SPELL-OUT: ALIGNING TRAJECTORIES OF CROSS-LINGUAL REALIZATION WHILE PRESERVING SEMANTIC PROGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-Lingual Alignment (CLA) of LLMs has sought to improve multilingual consistency by aligning representations across languages while preserving general capabilities. However, it remains underexplored how to preserve shared semantic progression without collapsing language-specific realization or undermining general capabilities. In this work, we hypothesize that cross-lingual trajectories are better characterized as language-specific realizations of a shared semantic progression, and that CLA should preserve this structure rather than collapse it. We introduce SPELL-OUT, a relational representation alignment objective that uses the source-side trajectory to anchor semantic progression while adapting the target-side realization trajectory. We evaluate SPELL-OUT across diverse multilingual settings and general-capability evaluations. Representation-level analyses further characterize how shared semantic progress is preserved while language-specific trajectory structure remains differentiated. Further analyses show that SPELL-OUT supports diverse target languages under high-resource-language anchors, suggesting broader applicability beyond individual language pairs.

1 INTRODUCTION

Multilingual capabilities of large language models (LLMs) have advanced substantially, with recent models supporting broad sets of languages and showing marked improvements in multilingual downstream performance (Team et al., 2025; Hernández-Cano et al., 2026). Nevertheless, substantial performance differences persist across languages (Singh et al., 2025; Bandarkar et al., 2024), motivating cross-lingual alignment (CLA) to improve capabilities in lower-performing languages. Such adaptation, however, introduces a complementary challenge: improving a target language can degrade capabilities already acquired in others. Prior work has therefore increasingly studied capability preservation during multilingual adaptation, including catastrophic forgetting and mechanisms for retaining general capabilities while training on target languages (Vu et al., 2022; Alexandrov et al., 2024; Liu & Niehues, 2025a; Yamaguchi et al., 2026).

Two established findings are particularly relevant to this problem. First, multilingual representations exhibit both broadly shared and language-sensitive structure (Chang et al., 2022; Xie et al., 2022), and language-sensitive representational variation can materially affect multilingual prediction and capability (Zhao et al., 2026; Ravisankar et al., 2026). Second, model representations evolve across the positions of a sequence rather than residing only in isolated states: relational changes along token-wise hidden-state trajectories carry task-relevant information (Tang et al., 2025), and preserving relational structure across evolving representations can itself provide an effective alignment signal (Chi et al., 2026). Together, these findings establish that multilingual computation contains language-conditioned structure and that relations across evolving representations are informative. What remains unclear is *what cross-lingual relation should be preserved when adapting a target language from a higher-performing source language*.

We approach this question through the hypothesis that parallel languages can express a **shared semantic progression through different language-specific realizations**. We use **semantic progression** to denote the ordered development of meaning as a linguistic sequence unfolds, following accounts of incremental semantic interpretation (Haddock, 1989; Sedivy et al., 1999). **Realization**

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.

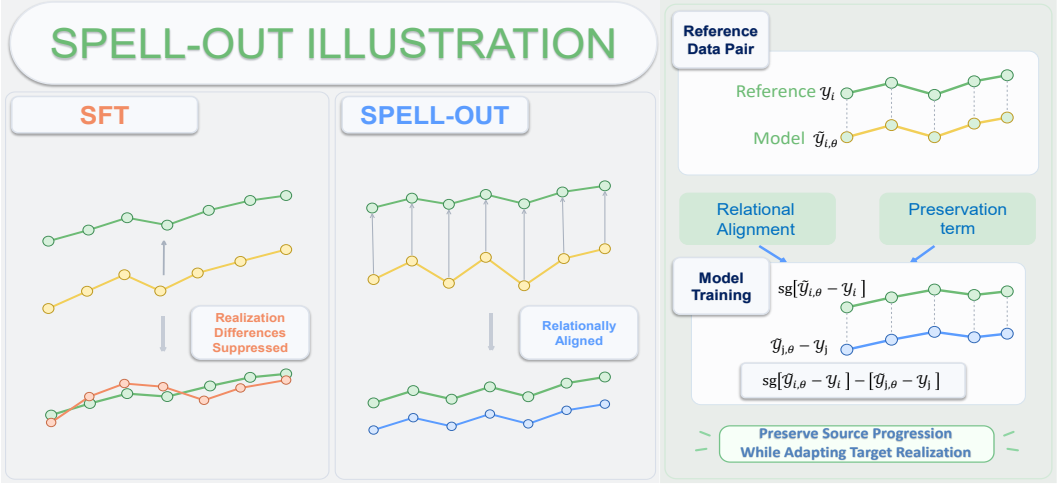


Figure 1: **SPELL-OUT Illustration.** Parallel languages can express corresponding meaning through different language-specific realizations. SFT adapts the target without explicitly preserving its relation to the source. SPELL-OUT instead harnesses teacher-forced representational trajectories over parallel reference sequences to preserve source–target correspondence while maintaining the source-side anchor.

denotes the language-conditioned lexical and grammatical form through which this progression is expressed (Matthiessen, 1999). Because languages differ systematically in how semantic information is organized through lexical choice and syntactic ordering (Dryer, 2005; Ahmad et al., 2019; Arviv et al., 2023), parallel realizations need not develop meaning at identical positions. We therefore use **correspondence** to denote the approximate ordered relation between their semantic progressions. In short, languages can express corresponding meaning differently.

This motivates our central hypothesis: **cross-lingual semantic-progress correspondence can serve as a training constraint that preserves shared semantic progression while adapting language-specific target realization**. Rather than requiring corresponding realizations to become identical, the source provides a relational anchor, the target realization is adapted, and parallel source–target realizations specify the correspondence to be preserved. This view shifts the objective of CLA from identity between realizations toward maintaining the relation between distinct realizations of corresponding meaning.

To test this hypothesis, we introduce **SPELL-OUT**, a source-anchored relational objective for cross-lingual alignment. SPELL-OUT operationalizes the proposed correspondence using teacher-forced representational trajectories over parallel reference sequences, while preserving the source-side anchor during target adaptation. The name draws an analogy to linguistic *Spell-Out* (Chomsky, 1998; Chomsky & Kenstowicz, 2001). Figure 1 illustrates the distinction between corresponding semantic progression and language-specific realization, together with the trajectory-based training signal used by SPELL-OUT. We evaluate SPELL-OUT across diverse multilingual settings and complement downstream evaluation with representation-level analyses that examine preservation of source-side and language-residualized trajectory structure together with adaptation of target-specific structure.

Our contributions are:

- **Preservation-aware cross-lingual alignment.** We formulate cross-lingual alignment as preserving shared semantic progression while adapting language-specific realization, rather than requiring source and target representations to collapse toward identical states.
- **Source-anchored relational alignment.** We introduce SPELL-OUT, which constrains reference-relative source–target trajectory relations so that the source provides an anchor for semantic progression while the target realization is adapted.
- **Behavioral and mechanistic validation.** We evaluate multilingual adaptation and general-capability retention, analyze representation-level evidence for preservation and target adap-

tation, and test whether disrupting cross-lingual semantic-progress correspondence weakens the alignment signal.

- **Cross-lingual generalization.** We evaluate SPELL-OUT across diverse target languages and multiple high-resource-language anchors, testing whether its correspondence-based formulation generalizes beyond a single source language.

2 METHODOLOGY

2.1 SPELL-OUT

Consider parallel realizations in a source language i and a target language j. For $k \in \{i, j\}$, let

$$\begin{aligned}
\mathcal{T}_k &= \left(w_k^{(1)}, \ldots, w_k^{(T_k)}\right), \\
\mathcal{Y}_k &= \left[y_k^{(1)}, \ldots, y_k^{(T_k)}\right], \qquad y_k^{(t)} = E[w_k^{(t)}], \\
\widehat{\mathcal{Y}}_{k,\theta} &= \left[\hat{y}_{k,\theta}^{(1)}, \ldots, \tilde{y}_{k,\theta}^{(T_k)}\right], \qquad \hat{y}_{k,\theta}^{(t)} = \sum_{v \in \mathcal{V}} p_\theta\left(v \mid w_k^{(<t)}\right) E[v],
\end{aligned}$$

where $\mathcal{T}_k$ is the reference token sequence, $E : \mathcal{V} \to \mathbb{R}^{d_e}$ is the model's token-embedding map, $\mathcal{Y}_k$ is the corresponding reference trajectory in token-embedding space, and $\widetilde{\mathcal{Y}}_{k,\theta}$ is the differentiable model-predicted trajectory obtained under teacher forcing. Its empirical construction is described in Section 3.2. Both $\mathcal{Y}_k$ and $\widehat{\mathcal{Y}}_{k,\theta}$ serve only as token-embedding trajectories for constructing the training signal; we do not assume that either explicitly decomposes into semantic or language-specific components of the model's hidden representation.

