Multilingual Steering by Design: Multilingual Sparse Autoencoders and Principled Layer Selection

Yusser Al Ghussin[1,2] **Daniil Gurgurov**[1,2] **Tanja Bäumel**[1,2,5]
Josef van Genabith[1,2] **Patrick Schramowski**[2,3,4,5] **Simon Ostermann**[1,2,5]

[1] Saarland University [2] German Research Center for Artificial Intelligence (DFKI)
[3] TU Darmstadt [4] hessian.AI [5] Centre for European Research in Trusted AI (CERTAIN)
yusser.al_ghussin@dfki.de

Abstract

Sparse autoencoders (SAEs) enable feature-level mechanistic interpretability and activation steering in large language models (LLMs), but SAE-based language control remains unreliable in multilingual settings: most SAEs are trained on English-only data, and steering layers are chosen heuristically. We address these limitations by advancing a principled, mechanistic account of multilingual language steering with SAEs. First, we show that training SAEs on multilingual data consistently strengthens cross-lingual representations and yields more reliable, quality-preserving language control across layers and model families. Second, we introduce an *a priori* steering layer-selection rule based on the intersection of multilingual alignment and language separability, which predicts effective intervention depths without exhaustive layerwise search. We evaluate our approach on LLaMA-3.1-8B and Gemma-2-9B across machine translation and cross-lingual summarization (CrossSumm), using SpBLEU, ROUGE-L, COMET, and LaSE. Our results show that multilingual SAEs combined with intersection-selected layers stabilize the trade-off between language identification accuracy and generation quality, providing a principled, predictive, representation-level account of multilingual SAE steering. We release all code and models for reproducibility. [1] [2]

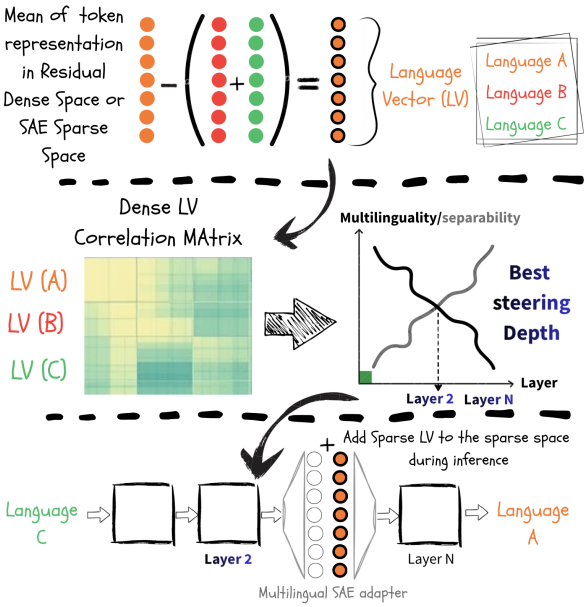

Figure 1: Overview of our language-control pipeline. A language-specific vector is constructed and used for layer selection and generation steering.

1 Introduction

Large language models (LLMs) can generate text in many languages, yet reliably *controlling* the output language remains challenging. While sparse autoencoders (SAEs) have emerged as a promising tool for interpreting internal activations and constructing steering vectors that causally influence model behavior (Cunningham et al., 2023; Templeton, 2024), SAE-based language steering in multilingual settings remains brittle and difficult to reproduce, with steering success varying unpredictably across models and layers: intervention depths are typically chosen heuristically (e.g., "mid-to-late" layers), requiring expensive layer sweeps and yielding inconsistent outcomes (Bayat et al., 2025; Chou et al., 2025). As a result, although SAE steering can work, it lacks a predictive, mechanistic account of *where* and *why* language control should be applied inside the model (Tang et al., 2024; Deng et al., 2025).

We argue that this haphazardness stems from the lack of a mechanistic perspective on how multilingual information is organized across model depth. We show that effective language steering requires access to two complementary signals: shared cross-lingual structure that supports fluent generation

[1] https://github.com/Yusser96/Multilingual-Steering-by-Design/
[2] https://huggingface.co/collections/Yusser/multilingual-steering-by-design

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 364–401
July 4, 2026 ©2026 Association for Computational Linguistics

across languages, and language-specific information that distinguishes one language from another. Prior work has shown that multilingual pretrained models learn shared latent representations across languages, facilitating cross-lingual transfer even in the absence of shared vocabularies or parallel data (Conneau et al., 2020). At the same time, language identity and language-specific features are differentially encoded across layers and can transition toward shared abstractions over depth in multilingual models (Riemenschneider and Frank, 2025; Zhang et al., 2025). If an intervention targets layers dominated by shared structure, steering lacks specificity; if it targets layers dominated by language-specific signals, the model often fails to recover generation quality. Our hypothesis reframes language steering as a problem of identifying representational balance points, rather than amplifying language-specific features in isolation, as is common in prior work (Tang et al., 2024; Deng et al., 2025; Gurgurov et al., 2025).

In this work, we operationalize this mechanistic hypothesis through two complementary contributions. First, we train SAEs directly on multilingual data for LLaMA-3.1-8B (Grattafiori et al., 2024) and Gemma-2-9B (Team et al., 2024), showing that multilingual training preserves the shared cross-lingual structure and language-specific distinctions required for predictable and interpretable steering in the sparse representation space. Compared to open-source SAEs (He et al., 2024; Lieberum et al., 2024), these multilingual SAEs yield more stable and quality-preserving language steering across layers and model families. Second, we introduce a principled, *a priori* rule for selecting steering layers based on the intersection of multilingual alignment and language separability, which predicts effective intervention depths without exhaustive layerwise search. Figure 1 provides an overview of the proposed language steering framework.

We validate this mechanistic framework across machine translation and cross-lingual summarization on LLaMA-3.1-8B and Gemma-2-9B, explicitly testing the prediction that balanced layers yield optimal language identification accuracy and generation quality trade-offs. Across both benchmarks, we find that multilingual SAEs combined with intersection-selected layers consistently stabilize language control and improve interpretability, supporting the view that effective steering depth is a property of the model's internal multilingual organization rather than a heuristic tuning choice.

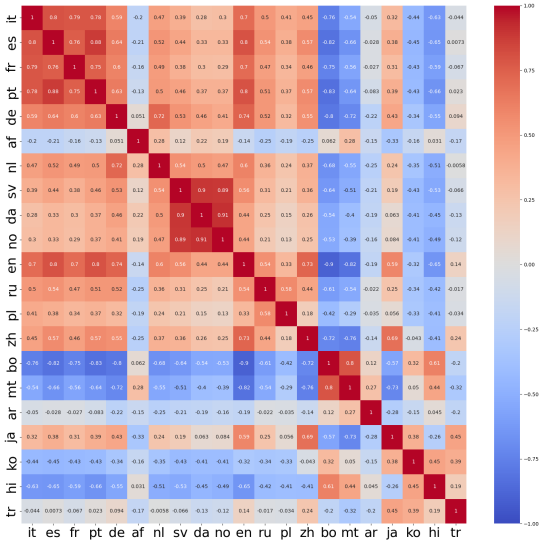

Figure 2: Correlation matrices of per-language contrast (DiffMean) vectors for Gemma-2-9B (Layer 23).

Our contributions are threefold:

- **Mechanistic characterization of language across depth.** We show that effective language steering arises at layers where cross-lingual alignment and language separability coexist.
- **Principled, *a priori* layer selection.** We introduce an intersection-based criterion that predicts effective steering depths without layer sweeps.
- **Multilingual SAEs as an interpretability enabler for language steering.** We show that multilingual SAE training preserves the representational structure required for reliable, interpretable language control.

2 Related Work

SAE-Based Activation and Language Steering. Sparse autoencoders (SAEs) have been widely used to interpret and steer internal activations in large language models (Templeton, 2024; Zhao et al., 2024; O'Brien et al., 2024; Wang et al., 2025; Zhao et al., 2026). Methods such as Sparse Activation Steering (SAS) (Bayat et al., 2025), Feature-Guided Activation Addition (FGAA) (Soo et al., 2025), and SAE-Targeted Steering (SAE-TS) (Chalnev et al., 2024) demonstrate that manipulating small sets of sparse features can causally influence model behavior. Applied to language control, prior work shows that editing individual SAE features can flip output language in models

such as Gemma-2-9B and LLaMA-3.1-8B (Chou et al., 2025; Deng et al., 2025; Gurgurov et al., 2026). However, effective steering depths are typically identified through manual exploration or fixed heuristics (e.g., mid-to-late layers), and many existing approaches rely on SAEs trained predominantly on English data. As a result, these methods do not provide a predictive, mechanistic account of where language steering should be applied across depth, nor how multilingual structure is preserved in sparse representations.

Evaluating and Training SAEs. Recent benchmarks such as SAE-Bench (Karvonen et al., 2025) and AxBench (Wu et al., 2025) evaluate SAE fidelity, interpretability, and intervention quality, reporting mixed results for SAE-based steering compared to simpler baselines. Other work emphasizes reconstruction fidelity as critical for causal interventions: Gemma-Scope (Lieberum et al., 2024) and LLaMA-Scope (He et al., 2024) report that high reconstruction error degrades steering effectiveness, while JumpReLU SAEs (Rajamanoharan et al., 2024) improve the fidelity–sparsity tradeoff via straight-through training. These findings suggest that insufficient SAE fidelity may disproportionately affect low-frequency or multilingual features, motivating our use of high-fidelity JumpReLU SAEs for multilingual language steering.

Language Features inside Models. Beyond SAEs, prior analyses point to strong layer-dependent language signals in multilingual models. Tang et al. (2024) identify language-specific neurons in BLOOM and LLaMA-2 and show that toggling them can switch the output language. Chang et al. (2022) study multilingual geometry in XLM-R, finding that languages occupy approximately parallel subspaces separated by linear "language vectors" particularly in middle layers; shifting hidden states along these directions flips predictions. Our findings echo these trends in the depth-wise distribution of multilingual structure and support treating language as a steerable direction in representation space (Gurgurov et al., 2026), while further revealing correlations among language families (Gurgurov et al., 2025).

Together, these lines of work motivate the need for a representation-level account of multilingual language steering that explains both how language information is organized across depth and how this organization can be exploited to guide interventions predictively.

3 Language Representations and Principled Steering

Our goal is not merely to improve language control, but to explain *where* and *why* language steering is possible inside multilingual LLMs, and to use this explanation to guide interventions *a priori*.

We define **language vectors** as directions in representation space that capture both the presence of individual languages and the directions along which they can be causally steered, building on prior evidence that language identity is linearly encoded as a direction or low-dimensional subspace within model representations (Park et al., 2024; Deng et al., 2025). Our layer-selection criterion is motivated by the observation that reliable language control requires access to two complementary signals: (i) *alignment*, corresponding to shared cross-lingual structure that supports generation across languages, and (ii) *separability*, corresponding to language-specific information that distinguishes one language from another. Only at depths where these signals are balanced can a small intervention reliably steer the output language.

3.1 Language Vectors

At each layer, we represent languages using contrastive *language vectors* constructed from model activations, either in the dense residual stream or in the sparse space induced by an SAE. Given activations from a target language and a set of other languages, we construct language steering vectors using the DiffMean method (Wu et al., 2025). For a given target language at layer ℓ, let $\mathcal{Z}^+$ denote the set of sparse codes corresponding to examples in the target language, and $\mathcal{Z}^-$ the set corresponding to all other languages. We compute the mean sparse representations by averaging SAE codes over all non-special tokens from all examples in that language

$$\bar{z}_\ell^+ = \frac{1}{|\mathcal{Z}^+|} \sum_{z \in \mathcal{Z}^+} z, \qquad \bar{z}_\ell^- = \frac{1}{|\mathcal{Z}^-|} \sum_{z \in \mathcal{Z}^-} z,$$

and define the steering vector as

$$w_{\text{DiffMean}}(\ell) = \bar{z}_\ell^+ - \bar{z}_\ell^- .$$

These vectors are then used additively in the SAE space to influence model outputs. Full mathematical definitions of the SAE representations,

DiffMean steering vectors, and the inference-time steering procedure are provided in Appendix C.

Beyond serving as steering directions, these language vectors exhibit meaningful linguistic structure. In particular, at the layers selected by our intersection-based criterion, pairwise correlations between per-language vectors reveal clear language-family groupings. As shown in Figure 2, languages from the same family (e.g., Romance or Germanic) exhibit high mutual similarity, while cross-family correlations remain lower. At the same time, a shared multilingual component persists across families, reflecting common cross-lingual structure. This coexistence of shared alignment and family-specific separation aligns with the intuition behind our layer-selection criterion and helps explain why these depths yield strong trade-offs between language identification accuracy and generation quality.

3.2 Multilingual SAEs for Language Steering

A central design choice in our framework is to train sparse autoencoders on multilingual data rather than English-only corpora. This choice is not merely pragmatic, but mechanistically important for reliable and interpretable language steering.

English-only SAEs preferentially encode monolingual structure: features that are frequent and salient in English dominate the sparse representation, while cross-lingual correlations and low-frequency language-specific features are weakly represented or collapse entirely. As a result, steering directions constructed from such representations are brittle. Language vectors may activate English-correlated features without cleanly isolating the intended target language, and the relationship between steering depth and downstream behavior becomes unstable. In contrast, multilingual SAE training exposes the autoencoder to systematic variation across languages, encouraging the sparse feature space to preserve both shared cross-lingual structure and language-specific distinctions.

From this perspective, multilingual SAEs can act as an *interpretability enabler* for representation-level language steering. They maintain the representational structure required to construct steering vectors whose effects can be predicted from representation-level statistics. The experimental comparisons in later sections empirically validate this claim, but the motivation for multilingual training arises directly from the mechanistic requirements of language steering.

3.3 Principled Layer Selection

A common assumption in prior work is that effective language control primarily relies on manipulating strongly language-specific features, which are believed to emerge in later layers, where the model is less able to recover from an intervention (Tang et al., 2024; Gurgurov et al., 2025). We find that effective steering emerges at depths where language-specific signals coexist with sufficient shared cross-lingual structure, motivating an *a priori* layer-selection strategy based on the depthwise evolution of language representations. At each layer, we quantify *multilinguality* as the degree to which language vectors share a dominant common direction, and *separability* as the extent to which languages remain distinct in representation space.

Let $\{\lambda_j\}_{j=1}^{N}$ be the eigenvalues of the language vectors pairwise Pearson correlation matrix C_ℓ; N is the number of languages. We define the *multilinguality* score as the explained-variance ratio of the first principal component,

$$f_\ell = \frac{\max_j \lambda_j}{\sum_{k=1}^{N} \lambda_k},$$

which measures the degree of shared alignment across languages. We define *separability* as the complementary quantity

$$s_\ell = 1 - f_\ell,$$

which reflects how distinct the language representations remain. We select steering layers at intersection points where these two signals are balanced, corresponding to depths that jointly preserve shared semantic structure while exposing discriminative language information.

We empirically validate this criterion across models, SAE variants, and tasks. Our contribution is to replace such heuristic choices with a principled, data-driven criterion that predicts these depths *before* training SAEs and steering experiments are run. Full definitions of the language correlation matrices, multilinguality and separability metrics, and the intersection-based layer-selection procedure are given in Appendix D.

4 Experiments

4.1 Models and Data

We evaluate on *LLaMA-3.1-8B* (Grattafiori et al., 2024) and *Gemma-2-9B* (Team et al., 2024) using 21 FLORES–200 languages (see Appendix B)

(Costa-Jussà et al., 2022). For each model, we train parallel English-only and multilingual JumpReLU SAE suites (Rajamanoharan et al., 2024) on 2.1B Wikipedia (Wikimedia Foundation, 2023) tokens with identical architectures and optimization settings, isolating the effect of multilingual training data. Full details are provided in Appendix A.

4.2 Evaluation

FLORES–200 Machine Translation. We evaluate language steering on machine translation using FLORES–200 (Costa-Jussà et al., 2022). We use the `dev` split to construct steering vectors and the `devtest` split for evaluation. Each `devtest` set contains approximately 1,000 sentences per language, providing a substantially large and clean evaluation set while ensuring strict separation between steering construction and evaluation.

