**0.624**	0.584	0.621	0.602	0.616
Indic	Tamil (ta)	0.617	0.543	0.613	**0.628**	0.581	0.623	0.623	0.614
Indic	Telugu (te)	0.568	0.509	0.584	0.592	0.564	0.584	0.587	**0.597**
Austronesian	Javanese (jv)	0.542	0.522	0.569	0.594	0.578	0.606	**0.637**	0.631
Austronesian	Cebuano (ceb)	0.658	0.570	0.633	0.673	0.614	0.680	0.688	**0.691**
Austronesian	Filipino (fil)	0.698	0.617	0.689	0.713	0.661	0.714	**0.723**	0.719
Austronesian	Indonesian (id)	0.706	0.613	0.702	0.710	0.680	0.710	**0.717**	0.693
Austronesian	Malay (ms)	**0.738**	0.613	0.710	0.721	0.678	0.730	0.724	0.726
Romance	Spanish (es)	0.732	0.638	0.737	0.729	0.709	**0.738**	0.727	0.732
Romance	Portuguese (pt)	0.759	0.647	0.748	**0.766**	0.697	0.759	0.760	0.756
Romance	French (fr)	0.779	0.644	0.774	0.779	0.732	0.782	0.780	**0.787**
Romance	Italian (it)	0.724	0.600	0.732	0.736	0.679	**0.748**	0.728	0.739
Romance	Romanian (ro)	**0.746**	0.597	0.739	0.738	0.661	0.738	0.744	0.738

Table 15: Per-language Belebele accuracy ↑ (2-shot). D.-Rev.$_{10}$, E.-Rev.$_{10}$, Freeze$_{10}$, and L2-SP$_{10}$ use the targeted layer window $[10, 16)$; Expert, E.-Rev.$_{10}$, and the targeted columns report the model trained on the language's family. **Bold** marks the best value at the displayed precision, including ties.

Family	Language	Base	Dense	Expert	Soup	D.-Rev.$_5$	E.-Rev.$_5$	Freeze$_5$	L2-SP$_5$
Slavic	Croatian (hr)	23.1	55.6	45.0	44.1	**58.3**	54.6	45.0	45.3
Slavic	Macedonian (mk)	27.8	57.8	49.5	40.8	**62.5**	58.3	51.2	52.0
Slavic	Russian (ru)	22.2	56.7	38.2	39.4	**57.9**	39.5	35.2	33.1
Slavic	Slovak (sk)	25.4	55.9	47.9	49.3	**60.3**	57.1	49.1	47.2
Slavic	Serbian (sr)	29.0	55.6	51.2	45.4	**59.9**	57.0	51.8	50.0
Slavic	Ukrainian (uk)	28.5	58.8	47.0	46.1	**59.8**	51.3	43.8	44.4
Germanic	Afrikaans (af)	20.2	69.8	65.4	46.8	**71.0**	61.3	62.0	64.3
Germanic	Danish (da)	21.9	64.6	59.1	49.4	**66.6**	64.9	61.9	62.1
Germanic	Luxembourgish (lb)	36.0	56.7	56.5	52.7	**65.2**	63.8	62.8	61.8
Germanic	Dutch (nl)	25.1	54.5	50.5	39.8	**56.1**	55.9	47.3	47.4
Indic	Bengali (bn)	21.2	**50.7**	30.2	44.2	45.0	30.8	32.9	32.6
Indic	Hindi (hi)	21.3	**51.5**	32.1	48.0	46.1	33.9	35.0	35.7
Indic	Kannada (kn)	34.2	45.2	36.0	**49.6**	40.9	34.1	34.1	34.8
Indic	Malayalam (ml)	31.7	43.8	36.8	**44.7**	39.9	35.4	34.6	36.5
Indic	Marathi (mr)	30.1	**48.9**	35.1	47.7	43.5	34.2	34.2	35.7
Indic	Nepali (ne)	24.8	**54.2**	38.5	48.9	50.2	40.8	41.1	41.5
Indic	Tamil (ta)	27.9	45.4	34.9	**46.5**	40.5	33.2	32.3	33.3
Indic	Telugu (te)	33.0	41.7	35.4	**45.3**	36.2	31.9	32.7	34.4
Austronesian	Cebuano (ceb)	40.0	57.3	12.0	56.2	**62.9**	9.7	10.9	10.1
Austronesian	Filipino (fil)	29.9	61.1	17.0	51.1	**63.7**	10.6	13.8	14.2
Austronesian	Indonesian (id)	19.1	59.8	20.5	50.5	**62.2**	13.3	17.7	17.1
Austronesian	Javanese (jv)	29.4	55.0	16.4	47.8	**57.6**	12.5	18.3	17.3
Austronesian	Malay (ms)	20.1	60.6	21.5	50.9	**62.2**	14.0	16.9	16.4
Austronesian	Samoan (sm)	35.5	**49.5**	28.2	39.4	48.9	25.3	29.7	30.0
Romance	Spanish (es)	28.2	55.2	42.8	36.5	**57.6**	40.4	38.3	36.6
Romance	French (fr)	25.4	63.6	56.1	49.9	**65.2**	57.6	51.2	51.2
Romance	Galician (gl)	25.4	58.0	54.1	46.6	**63.6**	57.5	53.2	53.1
Romance	Italian (it)	27.2	55.1	48.8	36.8	**59.3**	51.6	38.8	37.7
Romance	Portuguese (pt)	24.2	61.3	57.9	48.1	**66.4**	61.2	49.1	49.0
Romance	Romanian (ro)	30.8	58.5	60.5	51.5	**64.2**	62.0	54.3	53.7