We define the reference-relative prediction residual at each position as

$$r_{k,\theta}^{(t)} = \hat{y}_{k,\theta}^{(t)} - y_k^{(t)}, \qquad \mathcal{R}_{k,\theta} = \left[r_{k,\theta}^{(1)}, \ldots, r_{k,\theta}^{(T_k)}\right].$$

SPELL-OUT constrains these residuals relationally rather than matching source and target representations pointwise:

$$\mathcal{R}_{j,\theta} \approx \phi_{i \to j}\left(\mathcal{R}_{i,\theta}\right), \qquad \phi_{i \to j} : (\mathbb{R}^{d_e})^{T_i} \to (\mathbb{R}^{d_e})^{T_j}.$$

Here, $\phi_{i \to j}$ is an engineering approximation to cross-lingual progress correspondence, not a token- or word-alignment model. Because parallel realizations generally differ in tokenization and sequence length, we implement $\phi_{i \to j}$ as a progress resampling operator that maps the source residual trajectory onto the target reference grid. Thus, the objective assumes only coarse correspondence in relative progress across the two realizations, rather than one-to-one correspondence between individual linguistic units.

Thus, SPELL-OUT encourages the model-induced source–target relation to follow the relation exhibited by the paired references, rather than forcing source and target token embeddings themselves to coincide. The relational objective is

$$\mathcal{L}_{\mathrm{SPELL}}(i, j) = \frac{1}{T_j} \sum_{t=1}^{T_j} \left\| \mathcal{R}_{j,\theta}^{(t)} - \mathrm{sg}\left[\phi_{i \to j}\left(\mathcal{R}_{i,\theta}\right)^{(t)}\right] \right\|_2^2,$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The stop-gradient makes the alignment asymmetric by design. The source branch provides the relational anchor, whereas the target branch receives the alignment gradient. Without this asymmetry, the relational loss could also be reduced by moving the source-side prediction toward the target. Stop-gradient prevents such direct optimization through the source branch. However, because source and target share model parameters, target-side updates can still indirectly alter source representations. We therefore complement the relational objective with an explicit source-side preservation term.

Let $h_\theta^{(t)} \in \mathbb{R}^{d_h}$ denote the hidden state of the adapted model, and let $h_{\theta_0}^{(t)}$ denote the corresponding state from the frozen pre-adaptation model. For supervised source positions $\mathcal{I}_s$, we define

$$\mathcal{L}_{\mathrm{pres}} = \frac{1}{|\mathcal{I}_s|\, d_h} \sum_{t \in \mathcal{I}_s} \left\| h_\theta^{(t)} - \mathrm{sg}\left(h_{\theta_0}^{(t)}\right) \right\|_2^2.$$

When preservation is applied at a single layer, the sum includes the hidden states from that layer only. Unlike the relational term, which is constructed in token-embedding space, this penalty directly discourages source hidden-state drift during adaptation. It is a practical surrogate for maintaining a reliable source anchor, rather than a direct constraint on the source token-embedding residual.

The complete training objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SPELL-OUT}} + \lambda_{\text{pres}}\mathcal{L}_{\text{pres}}$. The two components of SPELL-OUT play distinct roles. $\mathcal{L}_{\text{pres}}$ is a *preservation* objective: it explicitly limits source-side hidden-state drift relative to the pre-adaptation model, thereby stabilizing the source anchor. $\mathcal{L}_{\text{SPELL-OUT}}$, in contrast, is an *alignment* objective: it constrains the target-side reference-relative trajectory with respect to the preserved source-side relation. Thus, SPELL-OUT does not preserve and align the same quantity. Preservation acts directly on the source hidden trajectory, whereas alignment acts on the source–target relation constructed in token-embedding space.

Under the conditional shared/language-associated decomposition developed in Appendix A.1, these two roles admit a more specific interpretation. If the source-side language-associated trajectory is approximately preserved, relational matching cancels the corresponding shared progression and leaves the target-side language-associated trajectory constrained toward its reference counterpart. This interpretation motivates the combination of preservation and relational alignment, without requiring the model representation itself to be explicitly decomposed during training. Further practical Approximation including embedding and interpolation is provided in Appendix A.3

2.2 Experimental Setup

Training data and language pairs. We construct training pairs from the Aya Dataset (Singh et al., 2024), a multilingual instruction-tuning resource spanning 65 languages. We use higher-resource languages as source anchors and select target languages spanning different resource levels, while controlling source–target relatedness using word-order distance derived from URIEL+ (Khan et al., 2025). We select five language pairs: EN→NL, ID→MS, RU→SR, HI→BN, and ES→RO. For each pair, we retain source and target realizations corresponding to the same underlying instruction, and all adaptation methods use the same paired examples. Further details on language selection and dataset are provided in Appendix B.2 and Appendix B.1, respectively.

Models and baselines. Our main experiments cover four multilingual models from distinct model families: Gemma 3 4B (Team et al., 2025), TranslateGemma 4B-IT (Finkelstein et al., 2026), TinyAya-Global (Salamanca et al., 2026), and Qwen3 4B (Yang et al., 2025). We additionally evaluate Apertus-8B-Base (Hernández-Cano et al., 2026) in a larger-scale setting.

We compare against five controls and baselines. VANILLA denotes the unadapted model. SFT applies conventional target-side cross-entropy supervision and measures supervised adaptation alone. SFT + PRES. and BISFT + PRES. supervise the target language only and both source and target languages, respectively, while applying the same source-side preservation term used by SPELL-OUT; these control for explicit preservation without relational alignment. MLC (Bu et al., 2026) applies multilingual consistency learning to the same paired data and serves as the primary cross-lingual alignment baseline without explicit source preservation. Specifically, it regularizes internal representations to align across semantically equivalent cross-lingual pairs, isolating the contribution of cross-lingual alignment from source-side regularization. Finally, SPELL-OUT combines relational cross-lingual alignment with source-side preservation.

Evaluation. We evaluate target-language adaptation and retention of previously acquired capabilities using Global-MMLU (Singh et al., 2025) and additionally report output-language fidelity. For Global-MMLU, we sample 2,000 examples from the approximately 14K questions available per language using a fixed seed for evaluation efficiency, and use the identical subset, prompt, and decoding configuration across methods. Complementing these behavioral evaluations, RQ1 examines representation-level evidence for the preservation and adaptation predicted by SPELL-OUT, with the corresponding analysis procedures detailed in Appendix C.

Method	EN→NL		ID→MS		RU→SR		HI→BN		ES→RO		Mean	
	Src.	Tgt.	Src.	Tgt.	Src.	Tgt.	Src.	Tgt.	Src.	Tgt.	Src.	Tgt.
VANILLA	**58.0**	51.5	**50.9**	47.5	**50.6**	46.6	46.4	43.8	53.7	50.3	**51.9**	48.0
MLC	57.5	50.7	47.6	44.2	48.2	44.4	43.3	42.3	50.4	46.0	49.4	45.5
SFT	57.6	50.9	48.2	44.8	49.5	45.7	44.6	42.9	51.2	47.5	50.3	46.4
SFT + PRES.	57.7	**51.6**	50.7	47.4	50.5	46.4	46.1	43.9	53.2	50.4	51.6	47.9
BISFT + PRES.	57.8	**51.6**	50.5	47.3	50.2	46.4	45.9	43.9	53.5	50.4	51.6	47.9
SPELL-OUT	57.7	**51.6**	**50.9**	**47.8**	**50.6**	**46.9**	**46.5**	**44.2**	**53.8**	**50.6**	**51.9**	**48.2**

Table 1: **Source- and target-language performance across language pairs.** Results are averaged across the five evaluated models. For each language pair, *Src.* and *Tgt.* report performance on the source and target languages, respectively. The final two columns report averages across the five language pairs.

3 RESULTS

3.1 MAIN RESULTS

Table 1 summarizes source- and target-side performance across five language pairs, averaged over model families. SPELL-OUT achieves the highest or tied-highest target accuracy on four of the five pairs and the highest overall target mean, reaching 48.2%, while retaining a source mean of 51.9%, equal to the pretrained Vanilla baseline. This contrasts with the standard adaptation baselines, which tend to improve or fit the target at the expense of greater source-side degradation.

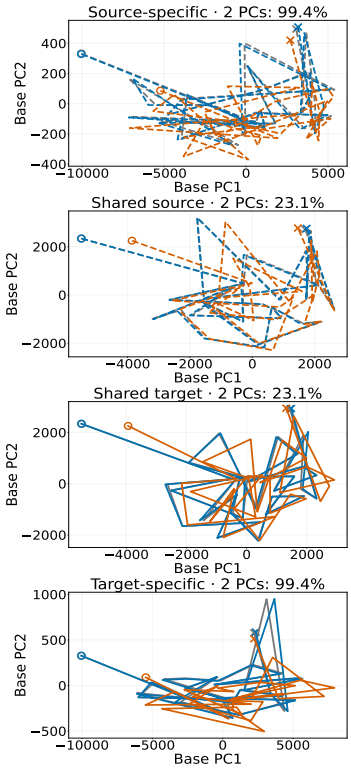
The preservation term itself accounts for an important part of this behavior. Adding source preservation to target SFT raises the aggregate source mean from 50.3% to 51.6% and, rather than suppressing target adaptation, also improves the target mean from 46.4% to 47.9%. A similar pattern is observed for the bidirectional SFT variant with preservation. SPELL-OUT further improves this adaptation–preservation balance, recovering the Vanilla source mean while achieving the strongest aggregate target performance. Thus, the main result is not simply that relational alignment improves the target: explicitly preserving the source is itself beneficial, while SPELL-OUT provides an additional gain by coupling preservation with cross-lingual relational alignment.