For each non-English target language i in our language set ($|i| = 20$), we define an English→`tgt_`i translation task, where English (`eng_Latn`) is always the source language. We construct a per-language steering vector using `dev` sentences, and apply this vector to steer generation into the intended output language, which we denote as `steer_`i.[3]

Prompts are written *in the target language* using natural translation instructions (e.g., German: "*Übersetze diesen Satz:*"), followed by a target-language answer cue (e.g., "*Übersetzung:*"). We provide prompt examples in the Appendix F.

`Translate this sentence:` "`<source text>`". (`Translate this sentence:`: instruction in target language; `<source text>`: Always English)

`Translation:` (answer cue in target language)

This setup biases the model toward both the translation task and the prompt language, so that any deviation toward a steering language can be attributed to the steering intervention rather than prompt ambiguity. We decode using greedy search with temperature 0, yielding a conservative and interpretable baseline that isolates the effect of steering from prompt engineering or decoding strategies. We report relative differences across SAE variants relative to open-source SAE baselines, which directly measure the effectiveness of multilingual training and layer selection, for three metrics. (1) **LangID**, computed by applying a fastText language identification classifier (Joulin et al., 2016) to the generated outputs, measures how reliably steering enforces the intended output language. (2) **SpBLEU** (Post, 2018), computed against the reference translation in the intended *steer language*, provides a script-agnostic measure of surface-level translation quality. (3) **COMET** (Rei et al., 2020), a neural evaluation metric that leverages cross-lingual pretrained encoders and both the source and reference sentences, estimates semantic translation quality and correlates strongly with human judgments.

We report results averaged across all 20 non-English prompt languages, where for each `tgt_`i we evaluate steering into every other target language `steer_`j with $j \neq i$. Concretely, the model is prompted in language `tgt_`i and steered toward `steer_`j, and results are averaged over the full cross-product of (i, j) pairs. This aggregation directly measures how reliably different SAE variants enable control over the output language, independent of any single prompt–language pairing. As our primary focus is the *relative* (delta) performance differences between SAE variants rather than absolute per-language scores, we present these averages in the main text, while detailed per-language and per-pair results are provided in Appendix O.

In addition, we report a restricted setting where steering is applied only when `steer_`j = `tgt_`i, allowing us to analyze the behavior of steering when the prompt language and intended output language coincide (Appendix N).

Cross-Lingual Summarization (CrossSumm). To evaluate whether our findings generalize beyond translation, we use the cross-lingual summarization dataset CrossSum (Park et al., 2025). We select document–summary pairs whose target languages intersect with our translation language set. The resulting dataset consists of 108 fully parallel English source documents paired with reference summaries in one of five target languages: Spanish (es), Russian (ru), Arabic (ar), Hindi (hi), and Turkish (tr).

We follow the same experimental design as in machine translation, reusing the same per-language steering vectors. The only change is the prompt, which is written in the target language and phrased as a natural summarization instruction in the target-language (e.g., "Summarize the following article")

[3] We use `tgt_`i for the prompt language and `steer_`i for the intended output language after steering. In our setup, `steer_`i is the language for which we construct the steering vector. We use the term "steer language" to emphasize that the output language is controlled via a steering intervention.

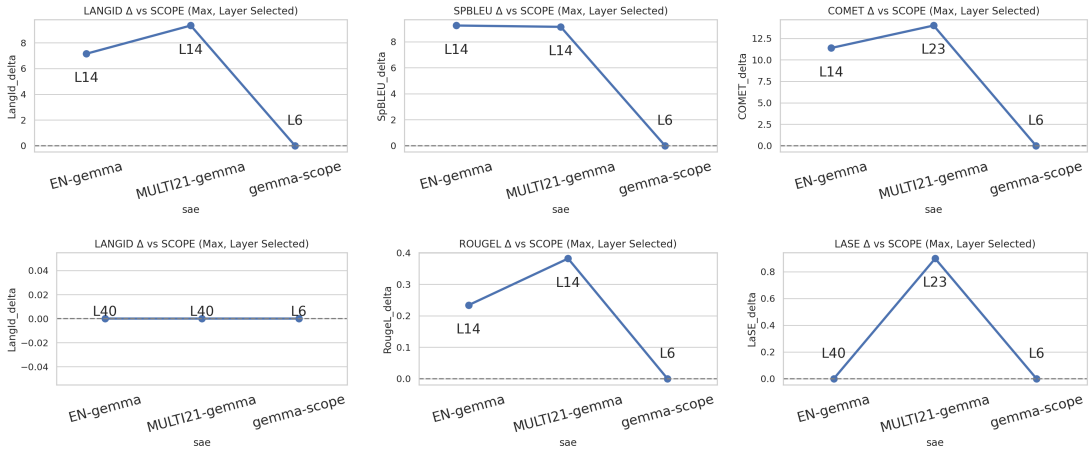

Figure 3: Performance deltas relative to Scope baselines for **Gemma-2-9B** at the best-performing steering layer. **Top:** FLORES machine translation (LangID, SpBLEU, COMET). **Bottom:** Cross-lingual summarization (LangID, ROUGE-L, LaSE).

with a target-language answer cue ("summary:"), thereby biasing the model toward both the summarization task and the target language. We provide prompt examples in the Appendix E.

We evaluate generated summaries using three metrics: **LangID**, **ROUGE-L** (Lin, 2004), and **LaSE** (Park et al., 2025), following the evaluation protocol of the original dataset. ROUGE-L measures content overlap with the reference summary, while LaSE evaluates cross-lingual semantic similarity between the generated and reference summaries. This setup allows us to test whether steering preserves semantic content while controlling the output language in a non-translation generative task.

5 Results

We assess steering performance and layer sensitivity, focusing on: (i) benefits from multilingual SAE training, (ii) optimal intervention layers, and (iii) comparisons with open-source SAEs.

5.1 Benefits of Multilingual Training for SAEs

We study the effect of training data on SAE steering, comparing monolingual (English-only) SAEs with multilingual SAEs across transformer layers. We evaluate multilingual steering across two generation tasks: (i) machine translation and (ii) cross-lingual summarization, focusing on language identification accuracy, surface quality, and semantic preservation.

Multilingual training improves steering. Figures 3 in the paper and 16 in the appendix summarize performance deltas relative to open-source Scope baselines at the best-performing steering layers (i.e., the layers with the overall highest performance), across both machine translation (FLORES) and cross-lingual summarization (CrossSumm). For Gemma-2-9B (Figure 3), multilingual SAEs outperform English-only SAEs across all reported metrics, yielding substantial gains in generation quality for both tasks. In FLORES, multilingual training improves LangID and COMET while maintaining strong SpBLEU; in CrossSumm, it yields higher ROUGE-L and LaSE, indicating better content preservation and semantic alignment. For LLaMA-3.1-8B (Figure 16), the improvements are smaller in magnitude but remain directionally consistent across tasks and SpBLEU, COMET and LaSE metrics while maintaining competitive LangID. Overall, these results demonstrate that multilingual SAEs induce more effective and semantically aligned steering directions, with consistent benefits across models and task families.

5.2 Optimal Layers

For each model, we identify steering layers as *intersection points* where multilingual alignment and language separability are jointly balanced (Figure 4). Importantly, the multilinguality–separability curves are computed independently of any downstream generation metrics, making the predicted intersection layers a falsifiable, pre-intervention hypothesis.

For Gemma-2-9B, these curves exhibit a characteristic *two-hump* shape, yielding intersection regions near **L14** and **L23**. Figure 3 shows that these

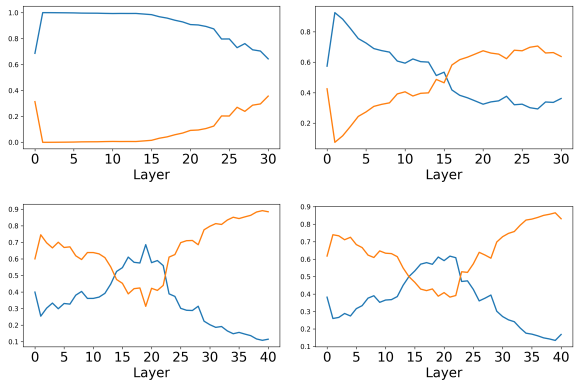

Figure 4: Layer-selection curves showing the balance between multilingual alignment and language separability across layers. **Blue** curves denote *multilinguality* (shared cross-lingual alignment), and **orange** curves denote *separability* (language-specific structure). **Top:** LLaMA-3.1-8B. **Bottom:** Gemma-2-9B. **Left:** Open-source SAEs (LLaMA-Scope, Gemma-Scope), where LLaMA-Scope shows no clear intersection and Gemma-Scope selects L14 and L23. **Right:** Residual representations, which exhibit clear balance points at L15 (LLaMA) and L14/L23 (Gemma).

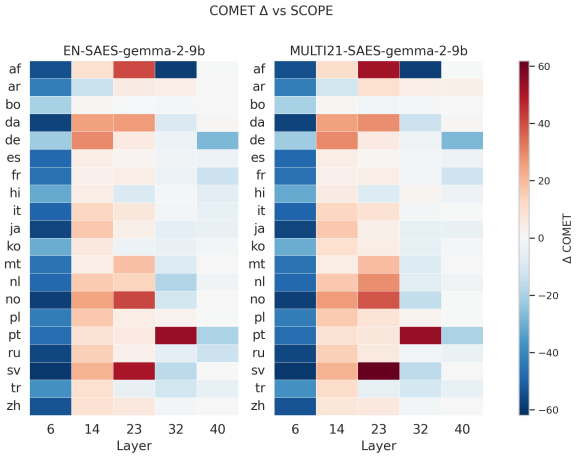

Figure 5: Per-language, per-layer COMET score deltas for **Gemma-2-9B** on FLORES under cross-lingual steering (`tgt_`$i \neq$ `steer_`j).

same layers achieve the strongest overall trade-offs between language identification accuracy and generation quality for both multilingual and English-only SAEs. Figure 5 further confirms this pattern at the per-language level, where **L14** and **L23** consistently emerge as the best-performing steering depths.

In LLaMA-3.1-8B, a pronounced increase in multilinguality near **L13** is followed by a rise in separability, yielding an intersection region spanning **L13–L15**. Figure 16 (Appendix) reports the best-performing layers across SpBLEU, COMET, LaSE, and LangID, and shows that steering within this intersection region achieves the strongest overall trade-offs. In particular, **L13** and **L15** consistently emerge as the empirically optimal steering depths across most metrics.

To further validate our layer-selection criterion, we analyze layerwise performance trends under different steering regimes. Figure 6 reports layerwise ΔCOMET and ΔLaSE averaged across SAE variants for two settings: (i) when the steering language matches the prompted target language, ΔCOMET and ΔLaSE increase monotonically with depth. This behavior is consistent with later layers exhibiting stronger language separability in LLaMA-3.1-8B and favoring same-language amplification. (ii) when steering toward a language different from the prompted target, performance follows a non-monotonic trend, peaking near the layers identified by our multilinguality–separability intersection.

This divergence highlights the role of representational balance: deeper layers benefit same-language reinforcement, whereas effective cross-language steering requires intervening at depths where shared cross-lingual structure is still preserved alongside language-specific distinctions. These results provide additional empirical support that the layers selected by our criterion correspond to optimal steering depths. Importantly, we show that layers selected by our criterion consistently outperform earlier and later layers when controlling for SAE architecture and training data. This indicates that effective steering depth is a structural property of the base model rather than an artifact of a particular SAE. We observe a different pattern in Gemma-2-9B: same-language steering favors earlier layers, where language separability is high, while cross-language steering again peaks near the layers identified by our intersection criterion.

5.3 Additional Experiments: Steering with Open-Source SAEs

We compare our multilingual SAEs against the open-source *LLaMA-Scope* and *Gemma-Scope* suites. For *Gemma-Scope*, our criterion identifies two intersection layers at **L14** and **L23**. Steering performance at these depths remains consistently below that of our multilingual SAEs (Figure 5). At other layers, Gemma-Scope can exhibit competitive or occasionally stronger results, highlighting the sensitivity of multilingual steering to intervention depth and suggesting that multilingual SAEs,

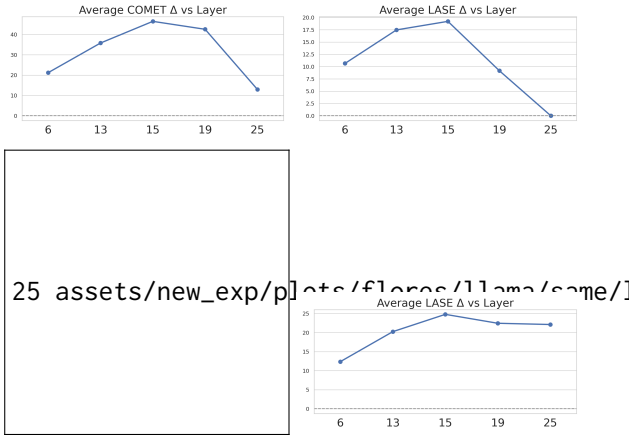

Figure 6: Layerwise ΔCOMET and ΔLaSE trends for LLaMA-3.1-8B averaged across SAEs under two steering regimes. **Top:** `steer_lang` ≠ `target_lang`. **Bottom:** `steer_lang` = `target_lang`.

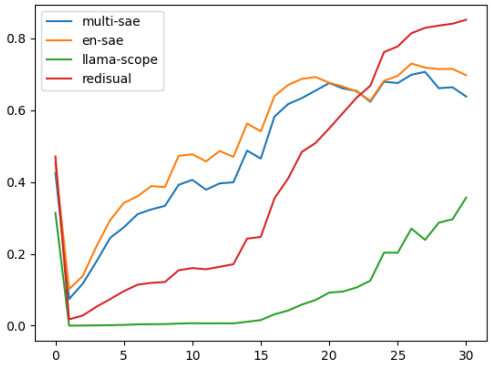

Figure 7: Separability across layers for LLaMA-3.1-8B, comparing different SAEs.

even when trained on comparatively less multilingual data, can surpass Gemma-Scope when applied at mechanistically appropriate layers.

In contrast, for *LLaMA-Scope*, we observe consistently negligible downstream gains across layers. Applying our multilinguality-separability analysis in the sparse space reveals that LLaMA-Scope does not exhibit a meaningful intersection layer: language separability remains weak across depth (Figures 4 and 20). The absence of a balance point between shared cross-lingual alignment and language-specific structure aligns with its poor steering performance.

Figure 7 further reproduces the early–late dynamics of multilingual representations previously reported for LLaMA-3.1-8B (Gurgurov et al., 2025; Tan et al., 2024): shared cross-lingual structure is strongest in early-to-mid layers, while language separability increases toward later depths. Notably, LLaMA-Scope exhibits substantially lower separability than even the dense residual stream across all layers, which likely reflects the combined effects of English-skewed training data and architectural choices in the SAE design, and helps explain its failure to support effective language control despite operating at similar depths. More generally, when the separability score approaches zero, we consistently observe steering failure, indicating that separability provides a simple and predictive signal of multilingual steering capability at a specific layer (Figure 20 in appendix).

6 Take-Aways and Conclusion

Our results show that reliable SAE-based multilingual steering emerges from the combination of multilingual training and principled layer selection. Crucially, our findings show that multilingual SAE steering is promising, and that its success can be predicted from representation-level structure.

(I) Multilingual SAE training strengthens language representations. Across both models, multilingual SAEs consistently outperform English-only SAEs on language identification accuracy and generation quality. These gains indicate that multilingual training does more than expand language coverage: it induces richer shared cross-lingual structure while preserving cleaner language-specific signals in the sparse feature space, yielding more reliable steering directions.

(II) Intersection points predict optimal steering depths. Balancing multilingual alignment and language separability identifies layers where language control and generation quality are jointly maximized. The intersection of these signals provides an *a priori* rule for layer selection that replaces heuristic mid–late choices and avoids exhaustive layer sweeps and repeated SAE training. Across both base models, the layers identified by this criterion consistently coincide with those yielding the strongest LangID–quality trade-offs, and outperform earlier and later layers even when controlling for SAE architecture and training data.