Table 16: Per-language FLORES-200 ChrF ↑ (xx→en, 2-shot). D.-Rev.$_5$, E.-Rev.$_5$, Freeze$_5$, and L2-SP$_5$ use the targeted layer window $[5, 11)$; Dense is the pipeline-matched control for D.-Rev.$_5$. Expert, E.-Rev.$_5$, and the targeted training columns report the model trained on the language's family. **Bold** = best per row; underline = within 1 ChrF point of best.

Family	Language	Base	Dense	Expert	Soup	D.-Rev.$_5$	E.-Rev.$_5$	Freeze$_5$	L2-SP$_5$
Slavic	Croatian (hr)	36.9	**53.1**	50.7	48.9	52.5	52.7	51.7	51.6
Slavic	Macedonian (mk)	40.3	47.0	51.4	50.6	50.1	**52.0**	50.1	49.2
Slavic	Russian (ru)	43.3	51.0	46.1	47.7	**51.4**	49.8	47.0	47.1
Slavic	Slovak (sk)	45.5	43.0	45.5	43.9	45.2	**45.6**	44.1	45.0
Slavic	Serbian (sr)	35.2	49.0	**49.2**	49.0	48.8	48.8	48.3	48.9
Slavic	Ukrainian (uk)	45.3	48.9	47.6	49.2	**49.8**	49.0	48.3	48.5
Germanic	Afrikaans (af)	40.9	**65.1**	63.9	61.4	64.8	64.8	63.3	64.5
Germanic	Danish (da)	55.5	61.6	62.1	60.5	62.7	**63.9**	63.1	63.1
Germanic	Luxembourgish (lb)	39.2	47.5	45.9	42.1	**49.8**	48.1	47.1	47.2
Germanic	Dutch (nl)	40.0	52.4	52.3	50.3	53.9	**54.5**	54.1	54.1
Indic	Bengali (bn)	34.1	**42.0**	38.1	39.5	41.0	38.7	36.7	37.0
Indic	Hindi (hi)	27.9	**43.6**	40.6	43.3	42.2	40.1	39.9	40.8
Indic	Kannada (kn)	35.8	37.5	38.4	38.0	40.4	39.0	**40.8**	39.6
Indic	Malayalam (ml)	33.0	33.7	35.3	34.0	36.8	36.8	**37.4**	36.2
Indic	Marathi (mr)	26.8	**38.5**	34.0	36.6	37.6	35.8	34.6	34.1
Indic	Nepali (ne)	29.8	45.6	42.6	40.1	**46.6**	44.1	42.8	42.7
Indic	Tamil (ta)	33.5	**41.1**	37.4	39.5	40.1	39.4	38.8	37.9
Indic	Telugu (te)	34.1	31.4	32.9	**35.5**	33.3	34.0	33.6	32.6
Austronesian	Cebuano (ceb)	38.9	54.3	25.3	46.7	**56.4**	26.9	30.6	24.5
Austronesian	Filipino (fil)	44.2	57.0	25.6	53.6	**57.1**	23.3	22.5	20.2
Austronesian	Indonesian (id)	41.4	**63.5**	28.4	**63.5**	**63.5**	28.2	35.8	35.3
Austronesian	Javanese (jv)	37.3	**39.9**	20.9	36.9	29.8	19.3	24.5	23.0
Austronesian	Malay (ms)	53.0	61.6	43.9	60.9	**62.7**	43.6	36.6	34.5
Austronesian	Samoan (sm)	16.6	47.1	40.4	19.3	**47.9**	39.4	38.9	38.3
Romance	Spanish (es)	41.3	**53.3**	51.2	48.8	52.8	51.6	45.5	44.3
Romance	French (fr)	36.7	**65.2**	63.2	61.6	**65.2**	64.7	58.4	56.2
Romance	Galician (gl)	37.9	56.2	54.7	48.1	**57.0**	56.5	48.5	49.1
Romance	Italian (it)	44.7	53.7	52.8	48.4	**54.3**	**54.3**	46.6	44.9
Romance	Portuguese (pt)	40.2	65.1	63.3	61.0	**65.2**	65.0	58.1	56.1
Romance	Romanian (ro)	51.2	49.2	53.2	53.5	52.1	**54.2**	51.5	50.3