The magnitude of these effects, however, varies substantially across model families. The pair-level averages in Table 1 therefore summarize a heterogeneous set of model-specific responses rather than a uniform gain of the same size in every setting. We report the complete model–language-pair breakdown in Appendix C (Table 3). We next examine whether the aggregate advantage of SPELL-OUT is accompanied by the representational preservation and ordered cross-lingual correspondence predicted by our hypothesis.

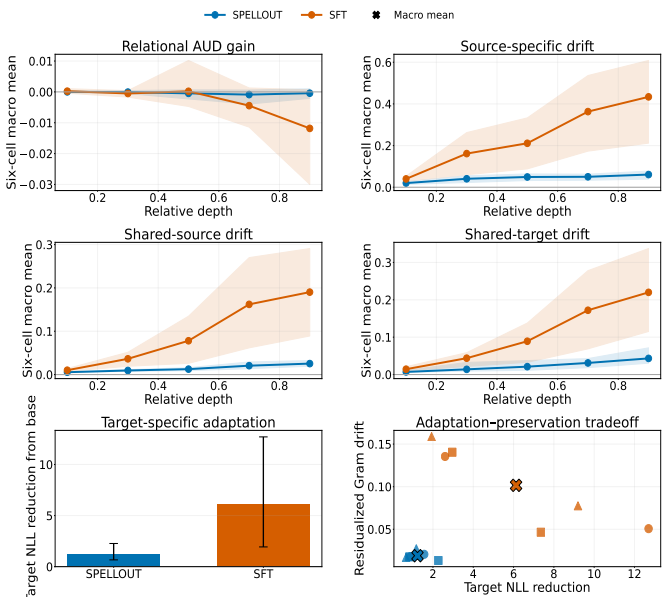
3.2 RQ1: DOES SPELL-OUT PRESERVE AND ALIGN THE INTENDED REPRESENTATIONAL STRUCTURE?

We next examine whether this empirical advantage is accompanied by the representational behavior intended by its objective. Following prior work that treats token-wise hidden-state evolution during generation as a representation trajectory (Tang et al., 2025; Sun et al., 2026), we extract generation-wise hidden states at several relative decoder depths before and after adaptation. A language-sensitive projection is estimated once from the pretrained model and then held fixed across methods and conditions. We analyze both the projected trajectory and its residualized counterpart, where the latter is obtained after removing the projected language-sensitive component. These views are diagnostic rather than ontological: we do not assume that hidden states decompose cleanly into language-specific and language-shared subspaces.

For each trajectory, we compare its pre- and post-adaptation geometry through normalized token–token Gram matrices and representational similarity analysis (RSA). To summarize these changes over generation progress, we compute an instance-level Area Under Difference (AUD): RSA-derived dissimilarity is evaluated along normalized generation progress and aggregated over the full trajectory, after which pre–post change is measured. AUD therefore provides a quantitative trajectory-level measure rather than relying on visual overlap alone. The PCA trajectories in Fig-



(a) PCA visualization of Gram-matrix trajectories in a fixed base-model coordinate system, shown for TranslateGemma on ID→MS at relative depth 0.7. Blue, orange, and gray denote SPELL-OUT, SFT, and VANILLA, respectively.



(b) Quantitative representation-level analysis. Blue denotes SPELL-OUT and orange denotes target-only SFT. The first four panels report six-cell macro means across relative decoder depth for relational AUD gain, source-specific drift, shared-source drift, and shared-target drift. The six cells comprise HI→BN, ID→MS, and ES→RO for TranslateGemma and Qwen3-4B. The lower-left panel measures target-specific adaptation as the reduction in held-out target-reference NLL relative to the pretrained model. The lower-right panel relates this target adaptation to residualized Gram drift, characterizing the adaptation–preservation trade-off. In the cell-wise summaries, ○, △, and □ denote BN, MS, and RO, respectively, and the large × denotes the macro-average across all six model–language cells.

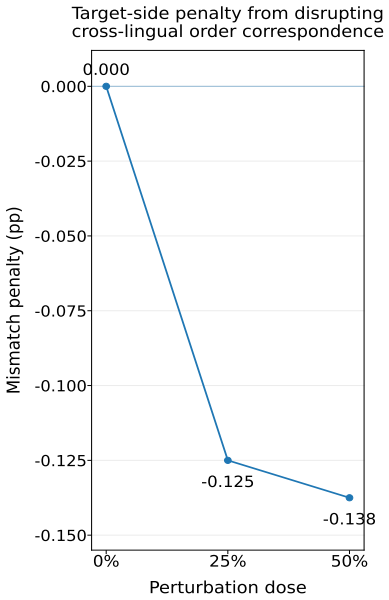
Figure 2: **Representation-level preservation and adaptation under SPELL-OUT.**

ure 2a are consequently used as qualitative illustrations in a fixed VANILLA coordinate system, showing that SPELL-OUT is not deviated from base, unlike SFT.

Figure 2b evaluates the representational properties predicted by SPELL-OUT. First, the source-specific drift panel tests whether adaptation preserves the source-side realization used as the relational anchor. SPELL-OUT remains close to the pretrained source trajectory across decoder depth, whereas SFT induces progressively larger drift, particularly in later layers. This separation is statistically significant across all three TranslateGemma cells at relative depth 0.9 ($p < 0.001$). For Qwen, the separation emerges earlier, reaching significance at depths 0.5 and 0.7 for HI→BN, 0.7 and 0.9 for ES→RO, and 0.5, 0.7, and 0.9 for ID→MS. Thus, although the depth at which the effect emerges is model-dependent, the overall pattern is consistent: SPELL-OUT constrains source-side drift, while SFT increasingly departs from the pretrained trajectory toward the deeper decoder layers.

Second, the shared-source and shared-target panels examine residualized trajectories after removing the language-sensitive projection. SPELL-OUT keeps both comparatively stable, whereas SFT produces substantially larger changes in residualized geometry. The relational AUD panel provides a complementary cross-lingual view: under SPELL-OUT, the source–target trajectory relation remains close to its pretrained configuration across depth, while SFT increasingly disrupts this relation. Together, these diagnostics indicate that SPELL-OUT preserves the source anchor, stabilizes residualized progression on both sides, and maintains the relational structure connecting them.

Preservation, however, does not preclude target adaptation. The lower-left panel shows that SFT achieves a larger reduction in held-out target-reference NLL, as expected from directly optimizing target tokens, whereas SPELL-OUT produces a smaller reduction while inducing substantially



(a) Order-correspondence disruption in JA→KO.

Pair	Method	Acc. (%) ↑	1 − Fidelity (%) ↓
HI→NE	Vanilla	41.48	**0.58**
	SFT+Pres	41.70	0.68
	SPELL-OUT	**41.90**	**0.58**
JA→NE	Vanilla	41.48	**0.43**
	SFT+Pres	41.70	0.68
	SPELL-OUT	**42.00**	**0.43**
HI→TE	Vanilla	41.25	1.85
	SFT+Pres	**41.90**	**1.20**
	SPELL-OUT	**41.90**	1.85
JA→TE	Vanilla	41.25	1.68
	SFT+Pres	41.90	**1.20**
	SPELL-OUT	**42.20**	1.68

(b) Source-language sensitivity on low-resource targets. Values are averaged across TranslateGemma-4B-IT and Qwen3-4B. Accuracy is better when higher, while 1 − Fidelity is the fraction of invalid outputs among 2,000 instances and is therefore better when lower.

Figure 3: **Correspondence and source-language sensitivity analyses. (a)** Increasing mismatch in source–target order correspondence produces a larger target-side performance penalty in JA→KO. **(b)** Accuracy and output failures over 2,000 instances per condition for low-resource NE and TE targets. SPELL-OUT improves over Vanilla in all eight settings. Replacing HI with the more genealogically distant JA improves SPELL-OUT accuracy.

less source-specific and residualized trajectory drift. The lower-right panel makes this trade-off explicit: across model–language cells, SPELL-OUT occupies the low-drift region, while SFT obtains stronger reference fitting at the cost of markedly larger representational change. Crucially, this stronger reference fitting does not translate into better downstream target performance: in the main results (Table **??**), SPELL-OUT outperforms SFT despite its smaller NLL reduction. These results suggest that SPELL-OUT does not merely weaken SFT, but instead constrains adaptation so that target improvement occurs while substantially more of the pretrained relational structure is retained.

Taken together, the relevant criterion is not whether the target representation becomes globally more source-like. Rather, SPELL-OUT is intended to permit target adaptation while preserving the source anchor and the relational trajectory structure linking the two realizations. The analyses above support this behavior without requiring the projected and residualized views to correspond to uniquely identifiable linguistic components. Full data of p-values and trajectories are provided in Appendix D

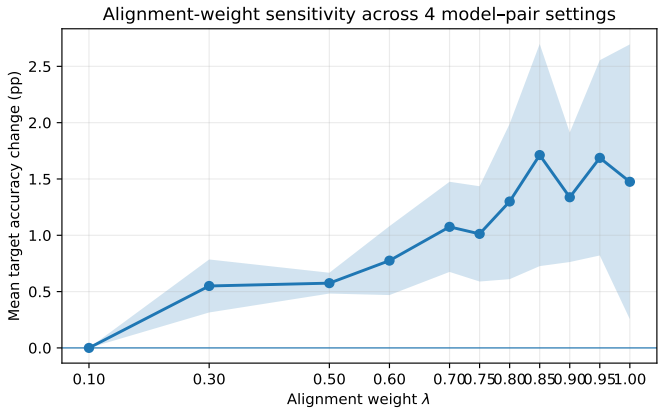
3.3 RQ2: Does Ordered Cross-Lingual Correspondence Matter?