(III) Open-source SAEs highlight the limits of heuristic depth choices. Open-source SAEs provide useful baselines but illustrate the importance

of principled layer selection and multilingual training. *LLaMA-Scope* does not exhibit a clear intersection between multilinguality and separability (Figure 4) and yields negligible steering gains across layers. Its sparse representations show weak language separability, often worse than the dense residual stream, suggesting that English-skewed training data and architectural choices collapse multilingual features (Figure 7).

Limitations

We evaluated two base models (LLaMA-3.1-8B and Gemma-2-9B); larger, instruction-tuned, or decoder–encoder architectures may exhibit different cross-lingual dynamics. Our evaluation focuses on automated metrics (LangID, SpBLEU, COMET, ROUGE-L, LaSE), which do not capture stylistic fidelity, code-switching behavior or robustness to ambiguous prompts. Additionally, our findings are based on JumpReLU SAEs trained on the residual stream; extending this analysis to other sparse architectures, intervention sites (e.g., attention or MLP activations), or alternative steering constructions remains an open direction. We do not claim that the intersection criterion is unique; alternative representational statistics may identify similar balance points. Future work should complement these automated evaluations with manual translation-error analysis and stronger comparisons to existing steering methods and state-of-the-art multilingual systems, in order to better characterize failure modes and clarify the remaining performance gap for SAE-based language control. Similarly, the 0.5 intersection threshold should be understood as an operational definition of equal multilingual alignment and language separability rather than as a uniquely optimal cutoff; future work should study adaptive or model-specific thresholds.

Acknowledgments

This research was supported by the German Federal Ministry for Economic Affairs and Energy (BMWE) as part of the project *"Souveräne KI für Europa (SOOFI)"* (13IPC040H), and by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005).

References

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. 2025. Steering large language model activations in sparse spaces. In *Proceedings of the Conference on Language Modelling (COLM)*.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. Improving steering vectors by targeting sparse autoencoder features. *arXiv preprint arXiv:2411.02193*.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Cheng-Ting Chou, George Liu, Jessica Sun, Cole Blondin, Kevin Zhu, Vasu Sharma, and Sean O'Brien. 2025. Causal language control in multilingual transformers via sparse feature steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Boyi Deng, Yu Wan, Yidan Zhang, Baosong Yang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. *Preprint*, arXiv:2505.05111.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daniil Gurgurov, Yusser Al Ghussin, Tanja Baeumel, Cheng-Ting Chou, Patrick Schramowski, Marius Mosbach, Josef van Genabith, and Simon Ostermann. 2026. Clas-bench: A cross-lingual alignment and steering benchmark. *Preprint*, arXiv:2601.08331.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. 2025. Language arithmetics: Towards systematic language neuron identification and manipulation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th*

Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, and 1 others. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart Mcdougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. 2025. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 29223–29264. PMLR.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. 2024. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*.

Gyutae Park, Jeonghyun Park, and Hwanhee Lee. 2025. Cross-lingual summarization for low-resource languages using multilingual retrieval-based in-context learning. *Applied Sciences*, 15(14).

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders, 2024b. *URL https://arxiv.org/abs/2407.14435*.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2025. Cross-lingual generalization and compression: From language-specific to shared neurons. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13470–13491, Vienna, Austria. Association for Computational Linguistics.

Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. 2025. Interpretable steering of large language models with feature guided activation additions. *arXiv preprint arXiv:2501.09929*.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Adly Templeton. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Anthropic.

Anyi Wang, Xuansheng Wu, Dong Shu, Yunpu Ma, and Ninghao Liu. 2025. Enhancing llm steering through sparse autoencoder-based vector refinement. *arXiv preprint arXiv:2509.23799*.

Wikimedia Foundation. 2023. Wikipedia dump, november 1, 2023. `https://dumps.wikimedia.org/`. Accessed: 2025-10-06.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*.

Haiyan Zhao, Xuansheng Wu, Fan Yang, Bo Shen, Ninghao Liu, and Mengnan Du. 2026. Denoising concept vectors with sparse autoencoders for improved language model steering. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 797–808, Rabat, Morocco. Association for Computational Linguistics.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2024. Steering knowledge selection behaviours in llms via sae-based representation engineering. *arXiv preprint arXiv:2410.15999*.

Appendix

A SAE Training

A.1 SAE Training Data

Following the mechanistic motivation outlined in Section 3.2, we train parallel English-only and multilingual SAE suites under a controlled setup to isolate the effect of training data on language steering. We train two SAE suites per base model using Wikipedia data (Wikimedia Foundation, 2023). For the multilingual suite (*MULTI21-SAEs*), we construct a balanced corpus covering the same 21 languages, and select a total of 2.1B tokens with a uniform distribution across languages. For the English-only suite (*EN-SAEs*), we select the same number of tokens (2.1B), drawn from English Wikipedia. This controlled setup ensures that both suites are trained on identical data volume with identical optimization parameters, isolating the effect of multilingual versus monolingual training data from corpus size or training duration.

A.2 SAE Training Procedure

For each layer of each base model, we train JumpReLU SAEs (Rajamanoharan et al., 2024) on the residual stream, matching the expansion factors used by the corresponding open-source SAE suites (8× for LLaMA-Scope and 16k for Gemma-Scope). We use identical architectures and optimization hyperparameters for our *EN-SAEs* and *MULTI21-SAEs*. To ensure a controlled comparison, we fix the number of optimization steps and therefore the total number of parameter updates across both suites. This setup cleanly isolates the impact of multilingual training from architectural and optimization confounds.

A.3 Hyperparameters

We train SAEs with `SAELens`[4] using a JumpReLU architecture on the residual stream at multiple layers. The base model is loaded in `float16`, while SAE training uses `float32`. Hook sites follow `blocks.{layer}.hook_resid_post`

Key hyperparameters (from code).

- **Architecture:** `jumprelu` (expansion factor $= 8$), L_1 coefficient $= 5.0$, JumpReLU bandwidth $= 10^{-3}$, init threshold $= 10^{-3}$, decoder init = zeros, transpose encoder init, decoder heuristic init enabled, sparsity penalty scaled by decoder norm.
- **Training length (#steps):** 30,000.
- **Batch size (tokens/step):** 4,096.
- **Context size:** 512.
- **Optimizer / schedule:** Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$), LR $= 5{\times}10^{-5}$, LR warmup $= 1{,}500$ steps, LR decay steps $= 3{,}000$, L_1 warmup $= 1{,}500$ steps.
- **Dead/feature refresh:** feature sampling window $= 2000$, dead-feature window $= 1000$, dead threshold $= 10^{-4}$.
- **Data loader:** streaming enabled, `prepend_bos=True`.

A.4 Computational Budget

For all experiments we used 1× H100 80GB. Each SAE model was trained for 30,000 optimization steps with a batch size of 4,096 tokens per step, corresponding to approximately 123M training tokens and about 3 GPU hours per run.

A.5 License and Availability

All trained SAE checkpoints produced in this work, including both English-only and multilingual variants for *LLaMA-3.1-8B* and *Gemma-2-9B*, are released under the **Apache License 2.0**. This license permits use, modification, and distribution of the models for both research and commercial purposes, provided that proper attribution is maintained.

We emphasize that the underlying base models (*LLaMA-3.1-8B*, *Gemma-2-9B*) remain subject to their original licenses as released by Meta and Google, respectively. Users of our checkpoints must therefore comply with both the Apache 2.0 license governing our SAEs and the terms of the corresponding base model licenses.

[4] https://github.com/jbloomAus/SAELens

B Language Labels

Language	Code
English	eng_Latn
Tibetan	bod_Tibt
Maltese	mlt_Latn
Italian	ita_Latn
Spanish	spa_Latn
German	deu_Latn
Japanese	jpn_Jpan
Arabic	arb_Arab
Chinese (Simplified)	zho_Hans
Afrikaans	afr_Latn
Dutch	nld_Latn
French	fra_Latn
Portuguese	por_Latn
Russian	rus_Cyrl
Korean	kor_Hang
Hindi	hin_Deva
Turkish	tur_Latn
Polish	pol_Latn
Swedish	swe_Latn
Danish	dan_Latn
Norwegian Bokmål	nob_Latn

Table 1: List of 21 target languages from FLORES–200 and their language codes.

C Formal Definitions of Language Vectors and Layer Selection

This appendix provides the full mathematical formulation of the language vectors, steering procedure, and layer-selection metrics summarized in the main text.

C.1 Representation Extraction

At each transformer layer ℓ, we extract the dense hidden representation from the residual stream,

$$h_\ell(x) \in \mathbb{R}^D,$$

for input x. To obtain a sparse and interpretable representation, we apply an encoder–decoder sparse autoencoder (SAE) trained at layer ℓ. The encoder maps dense activations to a high-dimensional sparse code,

$$z_\ell(x) = \mathrm{Encoder}_\ell(h_\ell(x)), z_\ell(x) \in \mathbb{R}^K,\ K \gg D$$

and the decoder reconstructs the activation,

$$\hat{h}_\ell(x) = \mathrm{Decoder}_\ell(z_\ell(x)).$$

Sparsity is enforced via the SAE objective, yielding sparse codes that isolate a small number of active features for each input.

C.2 DiffMean Steering Vectors

We construct language steering vectors using the DiffMean method (Wu et al., 2025). For a given target language at layer ℓ, let $\mathcal{Z}^+$ denote the set of sparse codes corresponding to examples in the target language, and $\mathcal{Z}^-$ the set corresponding to all other languages. We compute the mean sparse representations

$$\bar{z}_\ell^+ = \frac{1}{|\mathcal{Z}^+|} \sum_{z \in \mathcal{Z}^+} z, \qquad \bar{z}_\ell^- = \frac{1}{|\mathcal{Z}^-|} \sum_{z \in \mathcal{Z}^-} z,$$

and define the steering vector as

$$w_{\mathrm{DiffMean}}(\ell) = \bar{z}_\ell^+ - \bar{z}_\ell^-.$$

This vector amplifies features that are characteristic of the target language while suppressing features shared with other languages. Prior work applies DiffMean directly in the dense residual stream; in contrast, we primarily apply it in the SAE sparse space, which yields more disentangled and controllable steering directions.

C.3 Inference-Time Steering

Given a hidden activation $h_\ell(x)$ at inference time, we apply steering as follows:

1. Encode the activation into sparse space:

$$z_\ell(x) = \mathrm{Encoder}_\ell(h_\ell(x)).$$

2. Apply the steering vector:

$$z'_\ell(x) = z_\ell(x) + \alpha\, w_{\mathrm{DiffMean}}(\ell),$$

where α controls steering strength. We use fixed steering coefficients for all test examples within each model setting, with $\alpha = 5.0$ for LLaMA and $\alpha = 100.0$ for Gemma. These values were chosen in preliminary experiments as conservative values that improved target-language identification, and were fixed before final evaluation; they were not tuned per language, layer, or test example.

3. Decode back to dense space:

$$\hat{h}'_\ell(x) = \mathrm{Decoder}_\ell(z'_\ell(x)).$$

4. Correct for reconstruction error by adding the residual:

$$\tilde{h}_\ell(x) = \hat{h}'_\ell(x) + \big(h_\ell(x) - \mathrm{Decoder}_\ell(z_\ell(x))\big).$$

The corrected activation $\tilde{h}_\ell(x)$ is then passed to subsequent layers. This procedure preserves the original activation outside the SAE subspace while applying a targeted intervention along the language direction.

D Language Correlation and Intersection-Based Layer Selection

D.1 Per-Language Contrast Vectors

For each language i and layer ℓ, we construct a contrast vector using DiffMean. Let $\mathcal{H}_i^+$ denote dense codes from language i, and $\mathcal{H}_i^-$ dense codes from all other languages. The per-language vector is

$$\mathbf{v}_i = \frac{1}{|\mathcal{H}_i^+|} \sum_{h \in \mathcal{H}_i^+} h - \frac{1}{|\mathcal{H}_i^-|} \sum_{h \in \mathcal{H}_i^-} h.$$

These vectors represent languages in a shared feature space by emphasizing language-specific features and suppressing shared ones.

D.2 Correlation Matrix Across Languages

Given the set of language vectors $\{\mathbf{v}_i\}_{i=1}^{N}$ at layer ℓ, where N is the number of languages, we compute a pairwise Pearson correlation matrix

$$C_\ell \in \mathbb{R}^{N\times N}, \qquad C_{ij} = \mathrm{corr}(\mathbf{v}_i, \mathbf{v}_j).$$

This matrix captures how similarly different languages are represented at a given depth.

D.3 Multilinguality and Separability Metrics

Let $\{\lambda_j\}_{j=1}^{N}$ be the eigenvalues of C_ℓ. We define the *multilinguality* score as the explained-variance ratio of the first principal component,

$$f_\ell = \frac{\max_j \lambda_j}{\sum_{k=1}^{N} \lambda_k},$$

which measures the degree of shared alignment across languages. We define *separability* as the complementary quantity

$$s_\ell = 1 - f_\ell,$$

which reflects how distinct the language representations remain.

D.4 Intersection-Based Layer Selection

We select steering layers at depths where multilinguality and separability are balanced. Since $s_\ell = 1 - f_\ell$, an intersection occurs when $f_\ell \approx 0.5$, or equivalently when $2f_\ell - 1$ changes sign between adjacent layers. In practice, we detect these sign changes with a small tolerance and linearly interpolate between layer indices. These intersection points serve as *a priori* candidates for effective steering depths and consistently correspond to layers that yield strong language control while preserving generation quality.

E CrossSum Prompts

Prompt (Hindi, CrossSumm)

इस लेख का सारांश दें:
(English: "Summarize this article:")
``It is the right of every human being to choose their gender, ... [truncated] ... if they have consensual gay sex.''
सारांश:
(English: "Summary:")

Reference Summary (Spanish)

La Corte Suprema de India reconoció la existencia de un tercer género, además del femenino y el masculino.
(English: "India' s Supreme Court recognized the existence of a third gender, in addition to male and female.")

Model Output (Spanish, SAE-steered)

"Es el derecho de cada ser humano elegir su género", dijo la sentencia al otorgar derechos a quienes se identifican como hombres o mujeres. Ordenó al gobierno proveerles cuotas en los trabajos y educación con base en las minorías...
(English: "It is the right of every human being to choose their gender," the ruling said, granting rights to those who identify as men or women. It ordered the government to provide quotas in jobs and education...)

Figure 8: Example prompt and outputs for cross-lingual summarization (CrossSum). The model is prompted in Hindi and steered to generate a Spanish summary. English glosses are provided for clarity. Full article text omitted for readability.

Prompt (Spanish, CrossSumm)

Resume este artículo:
(English: "Summarize this article:")
``She said Britain needed certainty, stability and strong leadership following the EU referendum…[truncated] …''
Resumen:
(English: "Summary:")

Reference Summary (Arabic)

دعت رئيسة الوزراء البريطانية، تيريزا ماي، إلى إجراء انتخابات عامة مبكرة يوم 8 يونيو/حزيران المقبل.
(English: "British Prime Minister Theresa May called for early general elections on June 8.")

Model Output (Arabic, SAE-steered)

"الرئيسة البريطانية تيريزا ماي في خطابها أمس الأحد، أعلنت عن استدعاء الانتخابات العامة المقرر إجراؤها في المملكة المتحدة في شهر مايو المقبل، وذلك بعد أن حصلت على موافقة الملكة إليزابيث الثانية على هذا القرار..."
(English: "British Prime Minister Theresa May announced in a speech on Sunday the calling of early general elections in the UK next May, after receiving approval from Queen Elizabeth II...")

Figure 9: Example prompt and outputs for cross-lingual summarization (CrossSum). The model is prompted in Spanish and steered to generate an Arabic summary. English glosses are provided for clarity. Full article text omitted for readability.