Table 17: Per-language FLORES-200 ChrF ↑ (en→xx, 2-shot). D.-Rev.$_5$, E.-Rev.$_5$, Freeze$_5$, and L2-SP$_5$ use the targeted layer window $[5, 11)$; Dense is the pipeline-matched control for D.-Rev.$_5$. Expert, E.-Rev.$_5$, and the targeted training columns report the model trained on the language's family. **Bold** = best per row; underline = within 1 ChrF point of best.

A.9 Per-Language Held-Out Results Tables

Tables 18–21 report results for the same 22 languages excluded from CPT, complementing the family-averaged perplexity and Belebele results in Tables 11 and 12. Perplexity covers both targeted windows; Belebele uses $[10, 16)$ and FLORES uses $[5, 11)$. For each language, Expert, Freeze, L2-SP, and E.-Rev. report the model trained on its family; Base, Dense, D.-Rev., and Soup are shared models. Reverted models restore the indicated layers to Base weights. Family summaries average languages equally (and both directions for FLORES); overall summaries weight the five families equally.

		Shared baselines				$[5, 11)$				$[10, 16)$			
Family	**Language**	**Base**	**Dense**	**Expert**	**Soup**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**
Slavic	Bulgarian (bg)	6.46	8.13	7.21	**6.23**	8.87	7.51	7.12	7.14	8.07	7.09	7.14	7.15
Slavic	Czech (cs)	7.90	8.48	7.75	**7.56**	8.73	7.88	7.75	7.79	8.52	7.78	7.75	7.82
Slavic	Lithuanian (lt)	7.38	9.57	7.56	**7.17**	9.52	7.51	7.58	7.60	9.56	7.53	7.60	7.63
Slavic	Latvian (lv)	6.43	8.45	6.67	**6.29**	8.40	6.62	6.69	6.70	8.41	6.64	6.70	6.73
Slavic	Polish (pl)	7.87	9.20	7.82	**7.51**	9.21	7.80	7.88	7.91	9.24	7.86	7.89	7.95
Slavic	Slovenian (sl)	8.09	11.13	8.85	**7.82**	10.80	8.67	8.74	8.79	10.92	8.65	8.69	8.79
Germanic	Icelandic (is)	**8.05**	14.68	10.30	8.11	14.18	10.20	10.04	10.00	14.18	9.80	10.07	10.01
Germanic	Norwegian (no)	8.43	9.41	8.35	**8.03**	9.37	8.39	8.38	8.41	9.52	8.38	8.38	8.37
Germanic	Swedish (sv)	9.20	10.24	9.04	**8.74**	10.24	9.09	9.09	9.12	10.29	9.08	9.08	9.07
Indic	Assamese (as)	**10.28**	14.91	11.25	10.43	13.83	10.93	11.17	11.14	13.59	10.70	11.20	11.10
Indic	Gujarati (gu)	7.30	10.97	7.37	**7.04**	11.08	7.47	7.61	7.60	11.24	7.48	7.62	7.50
Indic	Odia (or)	4.62	5.45	4.86	**4.56**	5.43	4.86	4.81	4.86	5.43	4.79	4.88	4.85
Indic	Punjabi (pa)	5.12	6.32	5.31	**5.01**	6.29	5.31	5.40	5.40	6.29	5.28	5.41	5.36
Indic	Sindhi (sd)	**12.87**	26.61	14.94	13.20	26.66	15.04	14.50	14.58	24.11	14.68	14.70	14.58
Indic	Sinhala (si)	**7.47**	16.03	9.01	**7.47**	16.21	9.04	9.15	9.12	15.81	8.87	9.48	9.23
Indic	Urdu (ur)	15.29	22.61	15.55	**14.87**	23.07	15.88	15.79	15.83	23.36	15.79	15.76	15.68
Austronesian	Ilocano (ilo)	25.41	40.41	35.42	**25.13**	41.45	34.34	36.47	36.18	40.38	32.44	33.99	34.45
Austronesian	Malagasy (mg)	**9.97**	26.89	28.91	10.29	29.83	30.06	30.31	30.49	24.63	20.68	27.02	26.70
Austronesian	Māori (mi)	**14.10**	36.44	38.41	14.45	39.50	44.93	40.36	41.04	39.44	36.63	38.46	38.56
Austronesian	Sundanese (su)	**17.15**	34.05	35.29	17.57	34.87	37.00	39.18	39.94	35.33	32.01	35.32	34.54
Austronesian	Waray (war)	3.57	4.01	3.72	**3.39**	4.01	3.66	3.61	3.61	3.96	3.64	3.63	3.63
Romance	Catalan (ca)	7.90	9.07	7.93	**7.57**	9.31	8.01	7.93	7.89	9.21	7.94	7.97	7.97

Table 18: Per-language held-out validation perplexity ↓ for both targeted windows. D.-Rev. and E.-Rev. restore the indicated layers to Base after Dense and Expert CPT, respectively; Freeze and L2-SP preserve those layers during training. Expert, Freeze, L2-SP, and E.-Rev. use the model trained on the language's family; Base, Dense, D.-Rev., and Soup are shared models. **Bold** marks the best value at the displayed precision, including ties.

