Latent Mechanisms of Code-Switching in Large Language Models

Anonymous Authors[1]

Abstract

Multilingual large language models can exhibit *unintended code-switching* – unnecessarily alternating between languages during generation. We present a comparative study of three methods that identify language-controlling latents in cross-layer transcoders: activation value-based selection (`ValSel`), activation frequency-based selection (`FreqSel`), and LLM-generated latent annotation-based selection (`AnnSel`). To evaluate the efficacy of these methods in identifying language-controlling latents, we introduce two multilingual code-switching benchmarks designed for fine-grained analysis of language steering across seven languages. Through targeted intervention experiments on `Gemma-2-2B` and `Qwen3-4B`, we find that all three methods effectively manipulate generation language, with `FreqSel` achieving the strongest overall performance, while `AnnSel` offering interpretable latent selection through explicit language annotations. A knock-out analysis suggests the methods select non-overlapping but each-functional latent subsets, indicating redundancy rather than a single canonical language direction.

1. Introduction

Large language models have achieved remarkable multilingual capabilities, but this flexibility comes with an unintended consequence: *code-switching*, where models inappropriately alternate between languages mid-generation (Marchisio et al., 2024). This problem motivated explicit language consistency rewards during training of models such as DeepSeek-R1 (DeepSeek-AI et al., 2025).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Recent mechanistic interpretability work has identified neurons and latents that causally influence output language (Tang et al., 2024; Deng et al., 2025), suggesting that targeted interventions could provide fine-grained language control in generation. However, to our knowledge, no systematic comparison exists of language-controlling mechanism discovery, and it remains unclear whether different approaches uncover the same underlying latents or redundant representations.

We compare three approaches that identify language-specific latents in cross-layer transcoders (CLTs) (Ameisen et al., 2025): **`ValSel`**, based on activation value differences (Deng et al., 2025); **`FreqSel`**, based on language-specific activation frequency (Andrylie et al., 2025); and **`AnnSel`** that utilizes LLM-generated annotations to find latents whose descriptions explicitly reference the target language from an extracted circuit graph. Unlike prior work on single-layer SAEs, CLTs provide a unified sparse dictionary across all layers, enabling holistic analysis of language representations in latent space.

2. Preliminaries

Mechanistic Approaches to Language Control Prior work has investigated language control at multiple levels of abstraction. At the neuron level, Tang et al. (2024) identified language neurons whose selective activation steers output language. At the representation level, Goncharov et al. (2025) extracted language directions via PCA, enabling steering with directional ablation. More recently, sparse autoencoders (SAEs), which address *polysemanticity* by learning an overcomplete sparse basis of monosemantic features (Bricken et al., 2023), have been applied to isolate language-controlling features through activation magnitude (Deng et al., 2025) or frequency patterns (Andrylie et al., 2025).

Cross-Layer Transcoders (CLTs) extend SAEs by replacing MLP blocks across all layers with a jointly-trained sparse dictionary (Ameisen et al., 2025), capturing features spanning model depth. In this paper, we use CLTs trained for `Gemma-2-2B` (Lieberum et al., 2024) and `Qwen3-4B` (Hanna, 2025).

3. Manipulating Generation Language through Latent Interventions

We study three language-specific **latent selection methods**, and four **latent intervention strategies** that steer generation toward a target language.

3.1. Latent Selection Methods

Given a cross-layer transcoder with latent space $\mathcal{S}$ and a collection of languages L, we identify for language $l \in L$ a subset of latents $\mathcal{S}^l \subset \mathcal{S}$ that are specifically associated with the generation of l. We compare three selection criteria based on activation magnitude, activation frequency, and semantic annotation.

3.1.1. Value-Based Selection (ValSel)

Inspired by Deng et al. (2025), this method identifies latents whose activation magnitude is distinctively higher for a target language compared to others. For a latent s and a language-specific corpus D^l, we compute the mean activation:

$$\mu_s^l = \frac{1}{|D^l|} \sum_{t \in D^l} a_s(t) \tag{1}$$

where $a_s(t)$ denotes the activation of latent s for token t. The *l-monolinguality score* of latent s for language l is then defined as:

$$v_s^l = \mu_s^l - \frac{1}{|L|-1} \sum_{j \in L \setminus \{l\}} \mu_s^j \tag{2}$$

We select the top-K latents with the highest l-monolinguality scores for each language l.

3.1.2. Frequency-Based Selection (FreqSel)

Following Andrylie et al. (2025), this method identifies latents that activate more frequently for a specific language. We compute the *activation probability* for latent s in language l as:

$$P_s^l = \frac{1}{|D^l|} |\{t \in D^l : a_s(t) > 0\}| \tag{3}$$

A latent is considered *specific* to language l if: (1) $P_s^l = \max_{j \in L} P_s^j$ and (2) no language $l' \neq l$ satisfies $P_s^{l'} \geq T \cdot P_s^l$ for a threshold $T \in [0, 1]$. We select the top-K specific latents for l based on P_s^l.

3.1.3. Annotation-Based Selection (AnnSel)

We propose an approach that supports interpretability by leveraging LLM-generated latent annotations. This method operates in three stages: circuit tracing, path pruning and annotation filtering.

For each sentence in D^l, we trace the attribution graph from input tokens to the predicted output token using Circuit Tracer (Hanna et al., 2025). We truncate sentences at a midpoint ensuring the target token is a valid word in l. Then, we greedily prune the attribution graph by retaining only its high-importance paths from each token embedding to the top predicted logit by maximum bottleneck edge weight. From the pruned graph, we extract all latents contributing to the final token position, but retain only the latents whose annotations explicitly reference the language l (e.g., containing "Japanese" or "Japan" for $l =$'`ja`'), ensuring interpretable, language-related selection.

3.2. Intervention Strategies

For the sets of discovered language-specific latents $\{\mathcal{S}^l\}^{l \in L}$, we apply inference-time interventions to steer generation from a context language l_0 toward a target language l. We evaluate four strategies:

- **Target Amplification**: Set each latent $s \in \mathcal{S}^l$ to its mean nonzero activation computed over D^l. This naturally amplifies the signal for the target language l while avoiding out-of-distribution effects, without affecting other languages.
- **Distractor Zero Ablation**: Set to zero all latents associated with the distractor language l_0, or the latents of all non-target languages $\bigcup_{j \in L \setminus \{l\}} \mathcal{S}^j$. This suppresses competing signals from other languages.
- **Distractor Direction Ablation**: Ablate the latents of the distractor language $\mathcal{S}^{l_0}$ by *direction ablation*, where the the context language get suppressed by removing the projection of the residual stream activation along a latent's direction. We do this either for the single layer with most latents (*one-layer*), or for all-layers (*multi-layer*).
- **Combined Interventions**: Simultaneously apply a form of distractor ablation along with target amplification, boosting the signal of the target language while suppressing that of the context language. We test distractor zero ablation + target amplification (**Zero+Amp**), and distractor one-layer direction ablation + target amplification (**1L+Amp**).

4. Experiments

All the experiments are conducted on the transcoders of `Gemma-2-2B-pt`[1] from Gemma-scope (Lieberum et al., 2024) and `Qwen3-4b`[2].

For the activation value-based `ValSel` we use $K = 50$ (to parallel the 2-latent per layer choice in (Deng

[1] https://hf.co/mwhanna/gemma-scope-transcoders
[2] https://hf.co/mwhanna/qwen3-4b-transcoders

Table 1. Data format and examples of the two evaluation datasets. The sentence frame (in sans-serif) is in the context language l_0, while the target language l is in **bold**. The examples show selections of context/target languages (l_0, l).

	Context	Target
Antonyms	The opposite of "**[adj]**" is "	**[antonym]**
ex1 (en, fr)	*The opposite of "**grand**" is "*	***petit***
ex2 (de, ja)	*Das Gegenteil von "**大きい**" ist "*	**小さい**
Enumerations	The [enum] are: **[item 1], [item 2],**	**[item 3], [item 4], . . .**
ex3 (en, es)	*The four seasons are:* ***el verano, el otoño,***	***el invierno, la primavera***
ex4 (fr, zh)	*Les jours de la semaine sont:* **星期日, 星期一, 星期二,**	**星期三, 星期四,** . . .

et al., 2025) for the 26 layers of `Gemma-2-2B`). For the frequency-based selection `FreqSel`, we consider latents that are active in more than 98% of a language's corpus and set $T = 0.8$ following (Andrylie et al., 2025). For `AnnSel`, we utilize the Circuit Tracer tool (Hanna et al., 2025) and Neuronpedia annotations[3].