SPELL-OUT relies on approximate correspondence between the semantic progression of parallel source and target realizations. Although this correspondence does not require token-level alignment, it assumes that the two trajectories retain a meaningful ordering relation as their respective sequences unfold. We therefore ask whether this ordered correspondence contributes to cross-lingual transfer, rather than parallel semantic content alone.

To test whether SPELL-OUT uses ordered cross-lingual correspondence, we construct controlled reordering interventions on Japanese–Korean parallel data. We alter the surface order of selected subject and object constituents while preserving their grammatical roles and intended semantic content. Source and target realizations are perturbed independently, yielding *correspondence-preserving* conditions when the same reordering is applied to both sides and *mismatched* conditions when only one side is reordered. We repeat the intervention at increasing perturbation rates and manually audit the resulting examples for grammaticality and preservation of the intended interpretation. We sum-

p	Source	Target	Avg.
0.00	49.67	47.16	48.42
0.25	49.74	47.11	48.42
0.50	49.71	**47.23**	48.47
0.75	**49.84**	47.11	**48.48**
1.00	49.76	47.01	48.39
1.25	49.72	47.17	48.44
1.50	49.56	47.15	48.36

(a) Preservation-coefficient sweep averaged across Apertus 8B, Qwen3 4B, and TinyAya Global.



(b) Target-language performance under different mixtures of SFT and SPELL-OUT.

Figure 4: **Preservation and objective-mixture analyses. (a)** Global-MMLU accuracy under different preservation coefficients p, with other training settings held fixed. Moderate preservation improves source retention without consistently sacrificing target performance. **(b)** Target-language Global-MMLU accuracy under $(1-\lambda)\mathcal{L}_{\mathrm{SFT}}+\lambda\mathcal{L}_{\mathrm{SPELL}}$. Bars report mean accuracy across TinyAya-Global and TranslateGemma on ID→MS and HI→BN; error bars denote the standard error across the four model–pair settings.

marize the effect as the *correspondence advantage*, defined as the performance difference between correspondence-preserving and mismatched conditions. A positive advantage indicates that ordered source–target correspondence contributes beyond parallel semantic content alone. We evaluate this effect primarily on Korean, with Japanese serving as a source-side control for generic perturbation effects.

Figure 3a shows the opposite pattern on the KO target. Once ordering correspondence is disrupted, mismatched training pairs perform worse than their correspondence-preserving counterparts, with an average penalty of 0.125 %p at a 25% perturbation dose and 0.138 %p at 50%. The penalty therefore remains consistently negative and becomes slightly larger as a greater portion of the training data is perturbed. This pattern suggests that SPELL-OUT benefits not only from parallel semantic content, but also from preserving the ordered correspondence between source and target trajectories. The corresponding JA source-side control remains near zero and does not exhibit the same dose-dependent degradation, indicating that the effect is concentrated on cross-lingual target transfer rather than reflecting a generic cost of the perturbation. Details are provided in Appendix **??**.

3.4 How sensitive is SPELL-OUT to the choice of source language?

Table 3b examines whether SPELL-OUT depends on selecting a linguistically close source language. Across two models, two low-resource targets, and two source languages, SPELL-OUT improves over the unadapted model in all eight settings, indicating that its gains are not tied to a particular source choice. Replacing HI with the more genealogically distant JA source generally improves performance. This is especially notable for NE, for which HI is substantially more closely related, yet JA remains competitive or performs better. Fidelity analysis also shows a similar tendency. These results suggest that source selection may depend less on genealogical relatedness itself than on the quality of the source anchor and its cross-lingual correspondence with the target.

4 Discussion

4.1 Effect of Preservation Strength

SPELL-OUT is relatively robust to the preservation coefficient p, showing only modest variation across $p \in [0, 1.5]$. A moderate weight performs best overall: $p = 0.75$ yields the highest average accuracy, and most intermediate settings also improve source performance over $p = 0$. The trend

is non-monotonic and diminishes at $p = 1.5$, indicating that preservation is beneficial without requiring an aggressively strong constraint. Overall, the results favor a moderate preservation regime rather than fine tuning or maximizing p.

4.2 Interaction with Supervised Fine-Tuning

We vary the relative weights of $\mathcal{L}_{\mathrm{CE}}$ and $\mathcal{L}_{\mathrm{SPELL}}$ while fixing $\lambda_{\mathrm{pres}} = 0$, sweeping from SFT-only to relational-only training: $\mathcal{L}_{\mathrm{total}} = (1-\lambda_{\mathrm{CE}})\mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{CE}}\mathcal{L}_{\mathrm{SPELL}} + \lambda_{\mathrm{pres}}\mathcal{L}_{\mathrm{pres}}$. Target performance generally improves as training becomes more SPELL-OUT-dominant and remains strong at $\lambda = 1$, where SFT is removed entirely. Thus, SPELL-OUT need not serve only as an auxiliary regularizer to SFT; its relational signal can itself drive effective target adaptation. Because the preferred coefficient still varies across settings, we interpret this as evidence that dominant SFT supervision is unnecessary, rather than that larger λ is universally better.

5 Related Work

Multilingual Alignment and Preservation. Cross-lingual alignment aims to reduce representational or behavioral discrepancies and improve multilingual transfer, while adaptation can induce catastrophic forgetting, motivating preservation of source-language or general capabilities (Vu et al., 2022; Alexandrov et al., 2024; Liu & Niehues, 2025a; Yamaguchi et al., 2026). Related work broadens what should be preserved: Bu et al. (2026) promote representational consistency for multilingual safety while limiting utility degradation, whereas Han et al. (2025) show that stronger convergence can conflict with culturally localized behavior and use layer-specific interventions to balance transfer and localization. Thus, prior preservation targets capability, safety, utility, or cultural localization; SPELL-OUT instead preserves an explicit source-side semantic-progress anchor while relationally adapting the target-language realization.

Multilingual Representations and Trajectory-Based Alignment. Multilingual representations contain both cross-lingually shared and language-sensitive structure (Chang et al., 2022; Xie et al., 2022), with semantically related inputs approaching shared structure across languages (Wu et al., 2025). Layer-wise studies characterize this organization from different perspectives: Wendler et al. (2024) identify English-biased latent processing, while Dumas et al. (2025) provide causal evidence for language-agnostic concept representations. Other work links such structure to multilingual capability (Kargaran et al., 2025; Ravisankar et al., 2026) or intervenes through language-associated subspaces, middle-layer alignment, and activation steering (Zhao et al., 2026; Liu & Niehues, 2025b; Sundar et al., 2025). Trajectory-based methods move beyond isolated states: Tang et al. (2025) model token-wise hidden-state evolution during generation as task-relevant trajectories, whereas Chi et al. (2026) use relational structure across layer-wise transformation trajectories for distillation. In contrast, neither models correspondence between parallel linguistic realizations while preserving a multilingual source anchor; SPELL-OUT combines these directions through source-preserving relational alignment over cross-lingual sequence trajectories, using coarse progress correspondence without requiring representational coincidence.

6 Conclusion

We introduced **SPELL-OUT**, a source-anchored relational objective motivated by the hypothesis that parallel languages can express corresponding semantic progression through different language-specific realizations. Across multilingual settings, SPELL-OUT improves target-language adaptation while preserving capabilities established in higher-performing source languages. Our analyses further show that its behavior is consistent with the proposed mechanism: source-side and cross-lingually shared progression are better preserved, target-specific realization remains adaptable, and preserving ordered cross-lingual correspondence matters for target improvement. Together, these findings suggest that cross-lingual alignment need not make languages representationally identical; preserving the relation between distinct language-specific realizations provides an alternative principle for multilingual adaptation.

AI USE STATEMENT

In this work, we used OpenAI ChatGPT (GPT-5.6) to assist with manuscript drafting and language polishing, retrieval and discovery of related literature, research ideation, experimental design, and implementation-related problem solving. The tool was also used to suggest analyses and alternative formulations of experimental and representational results. Generative AI was not used to generate experimental data, benchmark labels, reported numerical results, or citations without subsequent verification, and it did not autonomously determine which results or claims were included in the paper. All AI-assisted text, references, research ideas, experimental suggestions, analyses, and interpretations were reviewed by the authors. Citations were checked against the corresponding literature, and reported results and numerical claims were checked against the authors' implementations and experimental outputs. The authors take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the model identifiers, language pairs, datasets, prompt templates, training objectives, loss coefficients, batch settings, checkpoint schedule, decoding parameters, and evaluation procedures used in all experiments. Unless otherwise stated, methods are trained on the same parallel data and with the same optimization budget, and are evaluated at fixed checkpoints (50 and 100 steps) under identical source- and target-side settings. Model-wise results underlying the aggregated main table are reported in Appendix **??**. For the representation analyses, we use the same layer-depth sampling, language-sensitive projection, residualization, and trajectory-comparison procedures across methods. For the correspondence intervention, Japanese–Korean examples are deterministically reordered into correspondence-preserving and mismatched conditions and manually audited for content and grammatical-role preservation. We retain the raw model outputs and analysis scripts used to compute all reported metrics, statistical tests, and figures, and provide the corresponding implementation and experiment configurations for reproduction.