		Shared baselines				$[10, 16)$			
Family	**Language**	**Base**	**Dense**	**Expert**	**Soup**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**
Slavic	Bulgarian (bg)	0.707	0.629	0.714	0.718	0.676	**0.730**	0.724	0.726
Slavic	Czech (cs)	**0.756**	0.628	0.734	0.748	0.683	0.746	0.730	0.736
Slavic	Lithuanian (lt)	**0.724**	0.614	0.721	0.716	0.668	0.723	0.718	0.710
Slavic	Latvian (lv)	**0.732**	0.576	0.699	0.707	0.636	0.714	0.696	0.712
Slavic	Polish (pl)	0.707	0.570	0.692	0.709	0.644	0.707	0.692	**0.712**
Slavic	Slovenian (sl)	**0.723**	0.569	0.699	0.716	0.637	0.708	0.688	0.702
Germanic	Icelandic (is)	**0.664**	0.524	0.636	0.656	0.554	0.649	0.639	0.644
Germanic	Norwegian (no)	**0.743**	0.613	0.714	0.737	0.697	0.731	**0.743**	0.732
Germanic	Swedish (sv)	**0.759**	0.638	0.748	0.756	0.706	0.743	0.756	0.751
Indic	Assamese (as)	**0.560**	0.460	0.539	0.548	0.516	0.549	0.552	0.550
Indic	Gujarati (gu)	**0.593**	0.500	0.578	0.571	0.551	0.588	0.584	0.590
Indic	Odia (or)	**0.542**	0.430	0.519	0.536	0.461	0.514	0.522	0.520
Indic	Punjabi (pa)	0.618	0.500	0.598	**0.626**	0.559	0.621	0.608	0.609
Indic	Sindhi (sd)	0.478	0.421	0.473	**0.501**	0.434	0.457	0.462	0.462
Indic	Sinhala (si)	**0.607**	0.470	0.586	0.604	0.548	0.606	0.602	0.604
Indic	Urdu (ur)	**0.637**	0.551	0.609	0.624	0.579	0.630	0.619	0.627
Austronesian	Ilocano (ilo)	**0.410**	0.349	0.384	0.398	0.348	0.382	0.407	0.387
Austronesian	Malagasy (mg)	**0.529**	0.391	0.481	0.527	0.409	0.494	0.506	0.499
Austronesian	Māori (mi)	**0.367**	0.317	0.304	0.351	0.321	0.308	0.319	0.323
Austronesian	Sundanese (su)	**0.578**	0.427	0.527	0.570	0.451	0.524	0.558	0.556
Austronesian	Waray (war)	**0.550**	0.431	0.462	0.530	0.464	0.493	0.493	0.510
Romance	Catalan (ca)	**0.752**	0.600	0.731	**0.752**	0.682	0.744	0.741	0.744