To extract the latent sets, we use FLORES+ (NLLB Team et al., 2024) in seven languages: English (en), French (fr), German (de), Spanish (es), Chinese (zh), Korean (ko), and Japanese (ja). We apply the three selection methods separately to both CLTs and each language subset. We report in the Appendix (Table 6) the count of latents extracted per method and language.

4.1. Datasets

We introduce **Antonyms** and **Enumerations**, two novel evaluation datasets designed to measure the effectiveness of language-steering interventions.[4] **Antonyms** task requires producing a target antonym in a desired language l given a sentence context in a different language l_0, creating a controlled code-switching scenario, while **Enumerations** asks for a continuation of a listing in the desired language, testing open-ended generation. The data format and some examples of the two datasets are shown in Table 1.

Dataset construction. We collect 100 antonym pairs per language for the seven languages (en, fr, de, es, zh, ja, ko) for **Antonyms** and 7 ordered categories for **Enumerations**. The complete lists of adjectives and words are provided in Appendix H and I. For each of the $7 \times 6 = 42$ ordered pairs of context and target languages (l_0, l) where $l_0 \neq l$, we generate a test example per antonym pair and per enumeration category.

Evaluation protocol. Given a context, we measure whether the model assigns a higher probability to the target antonym or enumeration in language l compared to its translations in other languages. For each example, we record the *logit* assigned to the target token of the antonyms for **Antonyms** and the length-normalized *logprob* for **Enumerations** across all seven languages. A successful intervention should increase the target logit or logprob post-intervention relative to the baseline before the intervention.

[3] https://www.neuronpedia.org

[4] The datasets will be publicly released.

4.2. Findings

4.2.1. Intervention strategy effectiveness

To assess language manipulation efficacy, we intervene on the extracted latents enforcing a target language l in a sentence with context language l_0 using **Antonyms**, where the target is a single token. We measure the target token's ***top-1 logit margin*** which is the separation of the logit of the target adjective from its best competitor: $m_l = \max_{i \in L \setminus \{l\}} a_i - a_l$, where a_l is the logit of the target adjective in language l. A more negative m_l value indicates a higher probability for target language l. We measure the change in m_l due to the intervention:

$$\Delta m_l = -(m_l^{after} - m_l^{before}) \tag{4}$$

hence a larger positive change indicates a more effective intervention. Summary results for `Gemma-2-2B` are in Table 2, with detailed results in the Appendix Tables 7 to 9, as well as results for `Qwen3-4B` (Tables 10 to 12).

The most effective intervention is **Zero+Amp**, combining zero ablation of distractor language l_0 with target language amplification, followed by **1L+Amp** that instead uses single-layer directional ablation of l_0. Zero ablation seems to outperform the more powerful direction ablation possibly due to latent overlap of related languages, which weakens the target language when direction-ablating the context language.

4.2.2. Selection method efficacy

We now apply the most effective intervention (**Zero+Amp**) to all selection methods and languages to contrast their performance per target language. Table 3 shows summary results for `Gemma-2-2B` on **Antonyms** and **Enumerations** (detailed results in Tables 13 to 15 in the Appendix), and results for `Qwen3-4B` are in Table 16.

Table 2. Intervention effectiveness per method using **Antonyms**, computed as the total change in top-1 logit margin under the intervention (Equation 4) averaged over all target languages (bigger is better) for `Gemma-2-2B`.

	Distractor (l_0) zero ablation	$L \setminus \{l\}$ zero ablation	Distractor (l_0) one-layer direction ablation	Distractor (l_0) multi-layer direction ablation	Target (l) amplification	Zero+Amp	1L+Amp
`ValSel`	-8.02	22.39	1.99	13.98	69.57	**79.83**	69.73
`FreqSel`	2.02	21.39	-2.95	-51.62	90.83	**99.07**	96.16
`AnnSel`	-0.20	11.90	0.30	64.05	87.00	**96.72**	89.33

Table 3. Selection method efficacy per language under the **Zero+Amp** intervention for `Gemma-2-2B`. *Top*: post-intervention total change in top-1 logit margin for **Antonyms**; *Bottom*: total change in normalized logProb of the target sequence for **Enumerations**.

		es	en	zh	de	ja	fr	ko	Avg
Antons.	`ValSel`	**14.19**	**10.16**	16.06	13.56	4.63	9.01	12.18	11.40
	`FreqSel`	8.95	4.34	**34.50**	7.85	8.65	6.45	**28.30**	**14.15**
	`AnnSel`	13.86	4.87	24.52	**19.67**	**11.12**	**13.54**	9.11	13.81
Enums.	`ValSel`	1.00	-0.08	1.13	0.93	1.54	0.72	1.87	1.02
	`FreqSel`	**2.08**	-0.33	**2.12**	**2.69**	**2.52**	**1.40**	**3.92**	**2.06**
	`AnnSel`	0.90	**0.30**	1.05	1.83	1.47	0.85	0.38	0.99

In total across all settings and models, all the selection methods are effective. `FreqSel` achieves the highest total absolute logit and logProb change across the tasks, with significant performance gains in Asian languages (especially Chinese and Korean). AnnSel, on the other hand, shows decent performance across all languages but is hindered by its weak performance in Korean which might be justified by the small number of extracted latents for Korean (Table 6).

4.2.3. Is there redundancy in latent language representation?

To study to what extent the latent sets identified by the three selection methods are equivalent we conduct a knock-out experiment that involves a pair of methods and the same target language. We utilize the amplification intervention on the latents of the first method, then we ablate the projection of the residual stream change vector on the representative direction of the latents of the second method. We aim by this experiment to remove the influence of the second methods' latents completely to see if the first method latents are still able to produce positive logit margin change.

We report in Table 4 the total logit margin change aggregated over the seven language per method pair for `Gemma-2-2B` and `Qwen3-4B` on **Antonyms**. The generally positive numbers indicate a form of robustness, however the asymmetry of the results for the same pair of methods require further investigation. We provide in the Appendix (Section J) additional observations related to the cosine similarity of the latent set representative residual-stream directions, and to latent annotations.

Table 4. Aggregate knock-out experiment results over all languages for `Gemma-2-B` and `Qwen3-4B`, where the latents of one selection method for the target language are amplified and the latents of the second method are used for direction ablation.

	Amplification method	Ablation method	Total Δm_l
Gemma-2-2B	AnnSel	FreqSel	5.20
	AnnSel	ValSel	38.82
	FreqSel	AnnSel	90.63
	FreqSel	ValSel	**92.16**
	ValSel	AnnSel	49.52
	ValSel	FreqSel	-1.19
Qwen3-4B	AnnSel	FreqSel	1.28
	AnnSel	ValSel	7.58
	FreqSel	AnnSel	**155.29**
	FreqSel	ValSel	85.82
	ValSel	AnnSel	107.55
	ValSel	FreqSel	33.86

5. Conclusion

We presented a comparative study of three methods for identifying language-specific latents in cross-layer transcoders: value-based (`ValSel`), frequency-based (`FreqSel`), and annotation-based (`AnnSel`). Through experiments on the controlled **Antonyms** and **Enumerations** tasks in seven languages, we demonstrated that all three methods can effectively steer generation language. `FreqSel` achieves the strongest overall performance, particularly for Asian languages, while `AnnSel` provides the added benefit of interpretable latent selection.

Future work could look into generalizing the findings of this work to larger models and to more languages. Real-world code-switching scenarios may present different challenges to the studied controlled settings. Deeper understanding of to what extent different methods identify similar or different latents has implications for both understanding multilingual representations and developing robust language-steering interventions.

References

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Ben Thompson, T., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Andrylie, L. M., Rahmanisa, I., Ihsani, M. K., Wicaksono, A. F., Wibowo, H. A., and Aji, A. F. Sparse autoencoders can capture language-specific concepts across diverse languages, 2025. URL https://arxiv.org/abs/2507.11230.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Deng, B., Wan, Y., Yang, B., Zhang, Y., and Feng, F. Unveiling language-specific features in large language models via sparse autoencoders. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4563–4608, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.229. URL https://aclanthology.org/2025.acl-long.229/.

Fiotto-Kaufman, J., Loftus, A. R., Todd, E., Brinkmann, J., Pal, K., Troitskii, D., Ripa, M., Belfki, A., Rager, C., Juang, C., Mueller, A., Marks, S., Sharma, A. S., Lucchetti, F., Prakash, N., Brodley, C., Guha, A., Bell, J., Wallace, B. C., and Bau, D. Nnsight and ndif: Democratizing access to open-weight foundation model internals, 2025. URL https://arxiv.org/abs/2407.14561.