F Flores Prompts

Prompt (Chinese, FLORES Translation)

请翻译这句话：
(English: "Please translate this sentence:")
``Siminoff said sales boosted after his 2013 appearance in a Shark Tank episode where the show panel declined funding the startup.''
翻译：
(English: "Translation:")

Reference Translation (Russian)

Симинофф сказал, что продажи выросли после его появления в выпуске шоу "Shark Tank" в 2013 году, где члены жюри отказались финансировать его стартап.
(English: "Siminoff said that sales increased after his appearance on a 2013 episode of the show Shark Tank, where the panel declined to fund his startup.")

Model Output (Russian, SAE-steered)

"Симинофф сказал, что продажи выросли после его появления в эпизоде шоу "Танк", где панель отказалась финансировать стартап."
(English: "Siminoff said that sales increased after his appearance in an episode of the show 'Tank', where the panel refused to fund the startup.")

Figure 10: Example prompt and outputs for machine translation. The model is prompted in Chinese and steered to generate a Russian translation. English glosses are provided for readability.

Prompt (German, FLORES Translation)

Übersetze diesen Satz:
(English: "Translate this sentence:")
``We now have 4-month-old mice that are non-diabetic that used to be diabetic, he added.''
Übersetzung:
(English: "Translation:")

Reference Translation (Arabic)

"أضاف قائلاً، لدينا الآن فئران تبلغ من العمر أربعة أشهر كانت تعاني في السابق من مرض السكري، لكنها لم تعد تعاني منه الآن."
(English: "He added, We now have four-month-old mice that previously had diabetes but no longer do.")

Model Output (Arabic, SAE-steered)

"نحن الآن لدينا فئران عمرها أربعة أشهر غير مصابة بالسكري والتي كانت سابقًا مصابة به."
(English: "We now have four-month-old mice that are non-diabetic and were previously diabetic.")

Figure 11: Example prompt and outputs for machine translation. The model is prompted in German and steered to generate an Arabic translation. English glosses are provided for readability.

G Multilingual Results for Gemma-2-9B

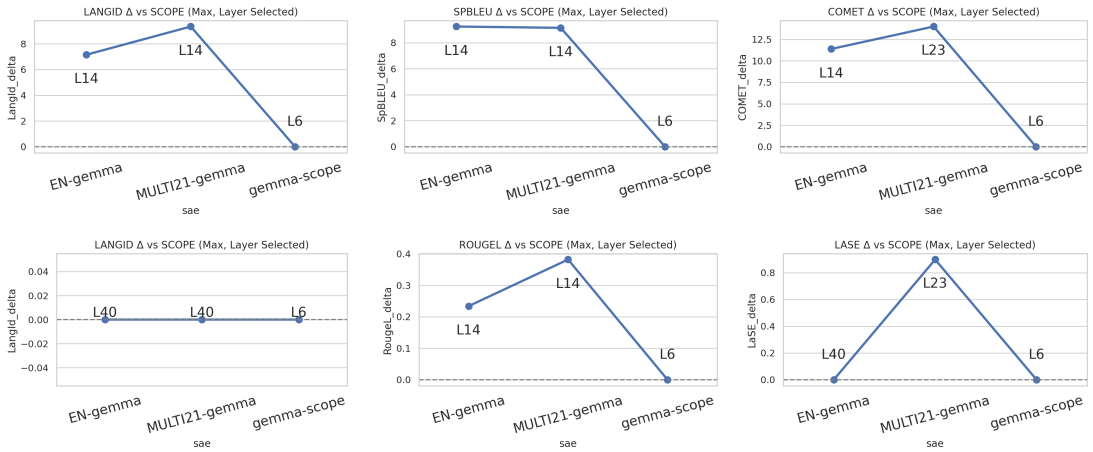

Figure 12: Performance deltas relative to Scope baselines for **Gemma-2-9B** at the selected steering layer. **Top:** FLORES machine translation (LangID, SpBLEU, COMET). **Bottom:** Cross-lingual summarization (LangID, ROUGE-L, LaSE). Improvements from multilingual training are smaller than in LLaMA but remain directionally consistent across tasks and metrics.

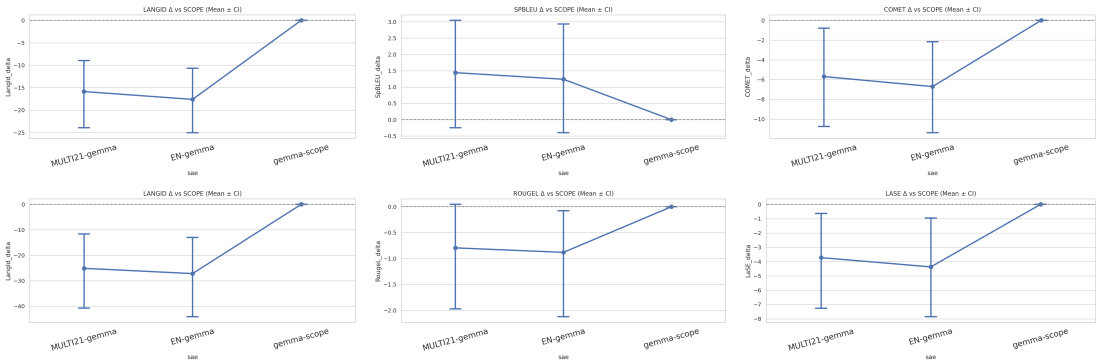

Figure 13: Performance deltas relative to Scope baselines for **Gemma-2-9B** averaged across layers. **Top:** FLORES machine translation (LangID, SpBLEU, COMET). **Bottom:** Cross-lingual summarization (LangID, ROUGE-L, LaSE).

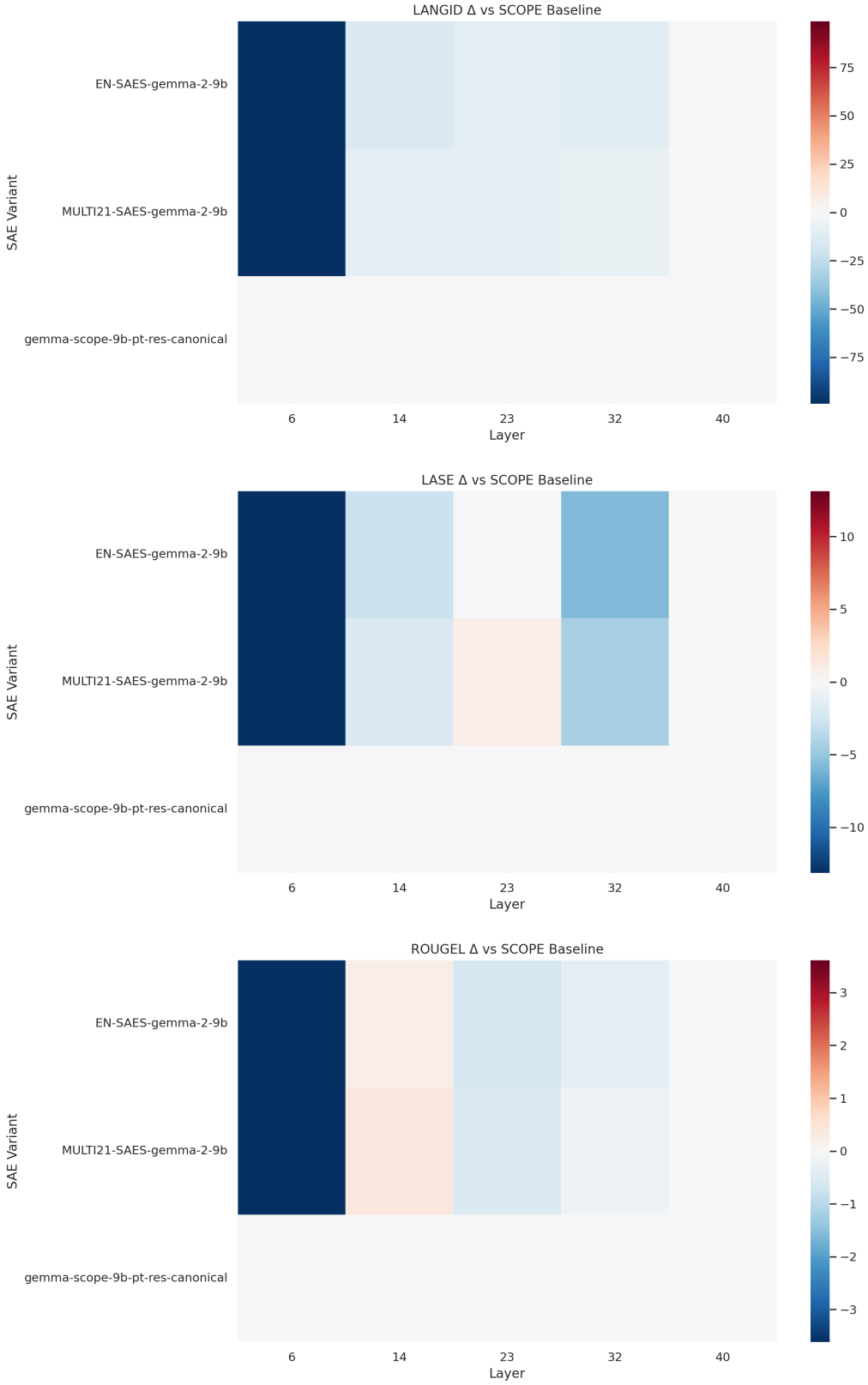

Figure 14: Layerwise heatmaps of performance deltas relative to *Gemma-Scope* for **Gemma-2-9B** on **cross-lingual summarization (CrossSumm)**. Columns show deltas in **LangID**, **LaSE**, and **ROUGE-L** as a function of steering layer. Regions of positive gain cluster around the intersection layers identified by our multilinguality–separability criterion (**L14** and **L23**), indicating that these depths support more reliable language control and semantic preservation, though gains remain smaller than those achieved by our multilingual SAEs.

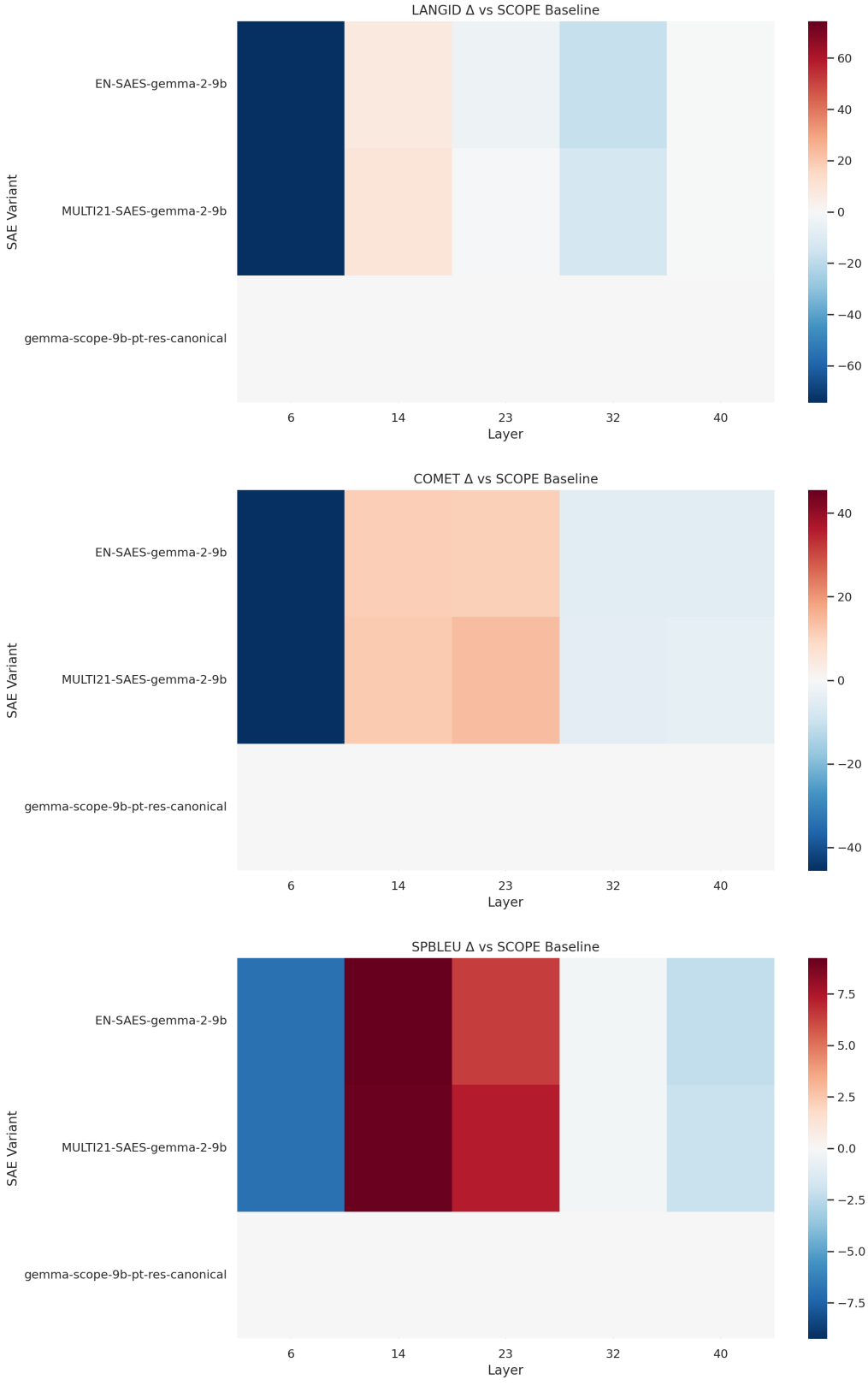

Figure 15: Layerwise heatmaps of performance deltas relative to *Gemma-Scope* for **Gemma-2-9B** on **machine translation (FLORES)**. Columns report deltas in **LangID**, **COMET**, and **SpBLEU** across steering layers. Improved performance concentrates near the predicted intersection layers (**L14** and **L23**), validating that these depths balance cross-lingual alignment and language separability, but still underperform compared to multilingual SAEs trained in our framework.

H Multilingual Results for LLaMA-3.1-8B

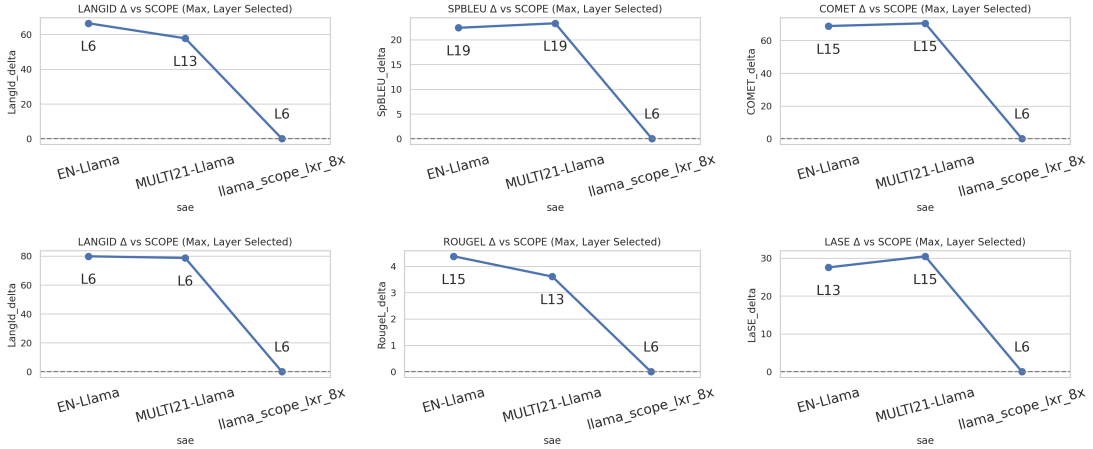

Figure 16: Performance deltas relative to Scope baselines for **LLaMA-3.1-8B** at the best-performing steering layer. **Top:** FLORES machine translation (LangID, SpBLEU, COMET). **Bottom:** Cross-lingual summarization (LangID, ROUGE-L, LaSE). Multilingual SAEs consistently outperform English-only SAEs across both tasks, with larger gains on semantic quality metrics.