REFERENCES

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2440–2452, 2019.

Anton Alexandrov, Veselin Raychev, Mark Niklas Müller, Ce Zhang, Martin Vechev, and Kristina Toutanova. Mitigating catastrophic forgetting in language transfer via model merging. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 17167–17186, 2024.

Ofir Arviv, Dmitry Nikolaev, Taelin Karidi, and Omri Abend. Improving cross-lingual transfer through subtree-aware word reordering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 718–736, 2023.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, 2024.

Yuyan Bu, Xiaohao Liu, ZhaoXing Ren, Yaodong Yang, and Juntao Dai. Align once, benefit multilingually: Enforcing multilingual consistency for llm safety alignment. *arXiv preprint arXiv:2602.16660*, 2026.

Tyler A Chang, Zhuowen Tu, and Benjamin K Bergen. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 119–136, 2022.

Pham Khanh Chi, Quoc Phong Dao, Thut Nguyn, Linh Ngo Van, Trung Le, and Thanh Hong Nguyen. Mta: Multi-granular trajectory alignment for large language model distillation. In *Pro-*

ceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 32669–32684, 2026.

Noam Chomsky. Minimalist inquiries: The framework. *MIT occasional papers in linguistics*, 15, 1998.

Noam Chomsky and Michael Kenstowicz. Derivation by phase. 2001.

Matthew S Dryer. 81 order of subject, object, and verb. *The world atlas of language structures, ed. by Martin Haspelmath et al*, pp. 330–333, 2005.

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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31822–31841, 2025.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, et al. Translategemma technical report. *arXiv preprint arXiv:2601.09012*, 2026.

Nicholas J Haddock. Computational models of incremental semantic interpretation. *Language and Cognitive Processes*, 4(3-4):SI337–SI368, 1989.

HyoJung Han, Sweta Agrawal, and Eleftheria Briakou. Rethinking cross-lingual alignment: Balancing transfer and cultural erasure in multilingual llms. *arXiv preprint arXiv:2510.26024*, 2025.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, et al. Apertus: Democratizing open and compliant llms for global language environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46877–46955, 2026.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. Mexa: Multilingual evaluation of english-centric llms via cross-lingual alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 27001–27023, 2025.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H Hoang, Eric Khiu, A Seza Doğruöz, and En-Shiun Annie Lee. Uriel+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6937–6952, 2025.

Danni Liu and Jan Niehues. Conditions for catastrophic forgetting in multilingual translation. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pp. 347–359, 2025a.

Danni Liu and Jan Niehues. Middle-layer representation alignment for cross-lingual transfer in fine-tuned llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15979–15996, 2025b.

MAK Halliday-Christian MIM Matthiessen. Construing experience through meaning. a language-based approach to cognition. *Általános Nyelvészeti Tanulmányok XXII.*, pp. 501, 1999.

Kartik Ravisankar, Hyojung Han, Sarah Wiegreffe, and Marine Carpuat. Can you map it to english? the role of cross-lingual alignment in the multilingual performance of llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4854–4872, 2026.

Alejandro R Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, et al. Tiny aya: Bridging scale and multilingual depth. *arXiv preprint arXiv:2603.11510*, 2026.

Julie C Sedivy, Michael K Tanenhaus, Craig G Chambers, and Gregory N Carlson. Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71(2):109–147, 1999.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, et al. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, 2024.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, 2025.

Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. Llm reasoning as trajectories: Step-specific representation geometry and correctness signals. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26872–26887, 2026.

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2375–2401, 2025.

Yichen Tang, Weihang Su, Yujia Zhou, Yiqun Liu, Min Zhang, Shaoping Ma, and Qingyao Ai. Augmenting multi-agent communication with state delta trajectory. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10230–10251, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9279–9300, 2022.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, 2024.

Zhaofeng Wu, Xinyan Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *International Conference on Learning Representations*, volume 2025, pp. 53705–53723, 2025.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5617–5633, 2022.

Atsuki Yamaguchi, Terufumi Morishita, Aline Villavicencio, and Nikolaos Aletras. Mitigating catastrophic forgetting in target language adaptation of llms via source-shielded updates. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18949–18975, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, et al. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *Advances in Neural Information Processing Systems*, 38:38608–38642, 2026.

A SPELL-OUT INTERPRETATION

A.1 SPELL-OUT ALIGNMENT TERM

We provide an idealized decomposition to clarify the role of the relational objective. This decomposition is used only for analysis; SPELL-OUT itself does not assume that semantic progression and language-specific realization occupy explicitly separable representational subspaces, nor that either representation space or parameter space admits a clean decomposition into such components. For readability, we consider source and target trajectories on a common normalized-progress grid, with source-side quantities understood to have been mapped through $\phi_{i\to j}$.

Suppose that the reference trajectory for language $k \in \{i, j\}$ can be written as

$$\mathcal{Y}_k = \mathcal{Y}_{\text{shared}} + \mathcal{Y}_{\text{spec},k},$$

and that the model-predicted trajectory after adaptation can analogously be written as

$$\hat{\mathcal{Y}}_{k,\theta} = \hat{\mathcal{Y}}_{\text{shared},\theta} + \hat{\mathcal{Y}}_{\text{spec},k,\theta}.$$

Here, i denotes a competent source language and j the target language. Following the notation in the main method, the reference-relative residual is

$$\mathcal{R}_{k,\theta} = \hat{\mathcal{Y}}_{k,\theta} - \mathcal{Y}_k.$$

Under the idealized decomposition,

$$\begin{aligned}\mathcal{R}_{k,\theta} &= \left(\hat{\mathcal{Y}}_{\text{shared},\theta} - \mathcal{Y}_{\text{shared}}\right) + \left(\hat{\mathcal{Y}}_{\text{spec},k,\theta} - \mathcal{Y}_{\text{spec},k}\right) \\ &= \Delta_{\text{shared}} + \Delta_{\text{spec},k}.\end{aligned}$$

We denote the magnitude of this residual over the normalized trajectory by

$$E_k = \|\mathcal{R}_{k,\theta}\|_{\mathcal{T}}.$$

Accordingly,

$$E_i = \left\|\hat{\mathcal{Y}}_{i,\theta} - \mathcal{Y}_i\right\|_{\mathcal{T}}$$

measures how far the source trajectory departs from its reference realization, whereas

$$E_j = \left\|\hat{\mathcal{Y}}_{j,\theta} - \mathcal{Y}_j\right\|_{\mathcal{T}}$$

is the corresponding target realization error. We characterize a competent source by the anchor assumption

$$E_i \le \epsilon_i, \qquad \epsilon_i \text{ small}.$$

This assumption requires only that the source remain a sufficiently reliable reference; it does not require its language-specific component to remain exactly fixed.

Ordinary target fitting directly encourages

$$\hat{\mathcal{Y}}_{j,\theta} \approx \mathcal{Y}_j,$$

or equivalently,

$$\mathcal{R}_{j,\theta} \approx 0.$$

However, under the decomposition above,

$$\mathcal{R}_{j,\theta} = \Delta_{\text{shared}} + \Delta_{\text{spec},j},$$

so a small total residual does not imply that either component has individually been preserved. For example,

$$\Delta_{\text{shared}} = \delta, \qquad \Delta_{\text{spec},j} = -\delta$$

still yields

$$\mathcal{R}_{j,\theta} = 0.$$

Thus, fitting the target trajectory alone permits compensating changes between shared and language-specific structure.

SPELL-OUT instead constrains the target residual relative to the corresponding source residual:

$$\mathcal{R}_{j,\theta} \approx \phi_{i\to j}\left(\mathcal{R}_{i,\theta}\right).$$

Under the common-grid notation used for this derivation, we suppress $\phi_{i\to j}$ and write this relation as

$$\mathcal{R}_{j,\theta} \approx \mathcal{R}_{i,\theta}.$$

Substituting the idealized decomposition gives

$$\begin{aligned}\mathcal{R}_{j,\theta} - \mathcal{R}_{i,\theta} &= (\Delta_{\text{shared}} + \Delta_{\text{spec},j}) - (\Delta_{\text{shared}} + \Delta_{\text{spec},i}) \\ &= \Delta_{\text{spec},j} - \Delta_{\text{spec},i}.\end{aligned}$$

The shared perturbation therefore cancels from the relational discrepancy. Importantly, this cancellation does not require

$$\hat{\mathcal{Y}}_{\text{spec},i,\theta} \approx \mathcal{Y}_{\text{spec},i}$$

as a hard assumption. Instead, the source need only remain sufficiently close to its reference trajectory for its residual to provide a useful anchor.

The same constraint can be rearranged as

$$\begin{aligned}\mathcal{R}_{j,\theta} \approx \mathcal{R}_{i,\theta} &\iff \hat{\mathcal{Y}}_{j,\theta} - \mathcal{Y}_j \approx \hat{\mathcal{Y}}_{i,\theta} - \mathcal{Y}_i \\ &\iff \hat{\mathcal{Y}}_{j,\theta} - \hat{\mathcal{Y}}_{i,\theta} \approx \mathcal{Y}_j - \mathcal{Y}_i.\end{aligned}$$

SPELL-OUT therefore preserves the reference-relative source–target contrast rather than forcing the two language trajectories toward the same representation.