Goncharov, A., Kondusov, N., and Zaytsev, A. Language steering in latent space to mitigate unintended code-switching, 2025. URL https://arxiv.org/abs/2510.13849.

Hanna, M. Qwen3-4b transcoders, 2025. URL https://huggingface.co/mwhanna/qwen3-4b-transcoders. Huggingface.

Hanna, M., Piotrowski, M., Lindsey, J., and Ameisen, E. circuit-tracer. https://github.com/safety-research/circuit-tracer, 2025. The first two authors contributed equally and are listed alphabetically.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R., and Nanda, N. Gemma scope: Open

sparse autoencoders everywhere all at once on gemma 2. In Belinkov, Y., Kim, N., Jumelet, J., Mohebbi, H., Mueller, A., and Chen, H. (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Marchisio, K., Ko, W.-Y., Bérard, A., Dehaze, T., and Ruder, S. Understanding and mitigating language confusion in llms. *arXiv preprint arXiv:2406.20052*, 2024.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07335-x. URL https://doi.org/10.1038/s41586-024-07335-x.

Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, W. X., Wei, F., and Wen, J.-R. 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)*, pp. 5701–5715, 2024.

A. `AnnSel` implementation details

The implementation of `AnnSel` can be decomposed into three steps: attribution graph construction, path pruning, and annotation-based latent selection.

A.1. Attribution graph construction

Since attribution graphs explain how the model predicts the next token, given a sentence from Flores+, we randomly cut the sentence in the middle so that the next token that the model should predict is a valid word in the target language, i.e. not a number or punctuation.

We used implementation from Circuit Tracer (Hanna et al., 2025) with parameters: max_n_logits $= 5$, desired_logit_prob $= 0.95$, max_feature_nodes $=$ None. It means the attribution graphs are computed to explain the top n logits, where $n = \min(5, m)$ and m is the smallest integer that satisfies $\sum_{i=1}^{m} prob_i \geq 0.95$. We then pruned graph with node_threshold $= 0.8$, edge_threshold $= 0.98$, meaning keeping nodes that explain 80% of total influence on the output logits and edges that explain 98% of total influence on the output logits.

A.2. Path Pruning

After computing the attribution graph, we pruned the graphs to prevent unimportant features from being included in the language supernodes. From each token to the top logit, we select important paths as follows. We first order the edges by the edge weight in descending order and keep choosing edges that are incident to already chosen nodes, starting from the token. When we reach the top logit, we can form a path from token to logit. Then, we remove that chosen path and start again. We keep doing this until either the edge weight chosen is less than or equal to 0.1 or we choose 75 distinct paths. We do this for every token position and obtain a set of paths.

After obtaining important paths, we further choose more important paths by setting threshold on the weights of the first edge and last edge since they represent the importance of the chosen paths in terms of the token and logit respectively and thus directly involved in the information flow from token to logit. We choose paths whose first edge has more than 0.5 edge weight and last edge has more than 0.25 edge weight. Then, out of the chosen features, we extracted all the features in the last token position and counted how many times each of them appear.

A.3. Annotation-based latent selection

After obtaining the set of features and the number of times they appear in the important paths, we first extract the description of each feature from Neuronpedia. We then remove all the features that does not mention the language name or highly relevant country name, e.g. ["Japanese", "japanese", "Japan", "japan"] for Japanese.

Then, what we have is a list of features that includes language name or country name in its description and their frequency. Let us denote the list by ℓ and the frequency of feature f by $freq_f$. We then choose f that satisfies $freq_f \geq 0.1 \max_{k \in \ell} freq_k$.

B. Applying SAE feature selection methods on CLT

We applied two methods, `ValSel` and `FreqSel`, from (Deng et al., 2025) and (Andrylie et al., 2025) respectively on the CLT. Since the original paper executes these methods on SAE, in order to justify the effectiveness of the methods on CLT, we conducted two experiments proposed in (Deng et al., 2025) and (Andrylie et al., 2025).

B.1. `ValSel` and code-switching experiment

(Deng et al., 2025) confirms the specificity of the language specific features by comparing the activation values on normal context and code-switching context. Following their experiments, we used the code-switching dataset of theirs and plotted the activation values of the top-50 features found and used for steering in our experiments. Keep in mind that we plotted the activations for all the languages used in our experiment except German because (Deng et al., 2025) did not include German in the experiments. Also, since we conducted the experiments on the CLT, the extracted features are not distributed uniformly across layers. The trend is the same as what is claimed

in (Deng et al., 2025) and the activation is generally highest in the original context and noun (Lang A Prefix + Lang A Noun), somewhat active in original context and modified noun (Lang A Prefix + Lang B Noun), and close to zero in modified noun (Lang B Noun).

B.2. `FreqSel` and text-generation experiment

(Andrylie et al., 2025) confirms the downstream effectiveness of the features by steering and doing unconditional text-generation. Their steering is $\boldsymbol{x}_{new} = \boldsymbol{x} + \alpha \boldsymbol{z}_j^{max} \boldsymbol{d}^j$, where $\alpha \in \mathbb{R}$ is a scaling factor and $\boldsymbol{z}_j^{max}$ is the maximum value of the feature of interest $\boldsymbol{d}^j$ in the multilingual corpora. We used the same multilingual corpora used to determine the language specific features when deciding on $\boldsymbol{z}_j^{max}$. The model was prompted with [BOS] and decoded with top-p sampling. Some examples are shown in Figure 5. The output is steered toward the target languages although the sentences might not be perfectly natural.

Table 5. Examples of unconditional text generation given [BOS] ($p = 0.0$ means greedy decoding)

Intervened Language	α	p	Output
de	0.8	0.9	package Produktion Kleidung refroidissement nôtre enfans hindurchholen nôtre bissunter Bedürf Flüssigkeit geben Flüssigkeit Kleidung
en	0.1	0.0	How to get the value of a variable in a function in python\n\n
es	0.8	0.9	How convert Activity anuales in española y y argentina\n pre acuerdos efectivos
fr	0.4	0.8	Que efficaces financières russes russes efficaces financières financières efficaces financières efficaces financières efficaces
ja	0.3	0.0	How toget指定したフォルダ内のファイル一覧をスクスクスクスク
ko	0.5	0.0	How to 3D 렌샷 것이다 렌샷 것이다 렌
zh	0.5	0.9	The特平在道路上有什用呢？什不把柏的

C. Latent counts

Table 6 shows the number of extracted token per method and language pair for `Gemma-2-2B` and `Qwen3-4B`. Note that `ValSel` extracts 50 latents by design, while the other two methods return a variable number of latents that relates to the CLT training data.

Table 6. Count of extracted latents per method and language for `Gemma-2-2B` and `Qwen3-4B`.

		en	fr	de	es	zh	ko	ja
Gemma	ValSel	50	50	50	50	50	50	50
	FreqSel	101	423	579	509	268	938	234
	AnnSel	17	41	59	19	41	7	45
Qwen	ValSel	50	50	50	50	50	50	50
	FreqSel	11	438	500	299	136	366	293
	AnnSel	12	16	13	10	16	8	13

D. Ablation intervention details

We tried several types of ablations. The one mentioned earlier is direction ablation.

D.1. Distractor zero ablation

Using Circuit Tracer (Hanna et al., 2025), we find the language supernodes as described above. In context zero ablation, we ablate the distractor language A by setting all the latents in the language A supernode to zero. In non-target zero ablation, we ablate all the languages except the target language B by setting all the latents in the language C supernode to zero, where C is all the languages except B.