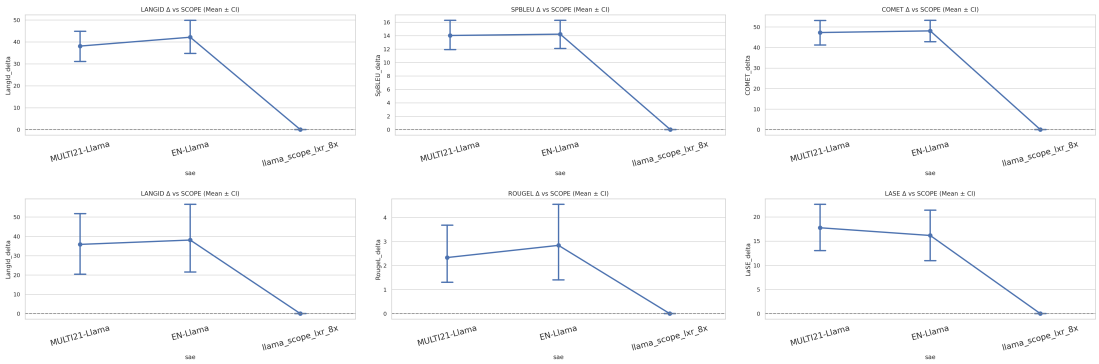

Figure 17: Performance deltas relative to Scope baselines for **LLaMA-3.1-8B** averaged across layers. **Top:** FLORES machine translation (LangID, SpBLEU, COMET). **Bottom:** Cross-lingual summarization (LangID, ROUGE-L, LaSE).

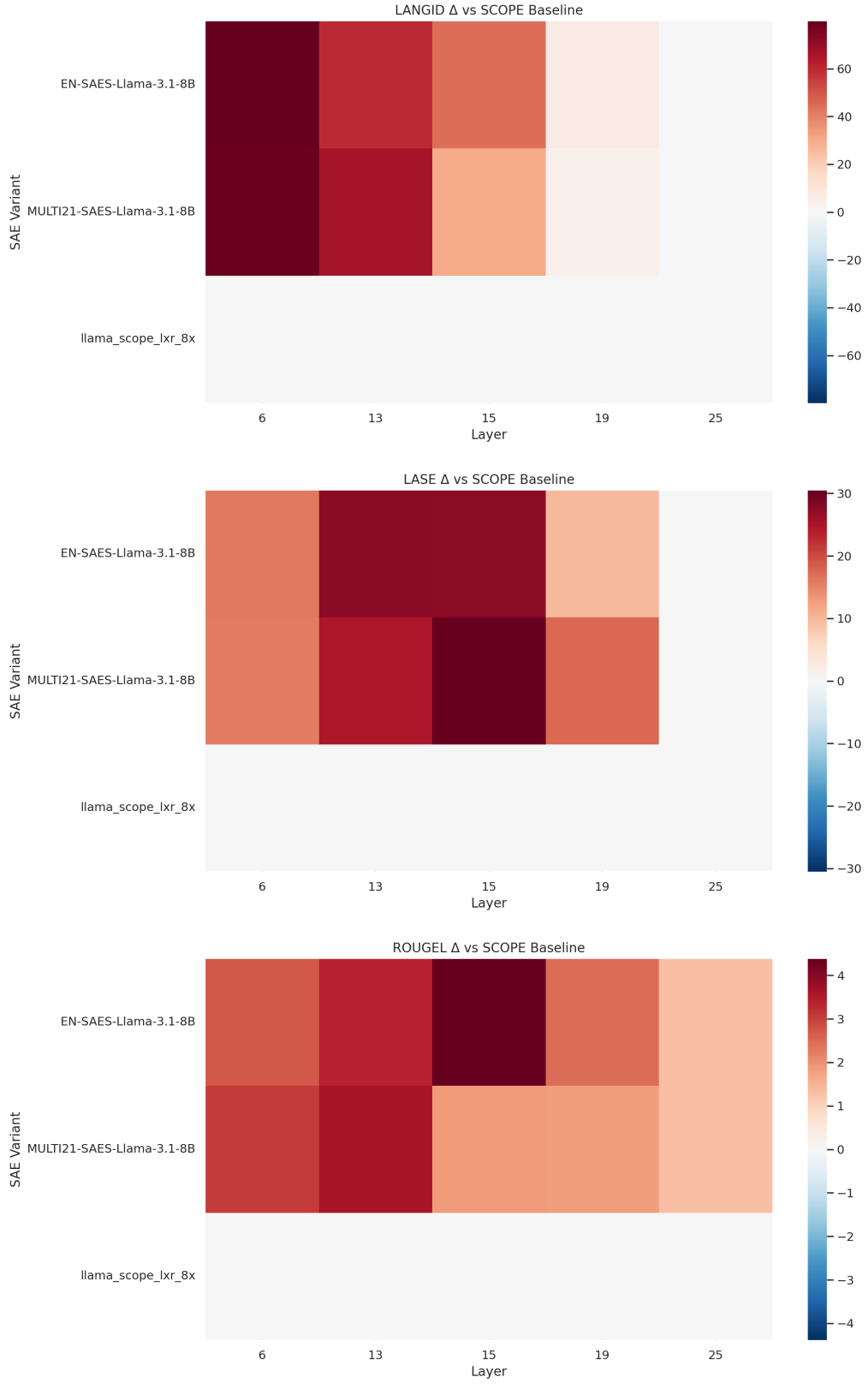

Figure 18: Layerwise heatmaps of performance deltas relative to *LLaMA-Scope* for **LLaMA-3.1-8B** on **cross-lingual summarization (CrossSumm)**. Columns show deltas in **LangID**, **LaSE**, and **ROUGE-L** as a function of steering layer. Unlike Gemma-Scope, LLaMA-Scope exhibits weak and diffuse gains across layers, with no clear concentration around an intersection depth, consistent with the absence of a strong multilinguality–separability balance and its limited downstream steering effectiveness.

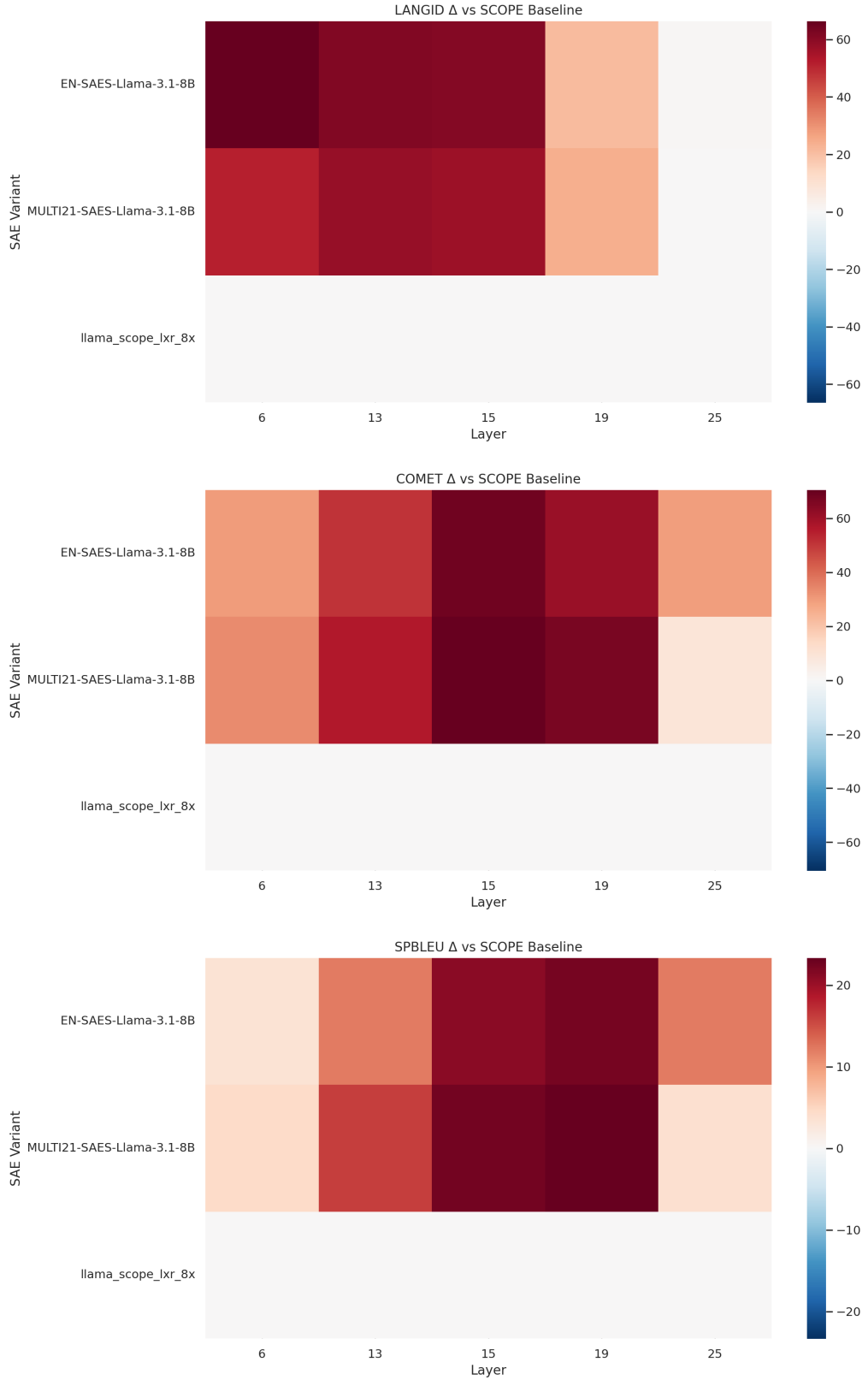

Figure 19: Layerwise heatmaps of performance deltas relative to *LLaMA-Scope* for **LLaMA-3.1-8B** on **machine translation (FLORES)**. Columns report deltas in **LangID**, **COMET**, and **SpBLEU** across steering layers. Performance improvements remain small and scattered across depth, with no distinct layer emerging as consistently effective, mirroring the lack of a clear intersection between multilingual alignment and language separability in LLaMA-Scope.

I Language vectors correlations and sparsity score

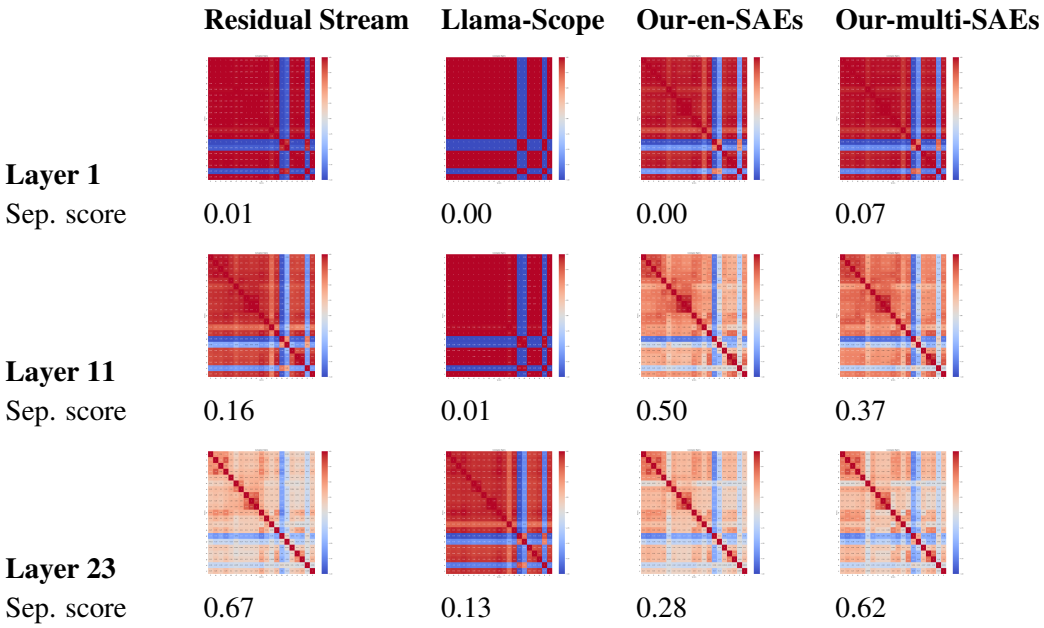

Figure 20: Comparison of LLama-3.1-8B model representation space using residual stream vectors, LLama-Scope sparse space vectors. and our trained SAEs sparse space vectors.

J Raw Results ($\texttt{tgt\_}i \neq \texttt{steer\_}j$) for Gemma-2-9B

layer	MULTI21-SAES			EN-SAES			gemma-scope			Base Model		
	LangID	ROUGEL	LASE	LangID	ROUGEL	LASE	LangID	ROUGEL	LASE	LangID	ROUGEL	LASE
6	0.0	0.52	0.0	0.0	0.53	0.0	**98.94**	4.14	13.13	-	-	-
14	48.33	**4.17**	16.55	42.92	4.02	15.75	57.73	3.78	**18.53**	-	-	-
23	11.81	1.25	12.38	10.79	1.15	11.29	21.39	1.78	11.48	-	-	-
32	9.35	1.27	11.39	5.46	1.12	10.03	17.04	1.47	15.79	-	-	-
40	0.0	0.59	0.0	0.0	0.56	0.0	0.0	0.6	0.0	-	-	-
Prompt (No) steering)	-	-	-	-	-	-	-	-	-	**36.48**	**3.47**	**22.87**

Table 2: **Gemma-2-9B CrossSumm, cross-lingual steering ($\texttt{tgt}_i \neq \texttt{steer}_j$).** Table entries report **LangID / ROUGE-L / LaSE** (column order). **No steering prompt results** (baseline; **LangID / ROUGE-L / LaSE**): **36.48**, **3.47**, **22.87**. *Prompt* scores are computed against the **prompt language** ($\texttt{tgt}_i$), whereas *steering* scores are computed against the **steering-vector language** ($\texttt{steer}_j$) and averaged over all mismatched pairs (i, j), averaging across target prompt languages for each steering language; therefore the prompt baseline and steering results are not directly comparable, but the baseline usefully characterizes the model's unsteered default behavior. **Strong shading** marks the best value *overall in the table* (per metric), while *light shading* marks the best value *within each SAE family* (per metric). Highlighted cells concentrate around the best layer, and the best overall results are often achieved by MULTI21-SAEs at that layer.

layer	MULTI21-SAES			EN-SAES			gemma-scope			Base Model		
	LangID	SpBLEU	COMET	LangID	SpBLEU	COMET	LangID	SpBLEU	COMET	LangID	SpBLEU	COMET
6	0.02	1.13	4.07	0.02	1.13	4.07	**74.39**	8.03	49.61	-	-	-
14	54.38	**24.80**	**73.55**	52.19	**24.90**	73.17	45.04	15.65	61.79	-	-	-
23	24.33	19.73	58.23	21.73	18.90	55.26	25.26	12.49	44.24	-	-	-
32	17.12	15.87	47.36	13.19	15.84	46.67	30.43	16.12	52.00	-	-	-
40	0.04	2.70	9.28	0.03	2.46	8.25	0.05	4.74	13.34	-	-	-
Prompt (No) steering)	-	-	-	-	-	-	-	-	-	**75.51**	**31.31**	**85.12**

Table 3: **Gemma-2-9B FLORES, cross-lingual steering ($\texttt{tgt}_i \neq \texttt{steer}_j$).** Table entries report **LangID / SpBLEU / COMET** (column order). **No steering prompt results** (baseline; **LangID / SpBLEU / COMET**): **75.51**, **31.31**, **85.12**. *Prompt* scores are computed against the **prompt language** ($\texttt{tgt}_i$), whereas *steering* scores are computed against the **steering-vector language** ($\texttt{steer}_j$) and averaged over all mismatched pairs (i, j), averaging across target prompt languages for each steering language; therefore the prompt baseline and steering results are not directly comparable, but the baseline usefully characterizes the model's unsteered default behavior. **Strong shading** marks the best value *overall in the table* (per metric), while *light shading* marks the best value *within each SAE family* (per metric). Highlighted cells concentrate around the best layer, and the best overall results are often achieved by MULTI21-SAEs at that layer.