Table 19: Per-language held-out Belebele accuracy ↑ (2-shot), with all targeted interventions at $[10, 16)$. Expert, Freeze, L2-SP, and E.-Rev. use the model trained on the language's family; Base, Dense, D.-Rev., and Soup are shared models. **Bold** marks the best value at the displayed precision, including ties.

		Shared baselines				$[5, 11)$			
Family	**Language**	**Base**	**Dense**	**Expert**	**Soup**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**
Slavic	Bulgarian (bg)	25.6	56.9	48.4	43.8	**62.0**	56.8	47.1	48.6
Slavic	Czech (cs)	25.3	57.7	49.5	49.5	**59.8**	57.6	49.1	49.3
Slavic	Lithuanian (lt)	28.9	50.6	44.7	45.1	52.0	**52.5**	45.1	48.0
Slavic	Latvian (lv)	37.6	**57.5**	50.2	52.0	56.6	56.6	51.7	51.9
Slavic	Polish (pl)	24.8	51.7	38.6	40.8	**54.4**	46.0	36.1	35.8
Slavic	Slovenian (sl)	25.9	55.4	43.6	44.5	**56.5**	54.1	46.6	47.6
Germanic	Icelandic (is)	42.4	50.6	51.7	49.0	52.2	**54.9**	53.4	53.5
Germanic	Norwegian (no)	24.0	59.8	46.7	40.6	**61.8**	60.4	58.0	56.6
Germanic	Swedish (sv)	21.9	61.2	57.0	48.5	63.0	**63.1**	59.1	60.2
Indic	Assamese (as)	33.5	40.1	31.0	**41.6**	41.2	31.2	32.5	33.8
Indic	Gujarati (gu)	28.8	**49.9**	35.2	47.4	49.3	35.5	37.7	39.6
Indic	Odia (or)	44.4	40.9	37.0	**44.9**	38.3	37.8	35.4	34.3
Indic	Punjabi (pa)	38.4	**50.0**	38.8	47.4	49.1	39.6	42.0	44.0
Indic	Sindhi (sd)	34.8	45.6	34.6	45.3	**46.8**	37.7	39.2	39.1
Indic	Sinhala (si)	38.8	39.9	36.0	**44.6**	37.2	35.8	34.0	36.2
Indic	Urdu (ur)	28.6	**50.7**	32.4	45.2	49.9	34.7	33.8	36.1
Austronesian	Ilocano (ilo)	40.1	36.9	28.8	38.6	**41.1**	29.6	30.2	29.0
Austronesian	Malagasy (mg)	35.0	31.1	24.8	**38.2**	34.5	26.3	29.8	29.3
Austronesian	Māori (mi)	28.0	24.9	21.9	**30.9**	28.8	23.2	23.1	22.3
Austronesian	Sundanese (su)	31.2	48.9	20.8	48.1	**51.7**	18.7	23.1	21.1
Austronesian	Waray (war)	41.1	53.7	19.3	53.9	**59.3**	15.3	20.6	19.8
Romance	Catalan (ca)	31.0	60.1	50.5	36.4	**64.8**	51.2	44.5	44.5