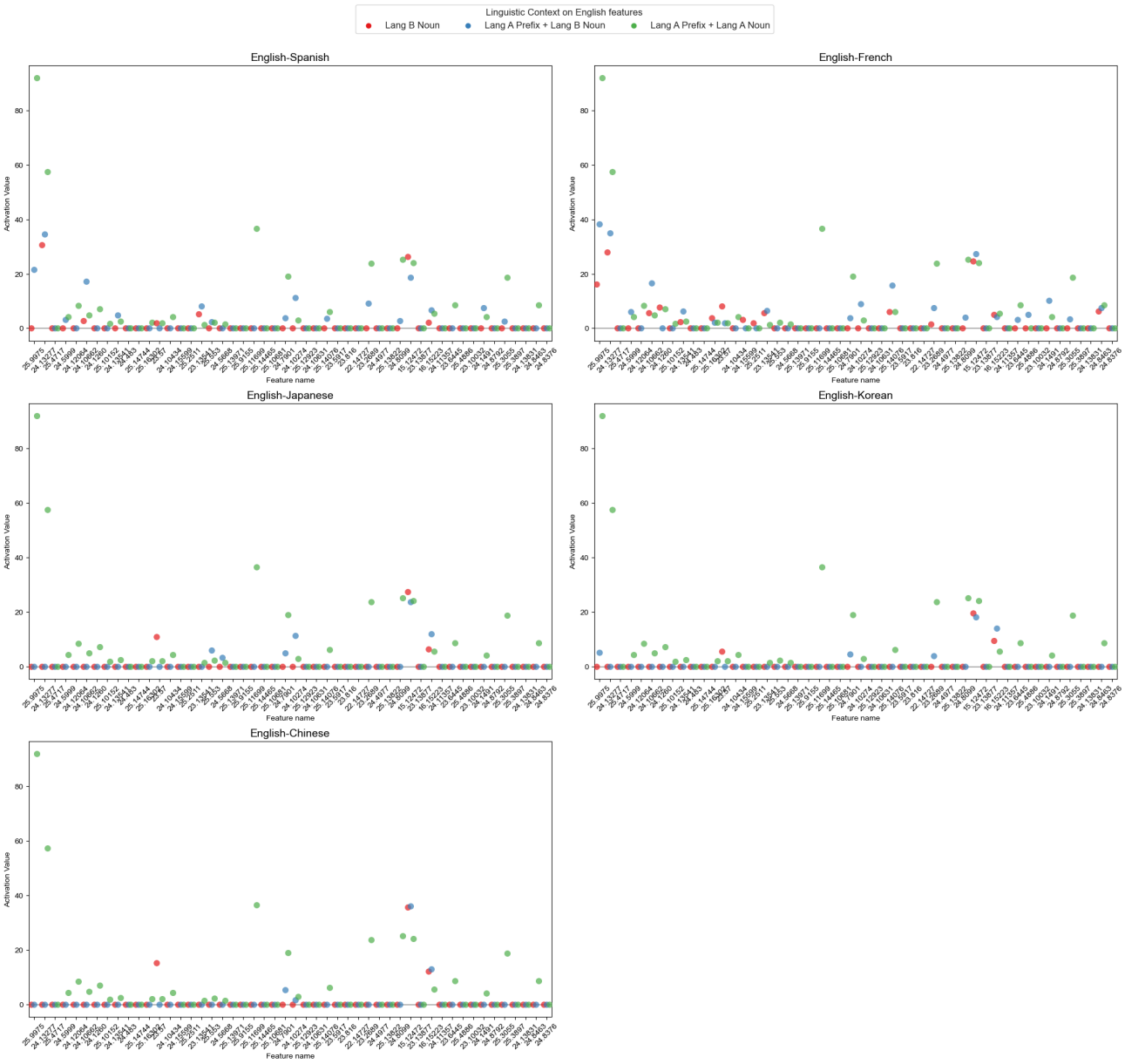

Figure 1. Activation values of English features on other language context and nouns

Figure 2. Activation values of French features on other language context and nouns

Figure 3. Activation values of Spanish features on other language context and nouns

Figure 4. Activation values of Chinese features on other language context and nouns

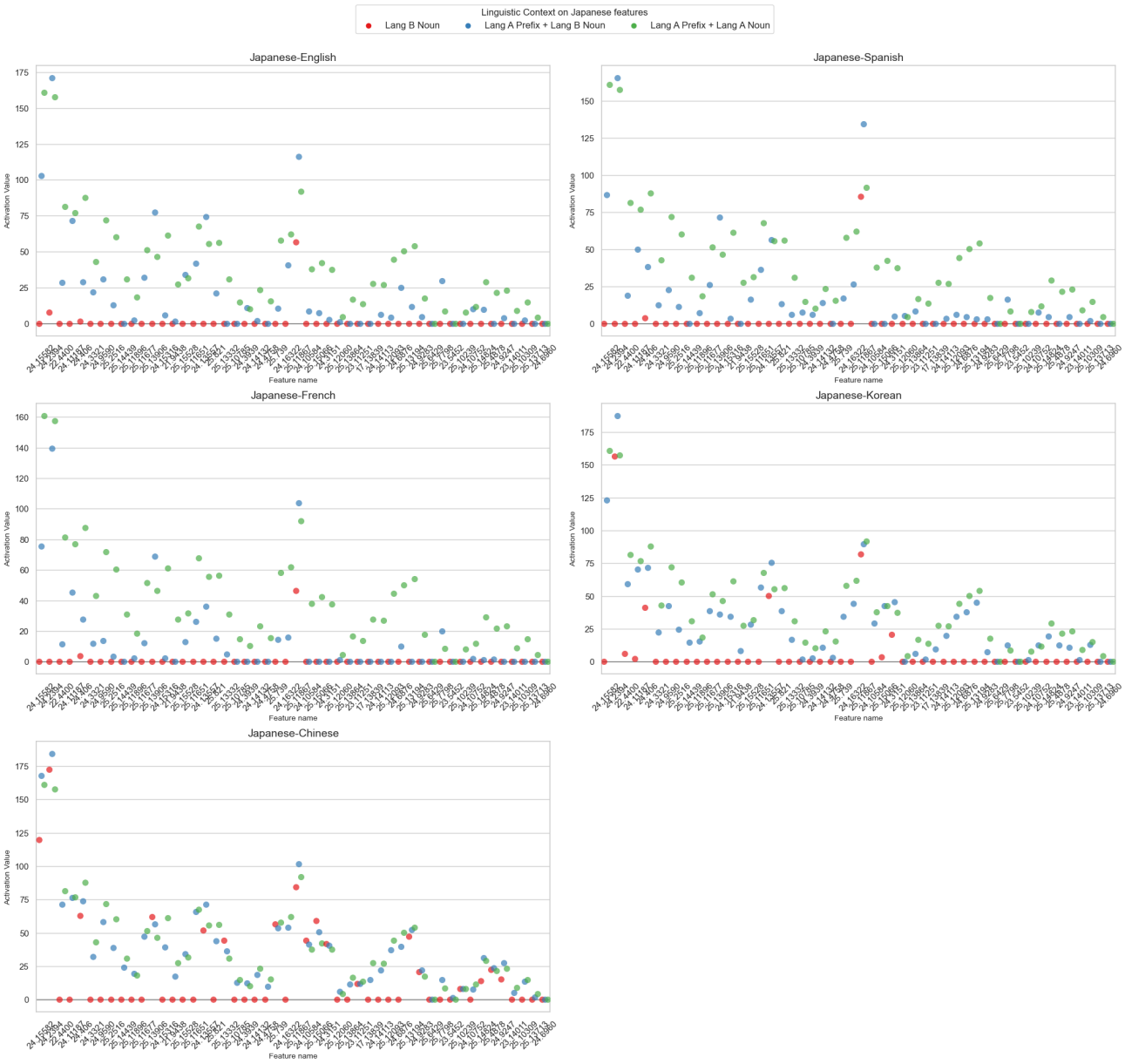

Figure 5. Activation values of Japanese features on other language context and nouns

Figure 6. Activation values of Korean features on other language context and nouns

D.2. One-layer direction ablation

The language features identified span across several layers, mostly in the last couple layers. Therefore, we chose the layer with the largest number of features for each language node with tie-breaking by the later layer, and did a direction ablation on that layer. Specifically, for each prompt, we run the attribution process to extract the activation values for the features in that layer and calculate the negative values we should insert into the features in the language node to completely get rid of the direction of the target features. Distractor one-layer direction ablation negates only one language, and one-layer direction ablation negates all the languages except one language.

The limited effectiveness of this approach might be due to the skip connections in CLTs and getting rid of the effects of one layer does not prevent other layers from influencing the logits through skip connections.

D.3. Multi-layer direction ablation

Instead of just getting rid of influence in one layer, we get rid of influence in all the layers. For all the features in a language supernode, we get the decoding weight matrices from CLT. Then, for each layer, we identify which direction to ablate using the weight matrices extracted based on which language features are in that layer. We finally do ablation in the activation space by extracting the projection onto that direction from the original activation. Since circuit tracer is a tool to manipulate the models in latent space, we used NNSight (Fiotto-Kaufman et al., 2025) instead to do ablation in activation space. Distractor (multi-layer) direction ablation intervenes on only one language, and (multi-layer) direction ablation intervenes on all the languages except one language.