This relation also gives a useful interpretation of source preservation. Define the remaining relational discrepancy as

$$D_{\text{SPELL}} = \left\|\mathcal{R}_{j,\theta} - \mathcal{R}_{i,\theta}\right\|_{\mathcal{T}}.$$

Within the idealized common-grid setting, the triangle inequality gives

$$\begin{aligned}E_j &= \left\|\mathcal{R}_{j,\theta}\right\|_{\mathcal{T}} \\ &\le \left\|\mathcal{R}_{j,\theta} - \mathcal{R}_{i,\theta}\right\|_{\mathcal{T}} + \left\|\mathcal{R}_{i,\theta}\right\|_{\mathcal{T}} \\ &= D_{\text{SPELL}} + E_i \le D_{\text{SPELL}} + \epsilon_i.\end{aligned}$$

Thus, the target realization error is bounded by

$$E_j \le D_{\text{SPELL}} + \epsilon_i,$$

where D_{SPELL} measures the remaining relational mismatch and ϵ_i bounds the source-anchor error. This inequality is exact under the common-grid quantities defined above; what is approximate in practice is the correspondence map $\phi_{i\to j}$ used to place the two trajectories on that grid.

A.2 SPELL-OUT Preservation Term

The bound also motivates the second part of the final objective. Minimizing the relational term reduces D_{SPELL} whereas the second term,

$$\left\|\mathcal{R}_{i,\theta}\right\|_{\mathcal{T}} = E_i,$$

shows why the source anchor itself must remain stable. Stop-gradient prevents direct optimization through the source branch, but it does not guarantee a small E_i, because source and target share model parameters. The explicit preservation term is therefore introduced as an engineering mechanism for maintaining the small-source-error condition required by the bound.

Equivalently, the relational constraint can be written as

$$\underbrace{\hat{\mathcal{Y}}_{j,\theta}}_{\text{target realization}} \approx \underbrace{(\mathcal{Y}_j - \mathcal{Y}_i)}_{\text{reference realization shift}} + \underbrace{\hat{\mathcal{Y}}_{i,\theta}}_{\text{source realization}},$$

with source-side quantities mapped through $\phi_{i\to j}$ when necessary. This yields the relational objective

$$\mathcal{L}_{\text{SPELL}} = \left\|\mathcal{R}_{j,\theta} - \text{sg}[\phi_{i\to j}\left(\mathcal{R}_{i,\theta}\right)]\right\|_{\mathcal{T}}^2.$$

The error bound $E_j \leq D_{\text{SPELL}} + \epsilon_i$ additionally requires the source to remain a reliable anchor. Stop-gradient prevents the relational loss from directly moving the source branch, but shared parameters allow target-side updates to alter source representations indirectly. We therefore regularize source drift with a hidden-state preservation surrogate.

Thus, The final objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SPELL}} + \lambda_{\text{pres}} \mathcal{L}_{\text{pres}}.$$

A.3 Practical Approximation

Teacher-forced soft token embeddings. For each reference position t, the model is evaluated under the reference prefix $w_{k,<t}$, producing logits $\ell_{k,\theta}^{(t)}$. An argmax token would provide a discrete prediction but would not yield a differentiable relational signal. We therefore construct a continuous summary of the predicted token distribution in the same token-embedding space as the reference:

$$\hat{y}_{k,\theta}^{(t)} = \sum_{v \in \text{TopK}_K(\ell_{k,\theta}^{(t)})} \hat{p}_{k,\theta}^{(t)}(v) E[v] \qquad s.t. \quad \hat{p}_{k,\theta}^{(t)}(v) = \frac{\exp(\ell_{k,\theta,v}^{(t)}/\tau)}{\sum_{u \in \text{TopK}_K(\ell_{k,\theta}^{(t)})} \exp(\ell_{k,\theta,u}^{(t)}/\tau)}.$$

We use $K = 16$ and $\tau = 0.3$ by default. The soft token embedding is therefore a differentiable readout-derived quantity, not a hidden representation. Its gradient propagates through the logits and the hidden states that produce them.

Teacher forcing ensures that source and target residuals are computed under their respective reference histories. Consequently, the relational signal reflects local prediction differences at each reference position rather than differences caused by accumulated autoregressive errors. The soft embedding is used only to compute the objective and is never fed back as the next model input.

Normalized-progress approximation. Correspondence in our hypothesis refers to a coarse relation in semantic progression between parallel linguistic realizations. It does not imply token identity, word alignment, or identical word order. To obtain a tractable engineering approximation, we assign each reference position a normalized progress coordinate and resample the source sequence onto the target grid. For target position t,

$$\phi_{i \to j}(\mathcal{X})^{(t)} = \text{LinearInterp}(\mathcal{X}, q_t) \qquad s.t. \quad q_t = 1 + \frac{t-1}{T_j - 1}(T_i - 1).$$

Linear interpolation is therefore only a monotonic reparameterization by normalized sequence position; it is not itself the linguistic correspondence assumed by SPELL-OUT. Likewise, word order is not identified with correspondence. Rather, differences in how languages order linguistic material may make normalized position a better or worse approximation to the underlying coarse correspondence. We examine this sensitivity experimentally by perturbing or comparing ordering correspondence, rather than encoding word-order alignment into $\phi_{i \to j}$.

The interpolation grid is determined entirely by the reference sequence lengths and is therefore fixed with respect to model predictions. The same mapping is applied to the source reference-relative residual before computing $\mathcal{L}_{\text{SPELL}}$. Rather than relying on a single end-to-end linear interpolation, we use a split interpolation that anchors four progress points—the beginning, end of instruction, end of input, and end of each trajectory—and interpolates separately within the two resulting intervals. This piecewise construction provides a simple safeguard against excessive distortion from a single global linear map while retaining a coarse, annotation-free notion of progress correspondence.

B Data

B.1 Training Data

For each language pair, we use 1,200 parallel instruction-tuning examples, with one source-language and one target-language realization for each underlying example. The data follow the `Instruction`/`Input`/`Output` structure of the Aya Dataset (Singh et al., 2024). During training, both language realizations are teacher-forced over all available fields, so the relational trajectory is constructed from the same supervised portions of the paired examples. All compared adaptation methods use the same manifests and data exposure.

Pair	Genetic	Geographic†	Syntactic	Phonological	Inventory
EN–NL	0.7068	0.0174	0.3625	NA‡	0.4861
ID–MS	0.0002	0.0600	0.0000	NA‡	0.3578
RU–SR	0.6162	0.1506	0.3255	0.0000	NA‡
HI–BN	0.7180	0.0542	0.3729	0.3030	0.3481
ES–RO	0.5804	0.0875	0.3246	0.3217	0.5114
Main-pair mean	0.5243	*0.0739*	0.2771	*0.2082*	*0.4259*
JA–KO	1.0000	0.0282	0.3968	0.7048	0.5204
HI–NE	0.6162	0.0355	0.4500	0.2301	0.3787
HI–TE	1.0000	0.0399	0.4752	0.2163	0.3131
JA–HI	1.0000	0.2209	0.5682	0.5064	0.4980
JA–NE	1.0000	0.1892	0.3763	0.5662	0.5003

Table 2: **URIEL+ distances for the language pairs used in our experiments.** Values closer to 1 indicate greater distance between the two languages. *Genetic* captures genealogical separation based on language-family relations; *Geographic* reflects physical separation between the languages; *Syntactic* measures differences in syntactic typological features; *Phonological* measures differences in phonological typological properties; and *Inventory* measures differences in the languages' phonological segment inventories. The first five rows are the language pairs used in the main experiments, selected to span heterogeneous genealogical and structural relationships; their mean distances are reported immediately below and are computed over the available values only. JA–KO is used for the controlled order-correspondence analysis. For the source-language sensitivity analysis, HI and JA provide contrasting source anchors, while NE and TE provide low-resource targets with substantially different relationships to those sources. In particular, HI–NE retains appreciable genealogical relatedness, whereas HI–TE is maximally distant genetically, and JA provides a substantially more distant alternative source. This allows us to test whether SPELL-OUT requires a linguistically close source or can transfer effectively from a more distant but suitable anchor. †Geographic distance is reported for completeness but is not used as a language-selection criterion, because physical distance alone provides limited information about the linguistic relationship between two languages. ‡NA indicates that URIEL+ cannot compute a reliable pairwise distance from the available comparable features: for *Phonological*, the two languages do not have sufficient overlapping phonological-feature coverage, while for *Inventory*, sufficient comparable segment-inventory information is unavailable for the pair.

B.2 Language Selection

Language pairs are selected under three constraints. First, the source language is chosen from a comparatively high-resource language that can provide a strong anchor. Second, target languages are selected to cover different resource regimes. Third, we control source–target ordering similarity using the word-order features provided by URIEL+ (Khan et al., 2025). We compute the corresponding continuous URIEL+ similarity for each candidate pair rather than relying only on coarse typological categories.

We additionally restrict the candidate languages to those covered by both Global-MMLU (Singh et al., 2025) and Belebele (Bandarkar et al., 2024), so that the same language pairs can be evaluated across both benchmarks. Table **??** reports the selected pairs together with their resource grouping and URIEL+ word-order similarity.