K Raw Results ($\texttt{tgt\_}i \neq \texttt{steer\_}j$) for LLaMA-3.1-8B

layer	MULTI21-SAES			EN-SAES			llama-scope			Base Model		
	LangID	ROUGEL	LASE	LangID	ROUGEL	LASE	LangID	ROUGEL	LASE	LangID	ROUGEL	LASE
6	78.70	3.26	15.79	**79.86**	2.92	16.17	0.00	0.18	0.00	-	-	-
13	66.25	3.90	24.89	59.31	3.65	27.54	0.00	0.29	0.00	-	-	-
15	30.46	2.12	**30.47**	44.49	**4.64**	27.18	0.00	0.26	0.00	-	-	-
19	3.80	1.84	17.61	6.76	2.45	9.89	0.00	0.01	0.00	-	-	-
25	0.00	1.50	0.00	0.00	1.49	0.00	0.00	0.18	0.00	-	-	-
Prompt (No) steering)	-	-	-	-	-	-	-	-	-	**16.85**	**2.87**	**32.92**

Table 4: **LLaMA-3.1-8B CrossSumm, cross-lingual steering ($\texttt{tgt}_i \neq \texttt{steer}_j$).** Table entries report **LangID / ROUGE-L / LaSE** (column order). **No steering prompt results** (baseline; **LangID / ROUGE-L / LaSE**): **16.85**, **2.87**, **32.92**. *Prompt* scores are computed against the **prompt language** ($\texttt{tgt}_i$), whereas *steering* scores are computed against the **steering-vector language** ($\texttt{steer}_j$) and averaged over all mismatched pairs (i, j), averaging across target prompt languages for each steering language; therefore the prompt baseline and steering results are not directly comparable, but the baseline usefully characterizes the model's unsteered default behavior. **Strong shading** marks the best value *overall in the table* (per metric), while *light shading* marks the best value *within each SAE family* (per metric). Highlighted cells concentrate around the best layer, and the best overall results are often achieved by MULTI21-SAEs at that layer.

layer	MULTI21-SAES			EN-SAES			llama-scope			Base Model		
	LangID	SpBLEU	COMET	LangID	SpBLEU	COMET	LangID	SpBLEU	COMET	LangID	SpBLEU	COMET
6	54.22	4.42	39.77	**68.77**	3.38	36.55	2.36	0.02	6.33	-	-	-
13	58.24	16.11	66.27	62.14	12.20	60.44	0.47	0.02	9.65	-	-	-
15	56.97	22.53	**73.25**	60.92	21.02	71.57	0.10	0.00	2.72	-	-	-
19	24.14	**23.35**	68.65	21.32	22.44	62.75	0.12	0.01	1.86	-	-	-
25	0.09	3.77	11.25	0.64	12.12	31.88	0.09	0.01	2.13	-	-	-
Prompt (No) steering)	-	-	-	-	-	-	-	-	-	**91.06**	**31.22**	**83.58**

Table 5: **LLaMA-3.1-8B FLORES, cross-lingual steering ($\texttt{tgt}_i \neq \texttt{steer}_j$).** Table entries report **LangID / SpBLEU / COMET** (column order). **No steering prompt results** (baseline; **LangID / SpBLEU / COMET**): **91.06**, **31.22**, **83.58**. *Prompt* scores are computed against the **prompt language** ($\texttt{tgt}_i$), whereas *steering* scores are computed against the **steering-vector language** ($\texttt{steer}_j$) and averaged over all mismatched pairs (i, j), averaging across target prompt languages for each steering language; therefore the prompt baseline and steering results are not directly comparable, but the baseline usefully characterizes the model's unsteered default behavior. **Strong shading** marks the best value *overall in the table* (per metric), while *light shading* marks the best value *within each SAE family* (per metric). Highlighted cells concentrate around the best layer, and the best overall results are often achieved by MULTI21-SAEs at that layer.

L Per-Language Results ($\texttt{tgt\_}i = \texttt{steer\_}j$) for Gemma-2-9B

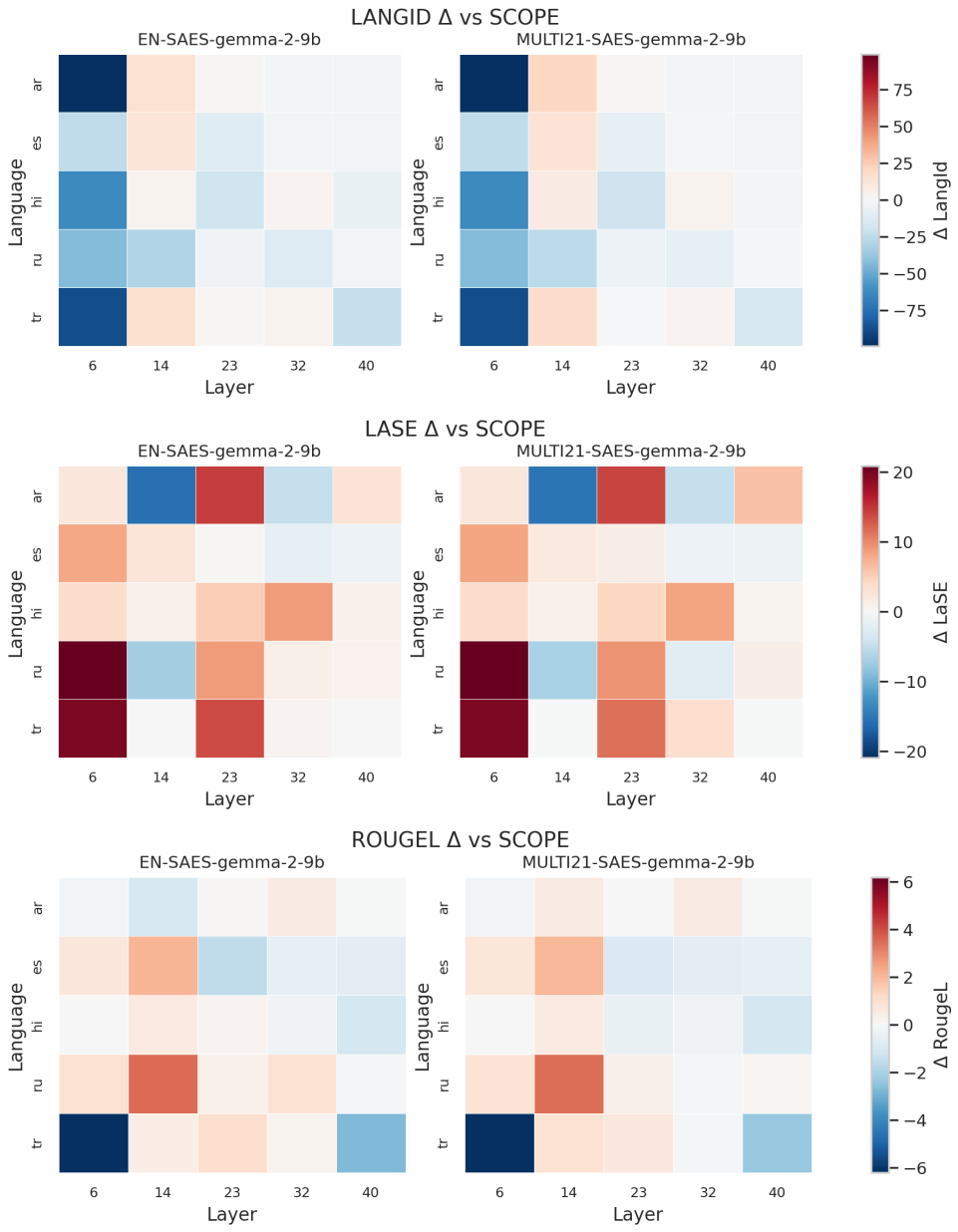

Figure 21: Per-language, per-layer performance deltas for **Gemma-2-9B** on the CROSSSUM task when the steering language matches the target language ($\texttt{tgt\_}i = \texttt{steer\_}j$). Each heatmap shows the change relative to the SCOPE baseline (excluded), with rows corresponding to target languages, columns to transformer layers, and separate panels for each SAE variant. Positive values indicate improvements over the baseline.

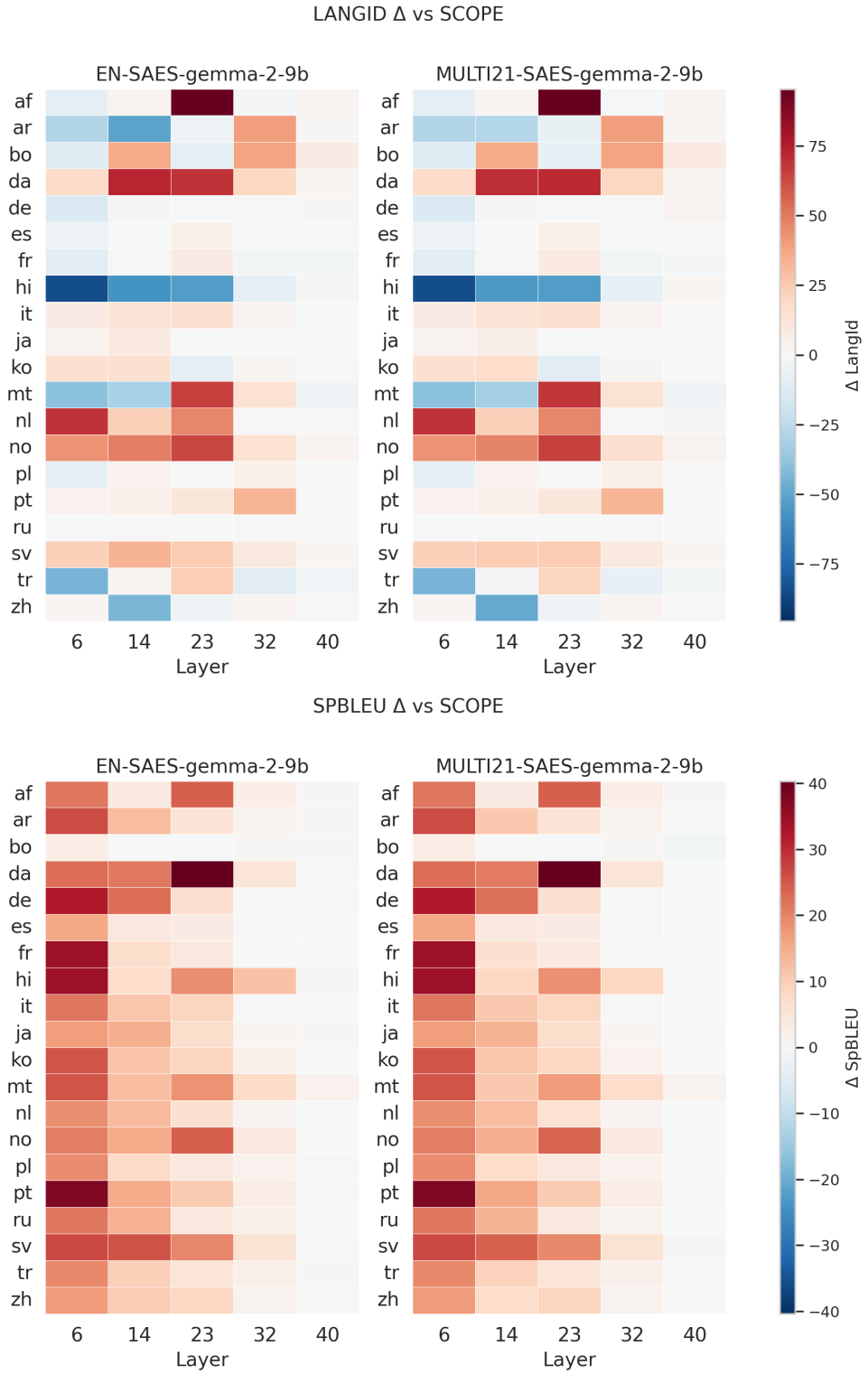

Figure 22: Per-language, per-layer deltas for **Gemma-2-9B** on FLORES under matched steering and target languages (`tgt_`i = `steer_`j). The heatmaps show the impact of SAE variants on language identification and translation quality across model depth.

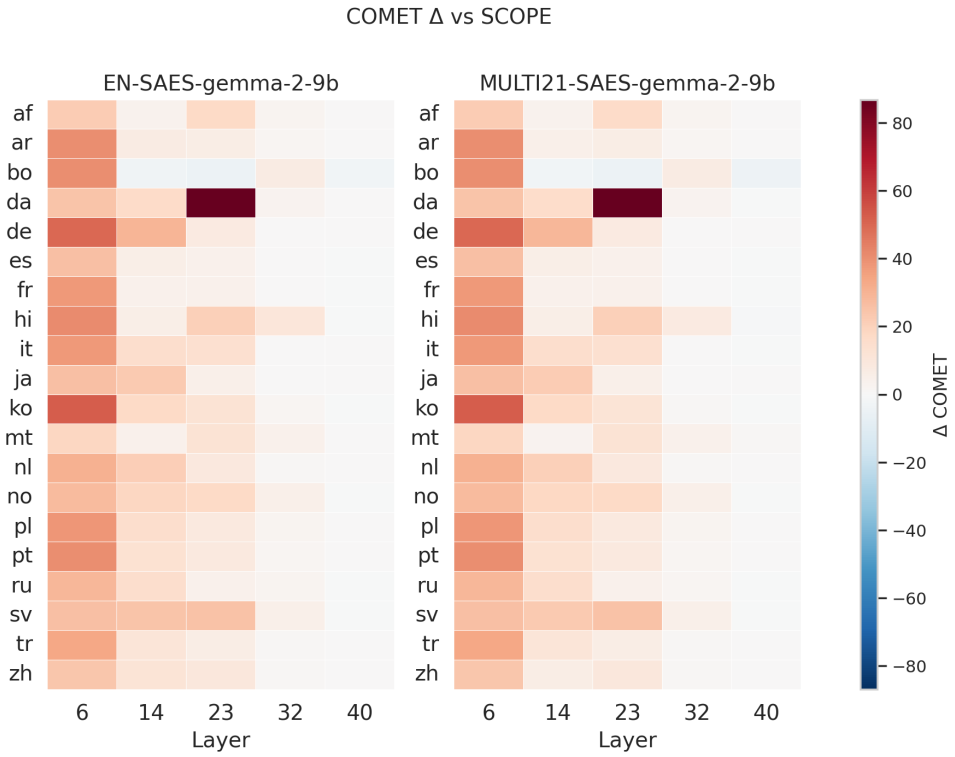

Figure 23: Per-language, per-layer COMET score deltas for **Gemma-2-9B** on FLORES with matched steering and target languages (`tgt_`i = `steer_`j). This figure emphasizes how semantic translation quality varies across languages, layers, and SAE configurations relative to the SCOPE baseline.