Table 20: Per-language held-out FLORES-200 ChrF (xx→en) ↑, with all targeted interventions at $[5, 11)$ and the postprocessing in Appendix A.6. Expert, Freeze, L2-SP, and E.-Rev. use the model trained on the language's family; Base, Dense, D.-Rev., and Soup are shared models. **Bold** marks the best value at the displayed precision, including ties.

		Shared baselines				$[5, 11)$			
Family	**Language**	**Base**	**Dense**	**Expert**	**Soup**	**D.-Rev.**	**E.-Rev.**	**Freeze**	**L2-SP**
Slavic	Bulgarian (bg)	35.7	43.1	48.5	**51.3**	40.3	47.1	48.6	49.0
Slavic	Czech (cs)	36.7	42.9	43.3	41.7	41.5	**44.7**	43.1	43.7
Slavic	Lithuanian (lt)	25.9	34.9	**41.4**	40.9	34.8	41.1	39.3	38.9
Slavic	Latvian (lv)	43.4	32.5	41.7	**45.1**	34.1	43.5	40.0	41.6
Slavic	Polish (pl)	35.2	34.0	35.5	**39.3**	32.1	38.0	36.4	37.0
Slavic	Slovenian (sl)	33.4	38.8	42.4	42.9	39.6	**44.7**	42.2	42.2
Germanic	Icelandic (is)	32.0	18.9	27.8	**34.9**	19.6	30.4	27.4	28.6
Germanic	Norwegian (no)	40.7	51.1	47.5	43.6	52.5	**54.2**	52.9	53.2
Germanic	Swedish (sv)	46.2	59.0	59.9	60.6	58.3	**63.2**	60.7	61.0
Indic	Assamese (as)	15.9	12.0	17.9	17.9	13.5	17.6	18.6	**18.7**
Indic	Gujarati (gu)	28.9	14.7	25.7	**34.7**	15.2	25.8	24.0	23.6
Indic	Odia (or)	**19.1**	8.1	14.5	17.8	9.3	14.1	13.3	12.2
Indic	Punjabi (pa)	33.5	19.5	27.3	**34.0**	19.4	28.5	24.7	25.6
Indic	Sindhi (sd)	**16.4**	4.5	9.2	14.0	3.4	8.0	9.3	9.0
Indic	Sinhala (si)	**25.4**	3.0	7.8	24.3	2.8	7.6	6.5	6.3
Indic	Urdu (ur)	26.5	19.0	27.2	**33.3**	19.5	29.0	26.1	26.6
Austronesian	Ilocano (ilo)	18.1	21.3	9.4	14.0	**26.6**	9.7	9.9	9.0
Austronesian	Malagasy (mg)	**23.1**	9.5	9.6	14.8	9.6	10.7	10.9	11.1
Austronesian	Māori (mi)	16.9	15.0	17.4	13.8	**18.0**	17.7	17.5	17.6
Austronesian	Sundanese (su)	**32.3**	29.4	15.6	29.3	29.3	15.4	15.7	15.6
Austronesian	Waray (war)	30.9	43.6	17.4	37.1	**43.7**	15.9	13.2	11.0
Romance	Catalan (ca)	49.3	56.9	58.5	55.3	57.8	**59.3**	54.1	53.6

Table 21: Per-language held-out FLORES-200 ChrF (en→xx) ↑, with all targeted interventions at $[5, 11)$ and the postprocessing in Appendix A.6. Expert, Freeze, L2-SP, and E.-Rev. use the model trained on the language's family; Base, Dense, D.-Rev., and Soup are shared models. **Bold** marks the best value at the displayed precision, including ties.