Direction ablation works when the prompt language is either Chinese, Korean, or Japanese and when the adjective language is either German or Spanish. It does not work, for example, when the adjective language is French. When the prompt language is Chinese, Korean, or Japanese with French as the adjective language, the direction ablation pushes English and Spanish higher than French while successfully pushes the prompt languages down. This could suggest that a large part of French direction overlaps with Spanish and English directions.

Note that because multi-layer direction ablation operates in activation space and is implemented in NNSight (Fiotto-Kaufman et al., 2025), it is hard to combine it with non-distractor amplification, which operates in latent space and is implemented in circuit tracer (Hanna et al., 2025). Hence, we do not have multi-layer direction intervention.

E. Zero+Amp Detailed Results

Tables 13 to 15 show the logit difference results of `ValSel`, `FreqSel` and `AnnSel`, respectively, before and after the **Zero+Amp** intervention. The numbers in the table are for *top-1 logit margin* to the target adjective, which is the separation in the logit space of the target token to its best competitor. A value ≤ 0 (green) indicates that the target adjective is the most likely token, while a positive value (red) means it lags behind other undesired tokens.

Specifically, rows show the context language and columns show the adjective language. For each language pair, the 'before' value shows the logit margin before any intervention, while 'after' is its value after the latents amplification is applied. The desired outcome is that the value becomes smaller (greener) after intervention. Red cells indicate failures.

While the three methods show reasonable efficacy in enforcing Spanish and Chinese, `FreqSel` is weaker in enforcing German and English, and `ValSel` shows fragility when the context language is Chinese (Table 13). The situation of French adjectives is unique where `ValSel` and `FreqSel` fail for some context languages while `AnnSel` shows better robustness. On the other hand, `AnnSel` shows fragility in Korean (where only 7 latents where identified (Table 6) — far fewer than for any other language, while `FreqSel` shows the most robust performance.

F. Detailed Intervention Strategy Comparison

Next, we provide detailed results for the different interventions under `Gemma-2-2B` and the **Antonyms** task in Tables 7 to 9.

Table 7. Intervention effectiveness per target language for `ValSel` latents of `Gemma-2-2B`

	Distractor (l_0) zero ablation	$L/\{l\}$ zero ablation	Distractor (l_0) one-layer direction ablation	Distractor (l_0) multi-layer direction ablation	Target (l) amplification	Zero+Amp	1L+Amp
de	-3.18	3.12	-0.46	-3.15	14.52	13.6	13.92
en	7.49	5.06	2.77	52.07	5.88	10.16	4.24
es	0.25	3.06	-0.06	-1.57	13.35	14.19	15.08
fr	-0.77	1.15	-0.28	-9.65	8.74	9.01	8.71
ja	-4.61	-1.57	-0.41	-11.59	3.11	4.62	5.09
ko	-0.08	12.13	-1.61	-9.4	10.38	12.19	7.58
zh	-7.12	-0.56	2.04	-2.73	13.59	16.06	15.11
Total	-8.02	22.39	1.99	13.98	69.57	**79.83**	<u>69.73</u>

Table 8. Intervention effectiveness per target language for `FreqSel` latents of `Gemma-2-2B`

	Distractor (l_0) zero ablation	$L/\{l\}$ zero ablation	Distractor (l_0) one-layer direction ablation	Distractor (l_0) multi-layer direction ablation	Target (l) amplification	Zero+Amp	1L+Amp
de	-3.45	0.8	-1.67	0.58	5.08	7.86	6.29
en	5.86	5.43	3.6	-0.21	-0.55	4.36	1.94
es	-0.7	3.39	-0.6	-5.12	6.68	8.95	7.09
fr	-1.51	0.65	-0.78	-5.33	4.45	6.45	5.09
ja	-0.97	1.38	-1.45	-27.91	8.65	8.65	9.49
ko	4.82	6.62	-3.3	0.06	29.54	28.31	27.6
zh	-2.03	3.12	1.25	-13.69	36.98	34.49	38.66
Total	2.02	21.39	-2.95	-51.62	90.83	**99.07**	<u>96.16</u>

Table 9. Intervention effectiveness per target language for `AnnSel` latents of `Gemma-2-2B`

	Distractor (l_0) zero ablation	$L/\{l\}$ zero ablation	Distractor (l_0) one-layer direction ablation	Distractor (l_0) multi-layer direction ablation	Target (l) amplification	Zero+Amp	1L+Amp
de	-3.87	0.68	-0.04	-3.62	20.15	19.68	20.6
en	4.35	4.83	0.82	43.07	1.12	4.88	1.74
es	-0.04	1.98	-0.15	6.6	13.91	13.86	14.05
fr	-1.01	0.62	0.02	-3.71	13.69	13.54	13.95
ja	-2.21	1.5	-1.01	-4.97	11.17	11.12	10.53
ko	6.76	5.97	0.44	12.19	4.27	9.13	4.81
zh	-4.18	-3.68	0.22	14.49	22.69	24.51	23.65
Total	-0.20	11.90	0.30	64.05	87.00	**96.72**	<u>89.33</u>

G. Detailed Selection Method Comparison

We show detailed results for `Gemma-2-2B` under the three selection methods and the **Zero+Amp** intervention with **Antonyms** in Tables 13 to 15.

Table 16 presents summary results for selection method efficacy for `Qwen3-4B` latents for **Antonyms** and **Enumerations**.

Table 10. Intervention effectiveness per target language for `ValSel` latents of `Qwen3-4B`

	Distractor (l_0) zero ablation	$L/\{l\}$ zero ablation	Distractor (l_0) one-layer direction ablation	Distractor (l_0) multi-layer direction ablation	Target (l) amplification	Zero+Amp	1L+Amp
de	5.24	0.83	-4.32	-21.71	4.58	8.47	4.99
en	1.05	6.6	-2.81	2.16	-14.3	-6.17	-14.71
es	3.35	1.7	-1.4	-20.7	2.79	7.4	4.44
fr	3.51	-0.23	-2.79	-16.11	2.85	7.51	4.05
ja	3.33	-4.71	-6.52	-19.02	-6.02	-1.2	-4.12
ko	18.59	2.5	-2.11	-5.64	3.44	11.81	7.29
zh	5.47	8.84	2.67	17.1	7.51	17.74	13.9
Total	<u>40.54</u>	15.53	-17.28	-63.92	0.85	**45.56**	15.84

Table 11. Intervention effectiveness per target language for `FreqSel` latents of `Qwen3-4B`

	Distractor (l_0) zero ablation	$L/\{l\}$ zero ablation	Distractor (l_0) one-layer direction ablation	Distractor (l_0) multi-layer direction ablation	Target (l) amplification	Zero+Amp	1L+Amp
de	-2.92	0.15	-0.28	-24.44	4.69	6.35	5.08
en	6.67	8.39	1.19	-9.42	-2.01	6.11	-1.55
es	1.41	3.67	0.35	-21.37	2.85	4.92	2.93
fr	-0.96	2.51	-0.3	-16.97	5.05	6.88	5.18
ja	-3.3	0.18	-3.63	-37.57	-0.27	0.37	-1.25
ko	3.99	7.29	-0.25	-30.15	19.97	21.29	20.5
zh	11.66	11.86	7.94	-25.09	26.47	29.76	27.71
Total	16.55	34.05	5.02	-165.01	56.75	**75.68**	<u>58.60</u>

Table 12. Intervention effectiveness per target language for `AnnSel` latents of `Qwen3-4B`

	Distractor (l_0) zero ablation	$L/\{l\}$ zero ablation	Distractor (l_0) one-layer direction ablation	Distractor (l_0) multi-layer direction ablation	Target (l) amplification	Zero+Amp	1L+Amp
de	-0.67	0	0.02	-2.51	6.73	6.49	6.89
en	4.62	4.9	0.6	8.84	0	0	0
es	1.2	0	0.04	-2.58	6.31	7.28	6.41
fr	-0.71	0	-0.02	-6.21	6.5	6.35	6.56
ja	0.4	0	-0.18	-13.1	1.76	1.83	1.72
ko	0.12	0	-2.42	-5	5.6	7.7	6.08
zh	3.46	0	3.08	-0.4	6.39	9.78	9.73
Total	8.42	4.9	1.12	-20.96	33.29	**39.43**	<u>37.39</u>