M Per-Language Results (`tgt_`$i \neq$ `steer_`j) for Gemma-2-9B

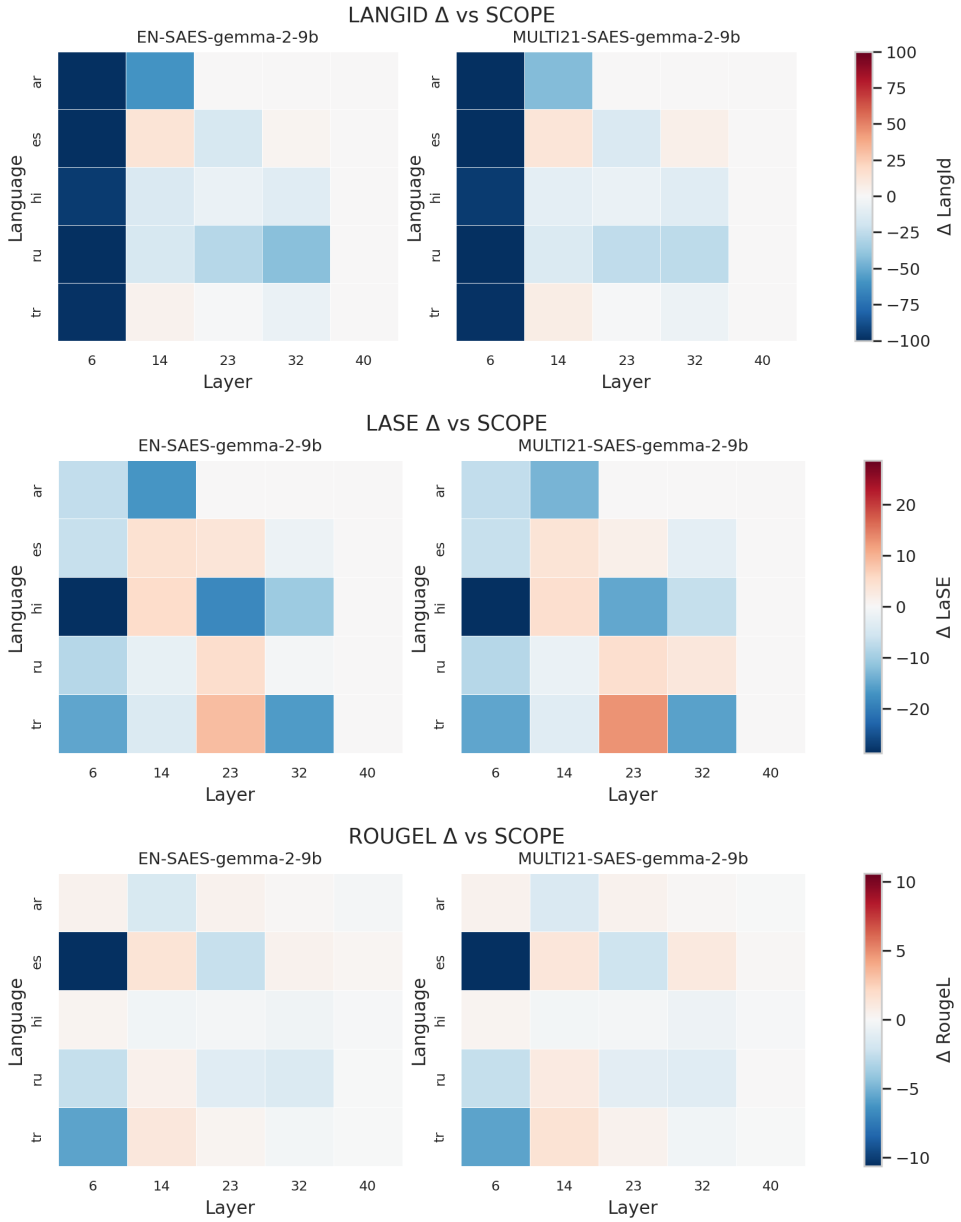

Figure 24: Per-language, per-layer performance deltas for **Gemma-2-9B** on the CROSSSUM task under cross-lingual steering (`tgt_`$i \neq$ `steer_`j). Each heatmap shows how steering in a different language affects summarization quality and language identification across layers and SAE variants, relative to the SCOPE baseline.

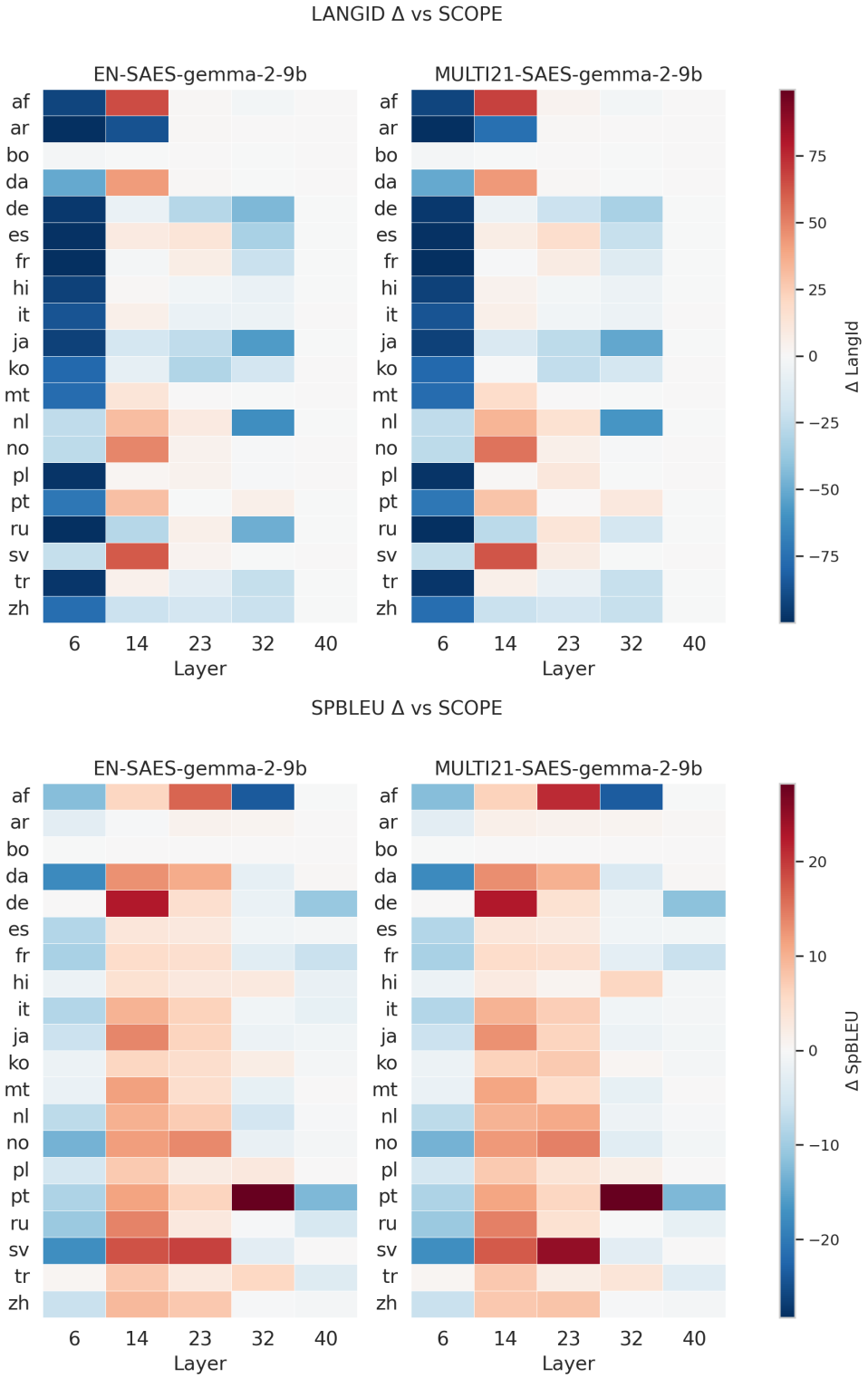

Figure 25: Per-language, per-layer performance deltas for **Gemma-2-9B** on FLORES with cross-lingual steering ($\texttt{tgt\_}i \neq \texttt{steer\_}j$). The figure highlights the degradation or transfer effects induced by mismatched steering languages across model depth.

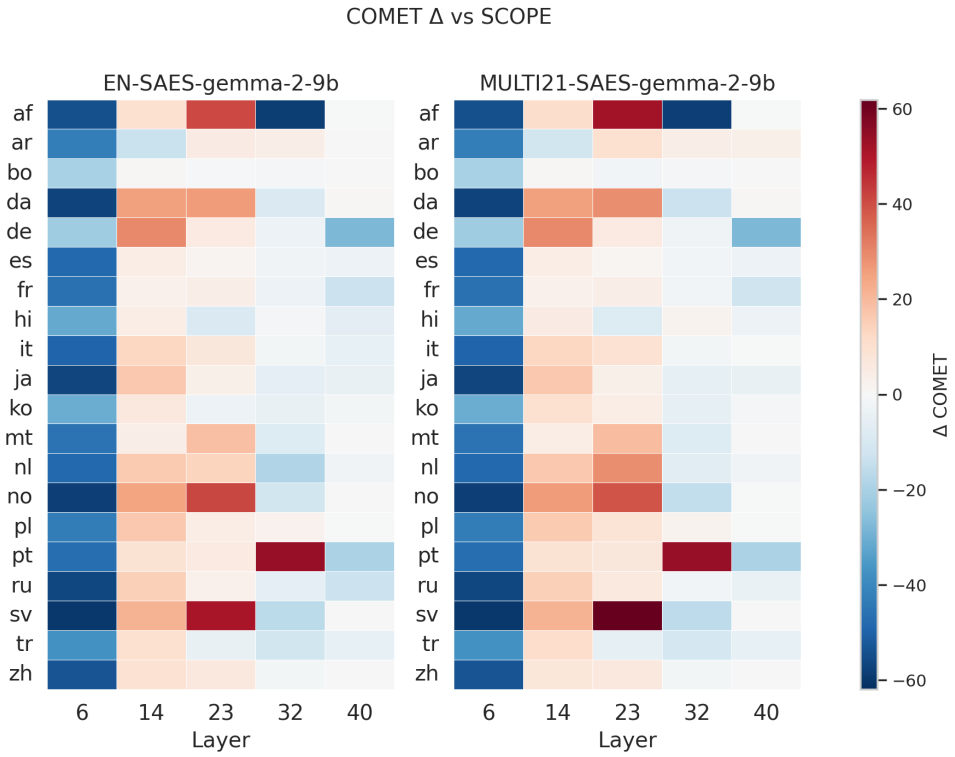

Figure 26: Per-language, per-layer COMET score deltas for **Gemma-2-9B** on FLORES under cross-lingual steering ($\texttt{tgt\_}i \neq \texttt{steer\_}j$). This visualization captures how semantic translation quality responds to cross-lingual steering at different layers and SAE variants, relative to the SCOPE baseline.

N Per-Language Results ($\texttt{tgt\_}i = \texttt{steer\_}j$) for LLaMA-3.1-8B

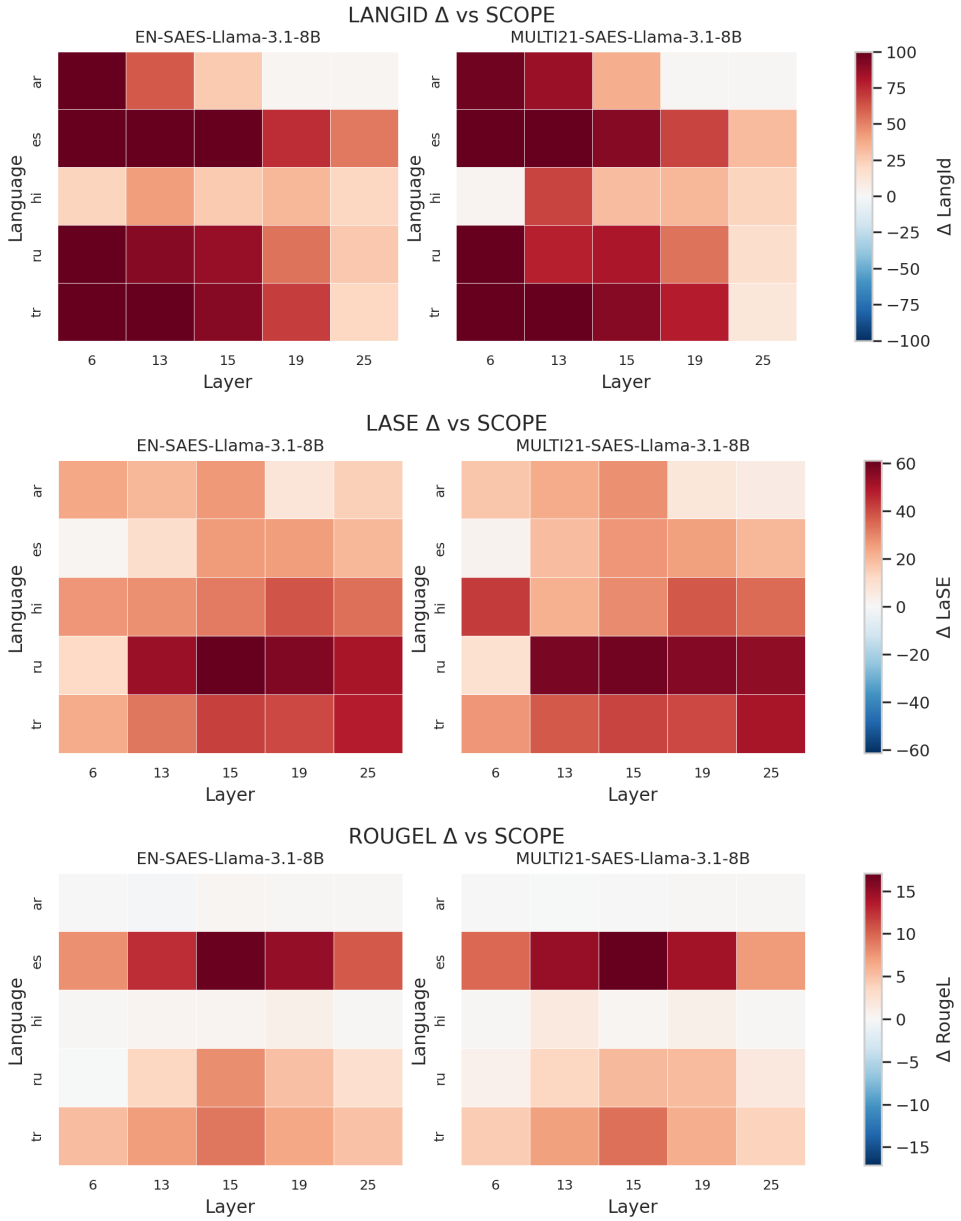

Figure 27: Per-language, per-layer performance deltas for **LLaMA-3.1-8B** on the CROSSSUM task when the steering language matches the target language ($\texttt{tgt\_}i$ = $\texttt{steer\_}j$). Each heatmap shows the change relative to the SCOPE baseline (excluded), with rows corresponding to target languages, columns to transformer layers, and separate panels for each SAE variant. Positive values indicate improvements over the baseline.