B.3 RQ2 Perturbation Data and Manual Audit

For the controlled correspondence experiment in RQ2, we construct 140 Japanese–Korean parallel examples from Aya dataset. Constituent-order perturbations are generated automatically using Stanza 1.10.1. For Japanese, we use the GSD Universal Dependencies models with Sudachi-pretokenized input; for Korean, we use the KAIST eojeol/UD models. The processing pipeline uses `tokenize`, `pos`, `lemma`, and `depparse` to identify constituents eligible for reordering.

At the highest perturbation condition, at most 70 of the 140 examples are reordered. Source and target realizations are manipulated independently, allowing correspondence-preserving conditions, in which both sides receive the same ordering intervention, to be contrasted with mismatched conditions, in which only one side is perturbed.

Because automatic dependency-based reordering can introduce unnatural or semantically disruptive examples, We manually audited the resulting Japanese and Korean data using proficiency in both languages. The audit checks that the intended proposition is retained, the manipulated constituents correspond to the intended syntactic roles, and the resulting realization remains interpretable in the respective language.

C Evaluation

C.1 Global MMLU Sampling

We first sample 160,000 Global MMLU examples using a fixed random seed of 42. From this pool, we resample 2,400 examples and use the first half (1,200 examples) for training.

C.2 Evaluation Prompts

For multiple-choice evaluation, all methods use the same prompt:

Prompt 1: Prompt for multiple-choice evaluation

```
<user>
Choose the correct option. Give its letter (A, B, C, or D). You may
briefly explain after the letter.
Question: {question}
A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}
Answer:
```

C.3 Model-wise Full Results

D Representation analysis

D.1 Layer-wise Interpretation

D.2 Full Trajectory Visualization

Model	Method	EN→NL		ID→MS		RU→SR		HI→BN		ES→RO		Mean	
		Src.	Tgt.	Src.	Tgt.	Src.	Tgt.	Src.	Tgt.	Src.	Tgt.	Src.	Tgt.
TranslateGemma-4B-IT	VANILLA	56.0	**49.4**	**48.9**	**44.1**	48.8	43.9	**45.9**	43.3	**51.7**	47.7	**50.2**	45.7
	MLC	54.3	46.2	42.1	35.4	42.5	37.9	37.9	39.3	43.1	37.7	44.0	39.3
	SFT	53.9	46.7	43.5	36.8	44.9	40.8	40.6	40.6	46.0	41.8	45.8	41.3
	SFT + PRES.	55.8	48.6	48.0	43.9	47.9	42.9	45.2	43.6	51.2	**48.0**	49.6	45.4
	BISFT + PRES.	**56.1**	48.5	48.1	43.6	46.8	42.6	44.7	43.5	51.1	47.8	49.4	45.2
	SPELL-OUT	55.6	49.1	48.8	43.9	**48.9**	**44.2**	45.7	**44.0**	51.3	**48.0**	50.0	**45.8**
Gemma3-4B	VANILLA	55.5	49.2	49.3	44.6	46.7	42.5	**45.7**	43.6	52.0	48.2	49.8	45.6
	MLC	55.5	48.4	41.3	37.3	41.4	37.2	38.7	39.3	44.3	37.1	44.2	39.9
	SFT	**55.8**	49.0	43.3	38.2	45.9	41.8	42.4	41.9	46.8	39.2	46.8	42.0
	SFT + PRES.	55.1	48.9	**49.3**	44.5	47.0	42.7	45.1	43.8	51.9	48.0	49.7	45.5
	BISFT + PRES.	55.5	49.1	49.1	44.6	**47.5**	42.5	45.6	**44.0**	52.1	48.1	**49.9**	45.6
	SPELL-OUT	55.4	**49.2**	49.1	**44.7**	46.9	**42.9**	45.6	43.8	**52.1**	**48.3**	49.8	**45.8**
TinyAya-Global	VANILLA	**55.7**	51.8	50.1	49.2	**50.7**	49.6	47.7	44.8	51.9	50.2	51.2	49.1
	MLC	55.4	51.6	50.1	48.8	50.6	49.7	**48.6**	44.8	52.0	49.9	**51.3**	48.9
	SFT	55.4	51.5	**50.3**	48.8	50.2	**49.8**	48.4	44.4	**52.1**	50.2	51.3	48.9
	SFT + PRES.	55.5	51.9	50.1	49.1	50.5	49.5	48.2	**44.9**	**52.1**	**50.3**	51.3	**49.1**
	BISFT + PRES.	55.3	**52.0**	**50.3**	**49.4**	50.5	49.6	48.3	44.7	52.0	50.1	51.3	49.1
	SPELL-OUT	55.3	51.8	49.9	49.2	50.5	49.4	48.4	44.9	51.8	50.2	51.2	49.1
Qwen3-4B	VANILLA	65.1	55.8	**57.6**	52.9	55.2	49.0	46.5	42.9	58.6	54.8	56.6	51.1
	MLC	65.0	56.4	57.3	53.2	55.3	49.2	46.4	**43.7**	58.7	55.1	56.5	51.5
	SFT	**65.7**	57.0	57.1	53.8	55.3	49.3	45.7	43.3	58.2	**56.1**	56.4	**51.9**
	SFT + PRES.	64.8	**57.1**	57.4	52.9	**55.8**	49.4	46.0	43.2	57.2	55.0	56.2	51.5
	BISFT + PRES.	64.4	56.7	57.3	53.0	55.5	**49.5**	46.4	**43.7**	58.2	55.0	56.3	51.6
	SPELL-OUT	64.6	56.2	**57.6**	**53.9**	55.4	49.4	**46.7**	43.3	**58.9**	55.1	**56.6**	51.6
Apertus-8B	VANILLA	57.6	51.6	48.8	46.8	**51.7**	48.0	**46.3**	44.7	54.3	51.0	51.7	48.4
	MLC	57.5	51.2	47.1	46.2	51.1	47.9	45.0	44.6	53.9	50.2	50.9	48.0
	SFT	57.3	50.6	47.0	46.3	51.6	47.2	46.1	44.2	53.3	50.3	51.0	47.7
	SFT + PRES.	**57.7**	51.6	48.5	46.8	51.5	47.6	45.9	44.2	53.8	50.8	51.5	48.2
	BISFT + PRES.	57.5	**51.8**	47.7	46.2	50.9	48.1	44.5	43.9	54.2	51.2	50.9	48.2
	SPELL-OUT	**57.7**	51.7	**49.0**	**47.4**	51.5	**48.5**	46.2	**45.0**	**54.9**	**51.3**	**51.8**	**48.8**

Table 3: **Source- and target-language performance across models and language pairs.** For each model and language pair, *Src.* and *Tgt.* report performance on the source and target languages, respectively. The final two columns report averages across the five language pairs for each model and method. Values are displayed to one decimal place by truncation.

Analysis	Model	Depth .1	Depth .3	Depth .5	Depth .7	Depth .9
Relational alignment	TranslateGemma	0/3	0/3	0/3	0/3	3/3
	Qwen3-4B	0/3	0/3	0/3	2/3	1/3
Source-specific preservation	TranslateGemma	3/3	3/3	3/3	3/3	3/3
	Qwen3-4B	3/3	3/3	3/3	3/3	3/3
Shared-source preservation	TranslateGemma	3/3	3/3	3/3	3/3	3/3
	Qwen3-4B	3/3	3/3	3/3	3/3	3/3
Shared-target preservation	TranslateGemma	3/3	3/3	3/3	3/3	3/3
	Qwen3-4B	3/3	3/3	3/3	3/3	3/3

Table 4: **Depth-wise consistency of representation-level effects.** Each entry reports the number of language pairs, out of three (HI→BN, ES→RO, and ID→MS), for which the corresponding effect satisfies $p < 0.001$ at each relative depth. Cell colors additionally summarize cross-pair significance: dark green indicates that all three pairs satisfy $p < 0.001$; light green indicates that all three satisfy $p < 0.05$ but not all satisfy $p < 0.001$; yellow indicates that exactly two pairs satisfy $p < 0.05$; and red indicates that at most one pair satisfies $p < 0.05$. Preservation effects are consistently supported across models and depths, whereas relational alignment is concentrated in later layers.