Table 13. Logit difference for `ValSel` and `Gemma-2-2B` before and after the intervention using **Antonyms**

	es		en		zh		de		ja		fr		ko	
	before	after	before	after	before	after	before	after	before	after	before	after	before	after
de	-1.79	-2.93	-4.57	-5.19	-2.97	-4.88	-	-	-2.62	-2.69	-0.24	-0.98	-2.38	-2.21
en	-2.29	-3.29	-	-	-3.77	-4.41	-1.45	-2.35	-2.96	-2.71	-0.10	-0.66	-2.87	-3.27
es	-	-	-4.15	-5.83	-3.77	-4.56	-1.30	-2.68	-2.27	-2.58	-0.08	-2.03	-2.30	-2.25
fr	-1.22	-3.19	-4.13	-5.08	-1.89	-4.22	-1.02	-2.26	-2.48	-2.65	-	-	-2.22	-2.18
ja	0.95	-1.87	-3.54	-5.90	0.34	-4.45	0.87	-1.73	-	-	1.65	0.25	1.69	-3.29
ko	1.54	-2.37	-3.28	-4.40	-0.44	-6.04	0.80	-2.53	-1.98	-2.54	1.75	-0.61	-	-
zh	2.62	-0.73	-2.29	-5.72	-	-	3.38	-0.77	2.25	-1.52	2.80	0.80	5.79	-1.27
Total	-0.19	-14.38	-21.96	-32.12	-12.50	-28.56	1.28	-12.32	-10.06	-14.69	5.78	-3.23	-2.29	-14.47
Change		+14.19		+10.16		+16.06		+13.6		+4.63		+9.01		+12.18

Table 14. Logit difference for `FreqSel` and `Gemma-2-2B` before and after the intervention using **Antonyms**

	es		en		zh		de		ja		fr		ko	
	before	after	before	after	before	after	before	after	before	after	before	after	before	after
de	-1.79	-1.59	-4.57	-4.97	-2.97	-7.58	-	-	-2.62	-3.13	-0.24	-0.18	-2.38	-5.82
en	-2.29	-1.55	-	-	-3.77	-7.99	-1.45	-1.17	-2.96	-3.34	-0.10	0.16	-2.87	-5.73
es	-	-	-4.15	-4.62	-3.77	-8.23	-1.30	-1.03	-2.27	-3.21	-0.08	-0.26	-2.30	-5.61
fr	-1.22	-1.75	-4.13	-4.70	-1.89	-7.79	-1.02	-1.04	-2.48	-3.30	-	-	-2.22	-5.90
ja	0.95	-1.45	-3.54	-4.06	0.34	-7.06	0.87	-1.05	-	-	1.65	0.19	1.69	-3.99
ko	1.54	-1.85	-3.28	-4.26	-0.44	-8.35	0.80	-1.30	-1.98	-3.13	1.75	-0.59	-	-
zh	2.62	-0.95	-2.29	-3.69	-	-	3.38	-0.98	2.25	-2.60	2.80	0.01	5.79	-3.54
Total	-0.19	-9.14	-21.96	-26.30	-12.50	-47.00	1.28	-6.57	-10.06	-18.71	5.78	-0.67	-2.29	-30.59
Change		+8.95		+4.34		+34.5		+7.85		+8.65		+6.45		+28.3

Table 15. Logit difference for `AnnSel` and `Gemma-2-2B` before and after the intervention using **Antonyms**

	es		en		zh		de		ja		fr		ko	
	before	after	before	after	before	after	before	after	before	after	before	after	before	after
de	-1.79	-3.04	-4.57	-5.32	-2.97	-5.67	-	-	-2.62	-3.63	-0.24	-1.67	-2.38	-2.54
en	-2.29	-3.11	-	-	-3.77	-6.46	-1.45	-3.01	-2.96	-3.69	-0.10	-1.04	-2.87	-2.57
es	-	-	-4.15	-4.84	-3.77	-6.52	-1.30	-3.37	-2.27	-3.62	-0.08	-2.26	-2.30	-2.50
fr	-1.22	-3.06	-4.13	-4.88	-1.89	-5.91	-1.02	-3.23	-2.48	-3.74	-	-	-2.22	-2.49
ja	0.95	-1.62	-3.54	-4.69	0.34	-6.00	0.87	-2.71	-	-	1.65	-0.58	1.69	-2.07
ko	1.54	-2.36	-3.28	-3.33	-0.44	-6.46	0.80	-3.55	-1.98	-3.53	1.75	-1.61	-	-
zh	2.62	-0.86	-2.29	-3.77	-	-	3.38	-2.52	2.25	-2.97	2.80	-0.60	5.79	0.77
Total	-0.19	-14.05	-21.96	-26.83	-12.50	-37.02	1.28	-18.39	-10.06	-21.18	5.78	-7.76	-2.29	-11.40
Change		+13.86		+4.87		+24.52		+19.67		+11.12		+13.54		+9.11

Table 16. Selection method efficacy per language under the **Zero+Amp** intervention for `Qwen3-4B`. Post-intervention total change in top-1 logit margin for **Antonyms**; Bottom: average change in normalized logProb of the target sequence for **Enumerations**.

		es	en	zh	de	ja	fr	ko	Avg
Antons.	`ValSel`	**7.36**	-5.31	17.56	**6.38**	-1.17	**7.49**	11.29	6.23
	`FreqSel`	4.90	**6.09**	**29.87**	4.13	-7.34	6.90	**20.86**	**9.34**
	`AnnSel`	7.35	5.12	9.84	4.95	**2.23**	6.95	6.79	6.17
Enums.	`ValSel`	0.89	-0.73	0.25	0.95	0.40	0.97	0.61	0.48
	`FreqSel`	**1.50**	**0.28**	**1.41**	**1.90**	**1.26**	**1.49**	**1.81**	**1.38**
	`AnnSel`	0.73		0.17	0.82	0.18	0.86	0.49	0.46

H. Adjectives Used in Antonyms

We prepared 100 pairs of antonyms to use for the antonym-structured sentences (*The opposite of "adj1" is ", adj2*). Keep in mind that there could be multiple suitable words for *adj2* (e.g. the antonym of small could be either big or large). We used Gemini 2.5 to generate them. All the antonyms are shown in Table 17. It is worth mentioning that the words or tokens can be exactly the same in some cases across different languages, most commonly between French and Spanish and between Chinese and Japanese.

Table 17. Multilingual Antonym Pairs in the 7 languages

English	French	German	Spanish	Chinese	Japanese	Korean
good *bad*	**bon** *mauvais*	**gut** *schlecht*	**bueno** *malo*	**好** 坏	**良い** 悪い	**좋은** *나쁜*
happy *sad, unhappy*	**heureux** *triste, malheureux*	**glücklich** *traurig, unglücklich*	**feliz** *triste, infeliz*	**开心** 难过, 不高兴	**嬉しい** 悲しい	**행복한** *슬픈, 불행한*
big *small*	**grand** *petit*	**groß** *klein*	**grande** *pequeño*	**大** 小	**大きい** 小さい	**큰** *작은*
hot *cold*	**chaud** *froid*	**heiß** *kalt*	**caliente** *frío*	**热** 冷	**暑い** 寒い, 冷たい	**더운** *추운, 차가운*
fast *slow*	**rapide** *lent, lente*	**schnell** *langsam*	**rápido** *lento*	**快** 慢	**速い** 遅い	**빠른** *느린*
light *heavy, dark*	**léger** *lourd*	**leicht** *schwer*	**ligero** *pesado*	**轻** 重	**軽い** 重い	**가벼운** *무거운*
easy *difficult, hard*	**facile** *difficile*	**einfach** *schwierig, schwer*	**fácil** *difícil*	**容易** 难	**簡単な** 難しい	**쉬운** *어려운*
new *old*	**nouveau** *vieux, ancien*	**neu** *alt*	**nuevo** *viejo*	**新** 旧	**新しい** 古い	**새로운** *오래된*
true *false, untrue*	**vrai** *faux*	**wahr** *falsch, unwahr*	**verdadero** *falso, incorrecto*	**真** 假	**正しい** 間違った	**진실한** *거짓의, 틀린*
alive *dead*	**vivant** *mort*	**lebendig** *tot*	**vivo** *muerto*	**活** 死	**生きている** 死んでいる	**살아있는** *죽은*
full *empty*	**plein** *vide*	**voll** *leer*	**lleno** *vacío*	**满** 空	**満杯** 空っぽ	**가득 찬** *빈*
bright *dark, dim*	**brillant** *sombre, obscur*	**hell** *dunkel*	**brillante** *oscuro, opaco*	**亮** 暗	**明るい** 暗い	**밝은** *어두운, 흐릿한*
strong *weak*	**fort** *faible*	**stark** *schwach*	**fuerte** *débil*	**强** 弱	**強い** 弱い	**강한** *약한*
clean *dirty*	**propre** *sale*	**sauber** *schmutzig*	**limpio** *sucio*	**干净** 脏	**きれいな** 汚い	**깨끗한** *더러운*
open *closed*	**ouvert** *fermé*	**offen** *geschlossen*	**abierto** *cerrado*	**开** 关	**開いた** 閉じた	**열린** *닫힌*
rich *poor*	**riche** *pauvre*	**reich** *arm*	**rico** *pobre*	**富裕** 贫穷	**裕福な** 貧しい	**부유한** *가난한*
beautiful *ugly*	**beau** *laid*	**schön** *hässlich*	**hermoso** *feo*	**美** 丑	**美しい** 醜い	**아름다운** *못생긴*
long *short*	**long** *court*	**lang** *kurz*	**largo** *corto*	**长** 短	**長い** 短い	**긴** *짧은*
wide *narrow*	**large** *étroit*	**breit** *eng*	**ancho** *estrecho*	**宽** 窄	**広い** 狭い	**넓은** *좁은*
hard *soft*	**dur** *mou*	**hart** *weich*	**duro** *suave*	**硬** 软	**硬い** 柔らかい	**딱딱한** *부드러운*
dry *wet, moist*	**sec** *humide, mouillé*	**trocken** *nass*	**seco** *mojado, húmedo*	**干** 湿	**乾いた** 濡れた, 湿った	**마른** *젖은, 축축한*
loud *quiet, silent*	**fort** *silencieux, doux*	**laut** *leise, still*	**ruidoso** *tranquilo, silencioso*	**大声** 安静, 小声	**うるさい** 静かな	**시끄러운** *조용한, 고요한*