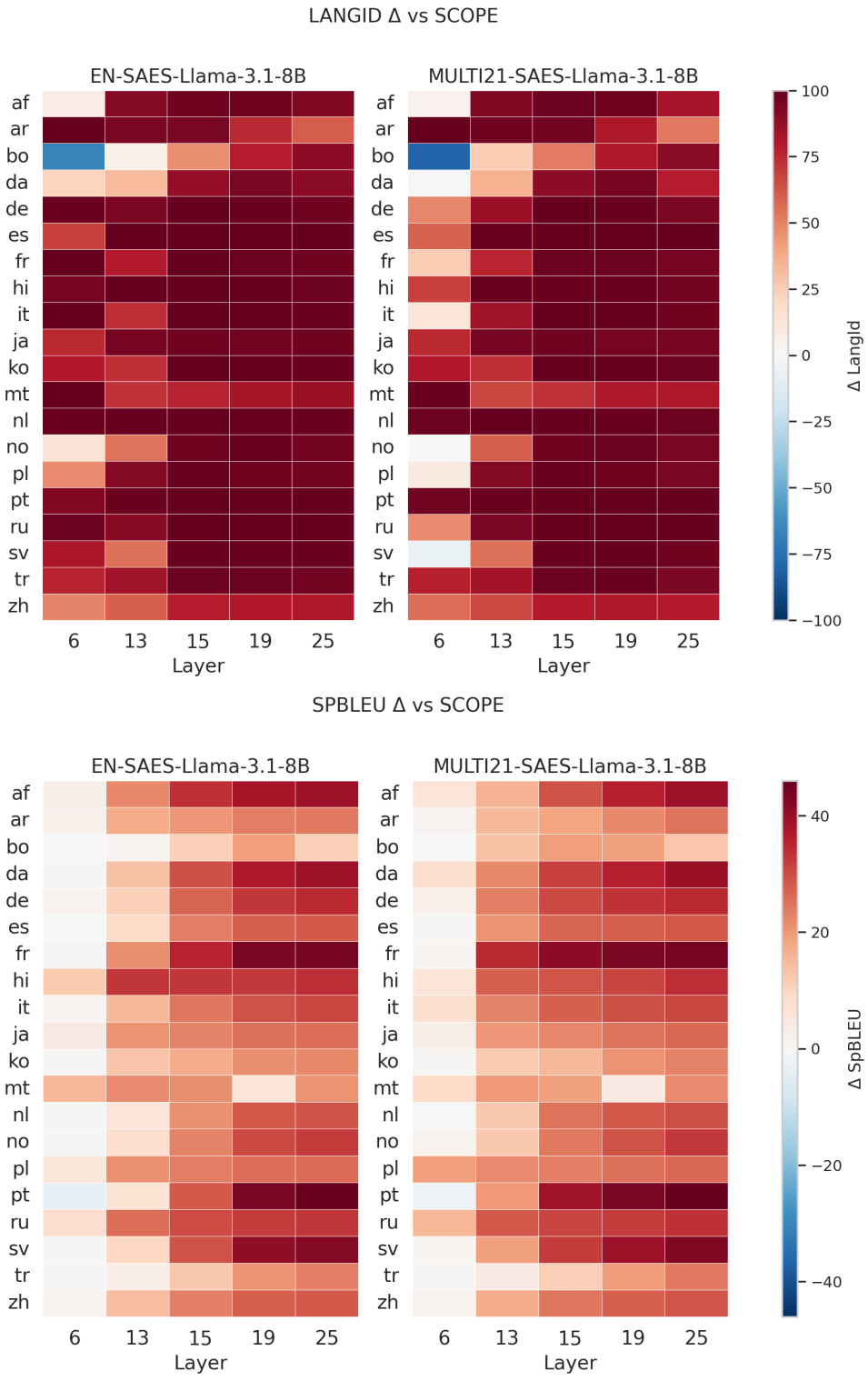

Figure 28: Per-language, per-layer performance deltas for **LLaMA-3.1-8B** on the FLORES benchmark when the steering language matches the target language ($\texttt{tgt\_}i = \texttt{steer\_}j$). Results are shown for language identification (LangID) and translation quality (SpBLEU), aggregated per SAE variant and measured relative to the SCOPE baseline.

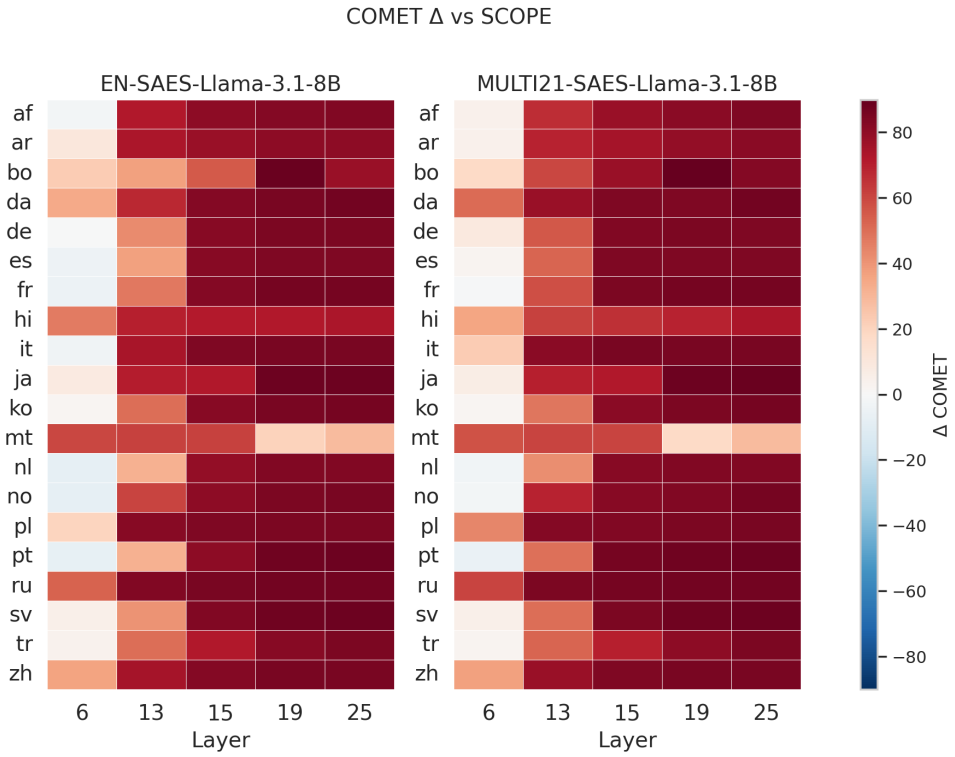

Figure 29: Per-language, per-layer COMET score deltas for **LLaMA-3.1-8B** on FLORES under matched steering and target languages ($\texttt{tgt\_}i = \texttt{steer\_}j$). The heatmap highlights how SAE interventions affect semantic translation quality across languages and model depth, relative to the SCOPE baseline.

O Per-Language Results ($\texttt{tgt\_}i \neq \texttt{steer\_}j$) for LLaMA-3.1-8B

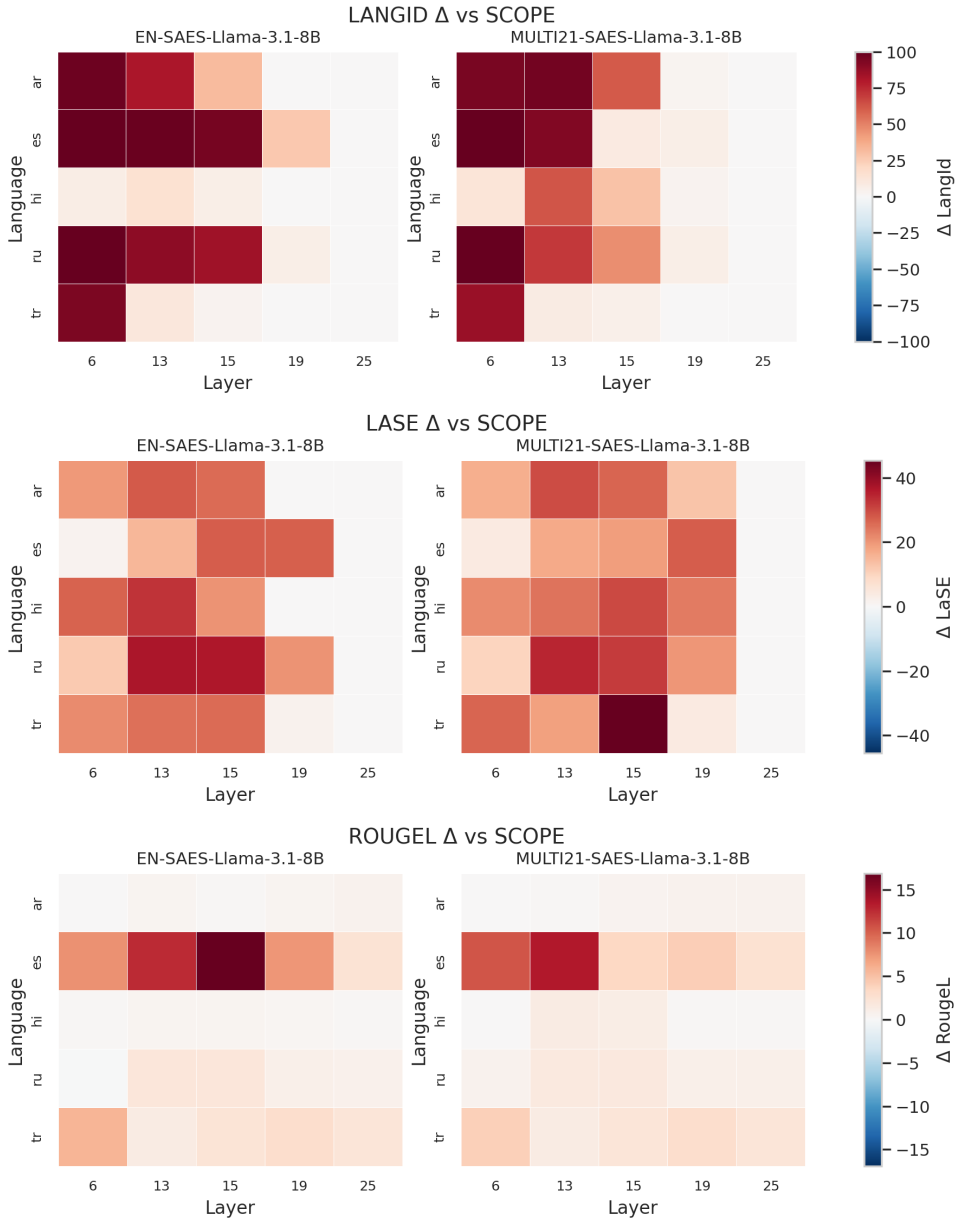

Figure 30: Per-language, per-layer performance deltas for **LLaMA-3.1-8B** on the CROSSSUM task under cross-lingual steering ($\texttt{tgt\_}i \neq \texttt{steer\_}j$). Each panel corresponds to a different SAE variant, showing how mismatched steering languages impact summarization quality and language identification across layers, relative to the SCOPE baseline.

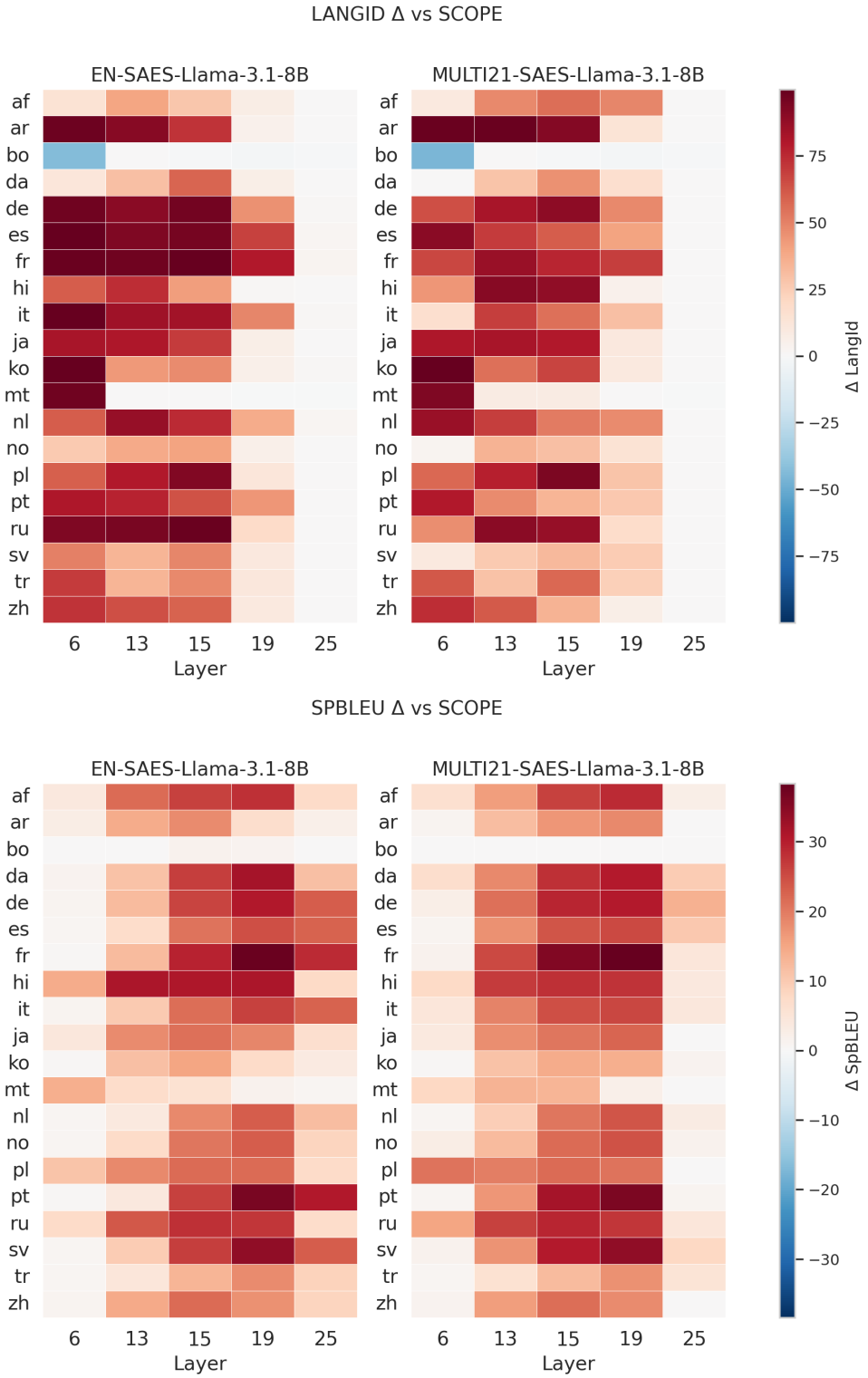

Figure 31: Per-language, per-layer performance deltas for **LLaMA-3.1-8B** on FLORES with cross-lingual steering ($\texttt{tgt\_}i \neq \texttt{steer\_}j$). The figure illustrates how steering in a different language affects language identification accuracy and translation quality across layers and SAE variants.

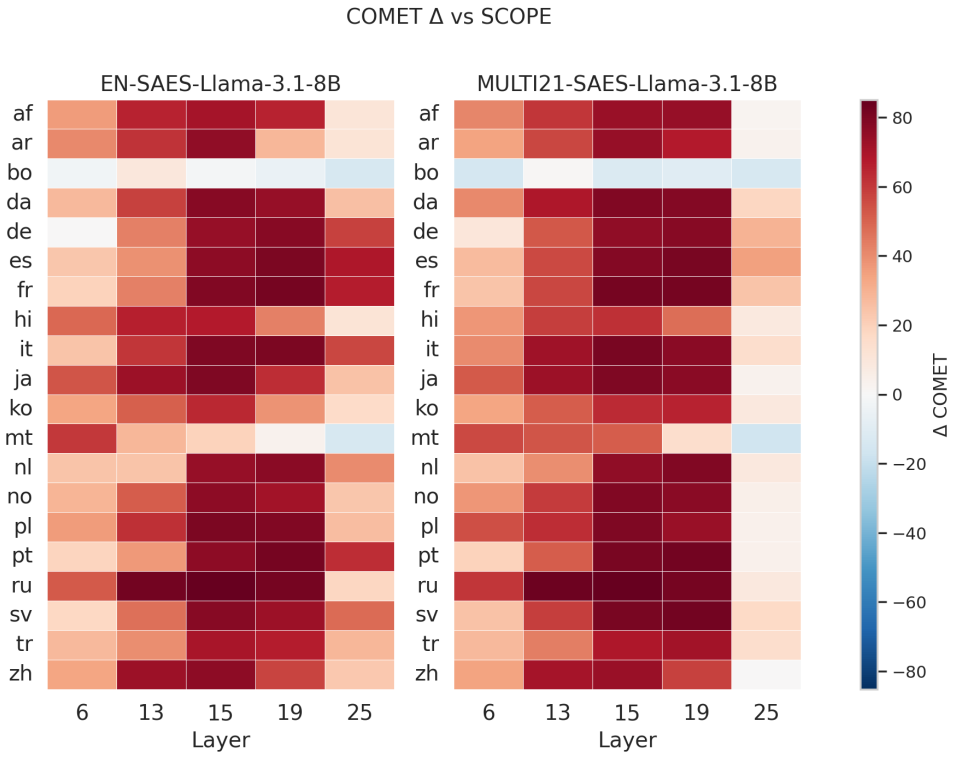

Figure 32: Per-language, per-layer COMET score deltas for **LLaMA-3.1-8B** on FLORES under cross-lingual steering ($\texttt{tgt\_}i \neq \texttt{steer\_}j$). Results highlight the sensitivity of semantic translation quality to steering language mismatches at different depths of the model.