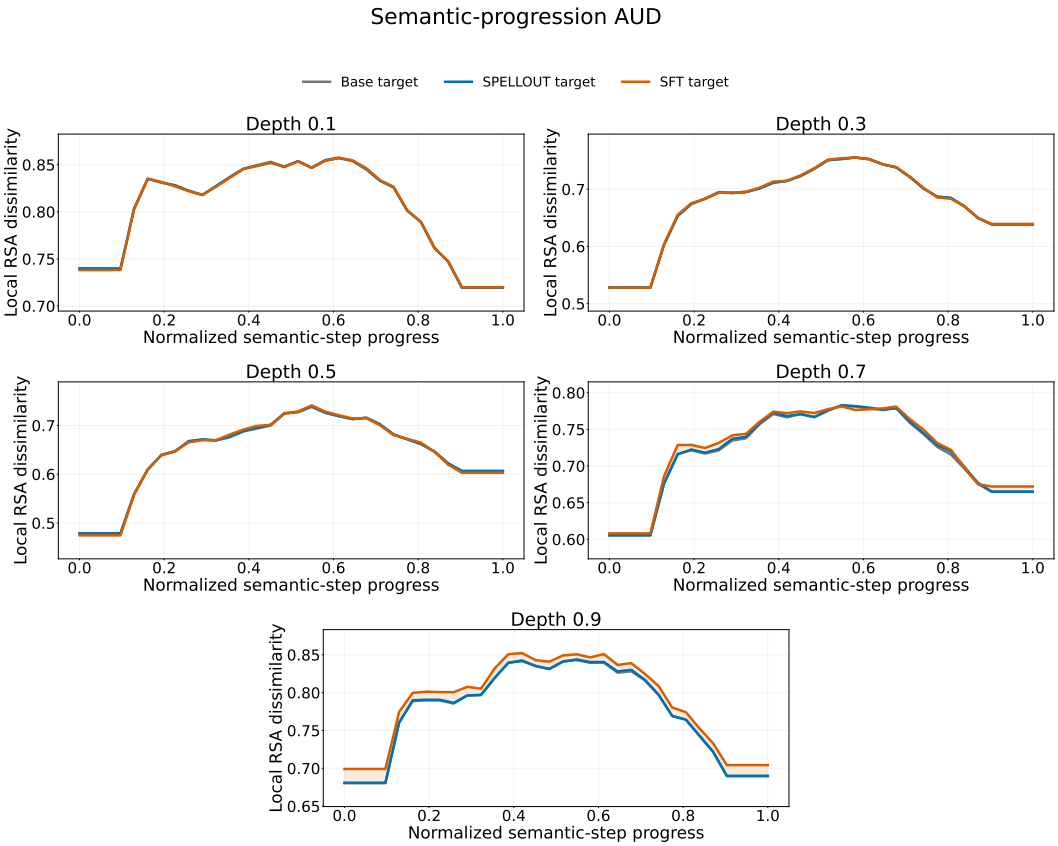


Figure 5: **Semantic-progression AUD across depth.** At each relative depth, curves show progress-local RSA dissimilarity ($1 - \rho_{\text{Spearman}}$) between the residualized source-pre trajectory geometry and the target trajectory geometry before and after adaptation. The reported ΔAUD summarizes the change from the pretrained target trajectory to the post-training trajectory, with positive values indicating closer agreement with the source-side progression. Each panel uses an independently scaled y-axis to expose small within-depth differences; absolute vertical positions should therefore not be compared across depths.

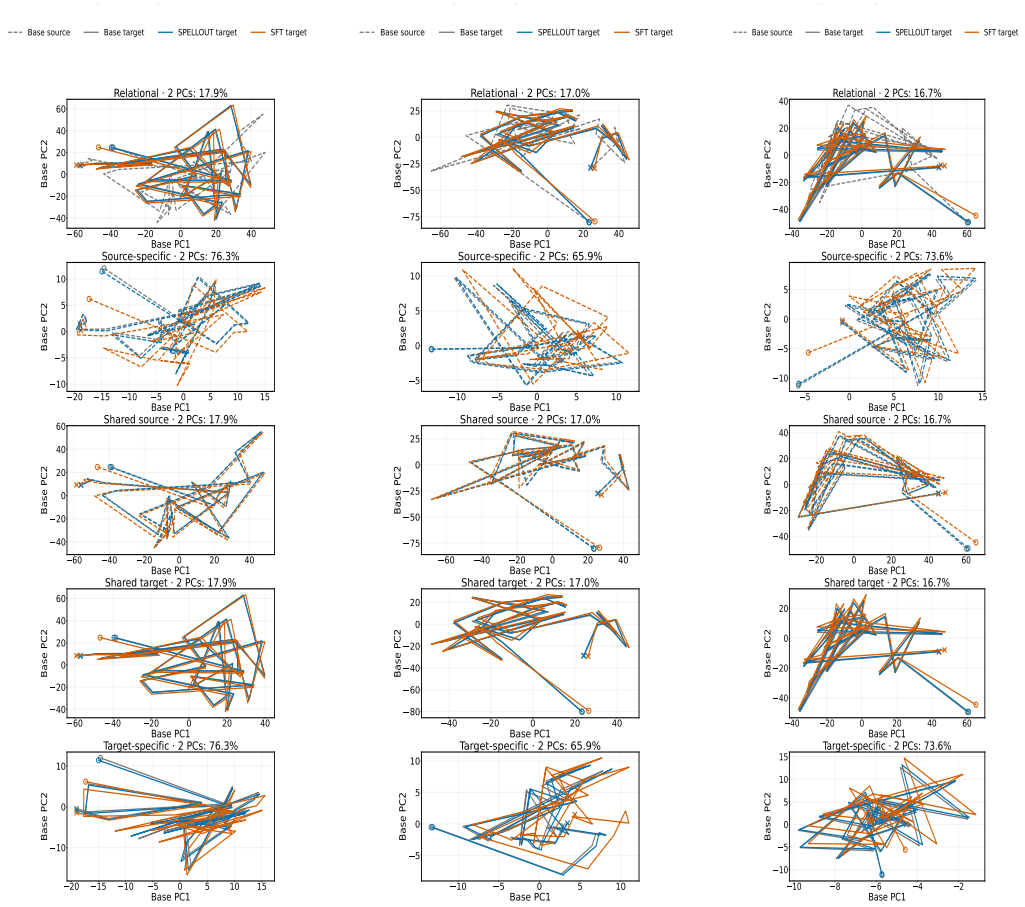


Figure 6: **Representation trajectories of Qwen3-4B.** From left to right, BN, MS, and RO. Depth is fixed at 0.7.

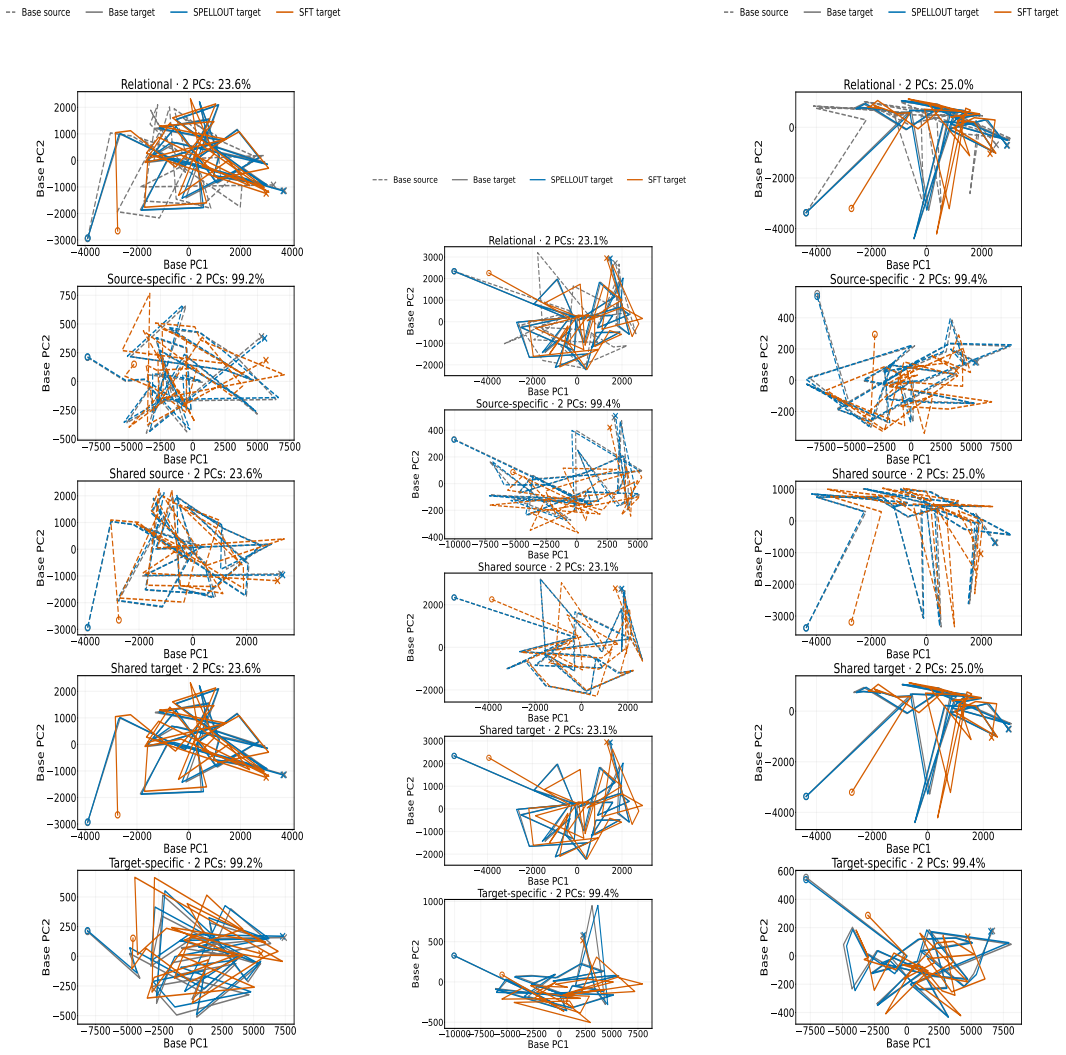


Figure 7: **Representation trajectories of Translategemma.** From left to right, BN, MS, and RO. Depth is fixed at 0.7.