English	French	German	Spanish	Chinese	Japanese	Korean
early *late*	**tôt** *tard*	**früh** *spät*	**temprano** *tarde*	**早** 晚	**早い** 遅い	**이른** 늦은
near *far*	**proche** *loin*	**nah** *fern, weit*	**cerca** *lejos*	**近** 远	**近い** 遠い	**가까운** 먼
deep *shallow*	**profond** *peu profond*	**tief** *flach*	**profundo** *superficial, poco profundo*	**深** 浅	**深い** 浅い	**깊은** 얕은
bad *good*	**mauvais** *bon*	**schlecht** *gut*	**malo** *bueno*	**坏** 好	**悪い** 良い	**나쁜** 좋은
sad *happy, joyful*	**triste** *heureux, joyeux*	**traurig** *glücklich*	**triste** *feliz, alegre*	**难过** 开心	**悲しい** 嬉しい	**슬픈** *행복한, 기쁜*
small *big, large*	**petit** *grand*	**klein** *groß*	**pequeño** *grande*	**小** 大	**小さい** 大きい	**작은** 큰
cold *hot, warm*	**froid** *chaud*	**kalt** *heiß*	**frío** *caliente, cálido*	**冷** 热	**寒い** 暑い, 熱い	**추운** *더운, 따뜻한*
slow *fast, quick*	**lent** *rapide*	**langsam** *schnell*	**lento** *rápido, veloz*	**慢** 快	**遅い** 速い, 早い	**느린** 빠른
heavy *light*	**lourd** *léger*	**schwer** *leicht*	**pesado** *ligero*	**重** 轻	**重い** 軽い	**무거운** *가벼운*
difficult *easy, simple*	**difficile** *facile*	**schwierig** *einfach*	**difícil** *fácil, sencillo*	**难** 容易, 易	**難しい** 簡単な, 簡単	**어려운** *쉬운, 간단한*
old *new*	**vieux** *nouveau, jeune*	**alt** *neu*	**viejo** *nuevo, joven*	**旧** 新	**古い** 新しい	**오래된** *새로운, 젊은*
false *true, correct*	**faux** *vrai, correct*	**falsch** *wahr, richtig*	**falso** *verdadero, correcto*	**假** 真	**間違った** 正しい	**거짓의** *진실한, 올바른*
dead *alive, living*	**mort** *vivant*	**tot** *lebendig*	**muerto** *vivo, viviente*	**死** 活, 生	**死んでいる** 生きている	**죽은** *살아있는, 생생한*
empty *full*	**vide** *plein*	**leer** *voll*	**vacío** *lleno*	**空** 满	**空っぽ** 満杯	**빈** *가득 찬*
dark *bright, light*	**sombre** *brillant, clair*	**dunkel** *hell*	**oscuro** *brillante, claro*	**暗** 亮, 光明	**暗い** 明るい	**어두운** *밝은*
weak *strong*	**faible** *fort*	**schwach** *stark*	**débil** *fuerte*	**弱** 强, 强壮	**弱い** 強い, 丈夫な	**약한** *강한*
dirty *clean*	**sale** *propre*	**schmutzig** *sauber*	**sucio** *limpio*	**脏** 干净	**汚い** きれいな, きれい, 綺麗	**더러운** *깨끗한*
closed *open*	**fermé** *ouvert*	**geschlossen** *offen*	**cerrado** *abierto*	**关** 开	**閉じた** 開いた	**닫힌** 열린
poor *rich, wealthy*	**pauvre** *riche*	**arm** *reich*	**pobre** *rico, adinerado*	**贫穷** 富裕	**貧しい** 裕福な, 豊かな	**가난한** *부유한, 풍부한*
ugly *beautiful, pretty*	**laid** *beau, joli*	**hässlich** *schön*	**feo** *hermoso, bonito*	**丑** 美	**醜い** 美しい	**못생긴** *아름다운, 예쁜*
short *long*	**court** *long*	**kurz** *lang*	**corto** *largo*	**短** 长	**短い** 長い	**짧은** 긴
narrow *wide, broad*	**étroit** *large*	**eng** *breit*	**estrecho** *ancho*	**窄** 宽	**狭い** 広い	**좁은** 넓은

English	French	German	Spanish	Chinese	Japanese	Korean
soft *hard, firm*	**mou** *dur*	**weich** *hart*	**suave** *duro*	**软** 硬	**柔らかい** 硬い	**부드러운** *딱딱한*
wet *dry*	**humide** *sec*	**nass** *trocken*	**mojado** *seco*	**湿** 干	**濡れた** 乾いた	**젖은** *마른*
quiet *loud, noisy*	**silencieux** *fort, bruyant*	**leise** *laut*	**tranquilo** *ruidoso*	**安静** 大声, 吵闹	**静かな** うるさい	**조용한** *시끄러운*
late *early*	**tard** *tôt*	**spät** *früh*	**tarde** *temprano*	**晚** 早	**遅い** 早い	**늦은** *이른*
far *near, close*	**loin** *proche*	**fern** *nah*	**lejos** *cerca*	**远** 近	**遠い** 近い	**먼** *가까운*
shallow *deep*	**peu profond** *profond*	**flach** *tief, hoch*	**superficial** *profundo*	**浅** 深	**浅い** 深い	**얕은** *깊은*
young *old*	**jeune** *vieux*	**jung** *alt*	**joven** *viejo*	**年轻** 老	**若い** 老いた	**젊은** *늙은*
kind *unkind, mean*	**gentil** *méchant*	**freundlich** *unfreundlich*	**amable** *desagradable, malo*	**善良** 不善良, 邪恶	**親切な** 意地悪な, 不親切な, 失礼な	**친절한** *불친절한, 못된*
brave *cowardly*	**courageux** *lâche*	**mutig** *feige*	**valiente** *cobarde*	**勇敢** 懦弱	**勇敢な** 臆病な	**용감한** *겁많은*
wise *foolish*	**sage** *insensé, stupide*	**weise** *dumm, töricht*	**sabio** *tonto, necio*	**明智** 愚蠢	**賢い** 愚かな	**현명한** *어리석은*
polite *impolite, rude*	**poli** *impoli, grossier*	**höflich** *unhöflich, grob*	**educado** *descortés, maleducado*	**礼貌** 不礼貌	**丁寧な** 失礼な, 無礼な, 粗略な, 粗雑な	**예의 바른** *무례한*
patient *impatient*	**patient** *impatient*	**geduldig** *ungeduldig*	**paciente** *impaciente*	**耐心** 不耐烦	**我慢強い** せっかちな	**인내심 있는** *성급한, 참을성 없는*
honest *dishonest*	**honnête** *malhonnête*	**ehrlich** *unehrlich*	**honesto** *deshonesto*	**诚实** 不诚实	**正直な** 不正直な	**정직한** *부정직한*
safe *dangerous, unsafe*	**sûr** *dangereux*	**sicher** *gefährlich, unsicher*	**seguro** *peligroso, inseguro*	**安全** 危险	**安全な** 危険な	**안전한** *위험한*
active *inactive, passive*	**actif** *inactif, passif*	**aktiv** *inaktiv, passiv*	**activo** *inactivo, pasivo*	**积极** 消极	**活動的な** 消極的な	**활동적인** *비활동적인, 소극적인*
straight *curved, bent*	**droit** *courbe, tordu*	**gerade** *gebogen, krumm*	**recto** *curvo, doblado*	**直** 弯	**まっすぐな** 曲がった	**곧은** *굽은, 휘어진*
whole *part, broken*	**entier** *partiel, cassé*	**ganz** *teilweise, gebrochen*	**entero** *parcial, roto*	**完整** 部分, 破	**全体の** 部分的な, 壊れた	**전체의** *부분적인, 부서진*
calm *agitated, stormy*	**calme** *agité, orageux*	**ruhig** *aufgeregt, stürmisch*	**calmado** *agitado, tormentoso*	**平静** 激动, 骚乱	**穏やかな** 荒れた, 興奮した	**차분한** *격앙된, 폭풍우 치는*
correct *incorrect, wrong*	**correct** *incorrect, faux*	**richtig** *falsch*	**correcto** *incorrecto, equivocado*	**正确** 不正确, 错误	**正しい** 間違っている	**정확한** *부정확한, 틀린*
complex *simple*	**complexe** *simple*	**komplex** *einfach*	**complejo** *simple, sencillo*	**复杂** 简单	**複雑な** 単純な	**복잡한** *간단한, 단순한*
effective *ineffective*	**efficace** *inefficace*	**effektiv** *ineffektiv*	**efectivo** *ineficaz*	**有效** 无效	**効果的な** 非効果的な	**효과적인** *비효과적인*

English	French	German	Spanish	Chinese	Japanese	Korean
famous *unknown, obscure*	**célèbre** *inconnu, obscur*	**berühmt** *unbekannt, unbedeutend*	**famoso** *desconocido, oscuro*	**著名** 无名，不为人知	**有名な** 無名の，知られていない	**유명한** *무명의, 잘 알려지지 않은*
generous *stingy, mean*	**généreux** *avare, mesquin*	**großzügig** *geizig*	**generoso** *tacaño, malo*	**慷慨** 吝啬	**寛大な** ケチな	**관대한** *인색한, 못된*
content *unhappy, dissatisfied*	**content** *mécontent, insatisfait*	**zufrieden** *unzufrieden*	**contento** *infeliz, insatisfecho*	**高兴** 不高兴，不满意	**幸せな** 不幸せな	**행복한** *불행한, 불만족스러운*
healthy *unhealthy, sick*	**sain** *malsain, malade*	**gesund** *ungesund, krank*	**saludable** *insalubre, enfermo*	**健康** 不健康，生病	**健康な** 不健康な，病気の	**건강한** *건강하지 않은, 아픈*
high *low*	**haut** *bas*	**hoch** *niedrig*	**alto** *bajo*	**高** 低	**高い** 低い	**높은** *낮은*
important *unimportant, trivial*	**important** *sans importance, insignifiant*	**wichtig** *unwichtig, trivial*	**importante** *sin importance, trivial*	**重要** 不重要，琐碎	**重要な** 重要でない，取るに足らない	**중요한** *중요하지 않은, 사소한*
innocent *guilty*	**innocent** *coupable*	**unschuldig** *schuldig*	**inocente** *culpable*	**无辜** 有罪	**無罪の** 有罪の	**무고한** *유죄의*
known *unknown*	**connu** *inconnu*	**bekannt** *unbekannt*	**conocido** *desconocido*	**已知** 未知	**既知の** 未知の	**알려진** *알려지지 않은*
male *female*	**masculin** *féminin*	**männlich** *weiblich*	**masculino** *femenino*	**男性** 女性	**男性の** 女性の	**남성의** *여성의*
normal *abnormal, unusual*	**normal** *anormal, inusuel*	**normal** *abnormal, ungewöhnlich*	**normal** *anormal, inusual*	**正常** 异常，不正常	**普通の** 異常な	**정상적인** *비정상적인, 이상한*
possible *impossible*	**possible** *impossible*	**möglich** *unmöglich*	**posible** *imposible*	**可能** 不可能	**可能な** 不可能な	**가능한** *불가능한*
private *public*	**privé** *public*	**privat** *öffentlich*	**privado** *público*	**私人** 公共	**個人の** 公共の	**사적인** *공적인*
right *wrong*	**juste** *faux*	**richtig** *falsch*	**correcto** *incorrecto, equivocado*	**对** 错	**正しい** 間違っている	**올바른** *틀린*
simple *complex*	**simple** *complexe*	**einfach** *komplex*	**sencillo** *complejo*	**简单** 复杂	**簡単な** 複雑な	**단순한** *복잡한*
sweet *sour, bitter*	**doux** *acide, amer*	**süß** *sauer, bitter*	**dulce** *agrio, amargo*	**甜** 酸，苦	**甘い** 酸っぱい，苦い	**달콤한** *신, 쓴*
visible *invisible*	**visible** *invisible*	**sichtbar** *unsichtbar*	**visible** *invisible*	**可见** 不可见	**見える** 見えない	**보이는** *보이지 않는*
warm *cool, cold*	**chaud** *frais, froid*	**warm** *kühl, kalt*	**cálido** *fresco, frío*	**暖和** 凉，冷	**暖かい** 涼しい，冷たい	**따뜻한** *시원한, 차가운*
smooth *rough, bumpy*	**lisse** *rugueux, bosselé*	**glatt** *rau, holprig*	**liso** *áspero, irregular*	**光滑** 粗糙	**滑らかな** 粗い	**매끄러운** *거친, 울퉁불퉁한*
thin *thick, fat*	**mince** *épais*	**dünn** *dick*	**delgado** *grueso*	**薄** 厚	**薄い** 厚い	**얇은** *두꺼운*
tall *short*	**grand** *petit*	**groß** *klein*	**alto** *bajo*	**高** 矮	**背が高い** 背が低い	**키가 큰** *키가 작은*

English	French	German	Spanish	Chinese	Japanese	Korean
married *single, unmarried*	**marié** *célibataire*	**verheiratet** *ledig, unverheiratet*	**casado** *soltero*	**已婚** 单身, 未婚	**既婚の** 独身の, 未婚の	**결혼한** *독신의, 미혼의*
optimistic *pessimistic*	**optimiste** *pessimiste*	**optimistisch** *pessimistisch*	**optimista** *pesimista*	**乐观** 悲观	**楽観的な** 悲観的な	**낙관적인** *비관적인*
permanent *temporary*	**permanent** *temporaire*	**permanent** *temporär*	**permanente** *temporal*	**永久** 临时	**恒久的な** 一時的な	**영구적인** *일시적인*
present *absent, past*	**présent** *absent, passé*	**gegenwärtig** *abwesend, vergangen*	**presente** *ausente, pasado*	**现在** 缺席, 过去	**現在の** 不在の, 過去の	**현재의** *부재의, 과거의*
public *private*	**public** *privé*	**öffentlich** *privat*	**público** *privado*	**公共** 私人	**公共の** 個人の	**공공의** *사적인*
real *fake, unreal*	**réel** *faux, irréel*	**echt** *falsch, unreal*	**real** *falso, irreal*	**真实** 假, 虚假	**本物の** 偽物の, 非現実の	**실제적인** *가짜의, 비현실적인*
responsible *irresponsible*	**responsable** *irresponsable*	**verantwortlich** *unverantwortlich*	**responsable** *irresponsable*	**负责** 不负责	**責任** 無責任な, 無責任	**책임 있는** *무책임한*
single *married*	**célibataire** *marié*	**ledig** *verheiratet*	**soltero** *casado*	**单身** 已婚	**独身の** 既婚の, 結婚している	**독신의** *결혼한*
sour *sweet*	**acide** *doux*	**sauer** *süß*	**agrio** *dulce*	**酸** 甜	**酸っぱい** 甘い	**신** *달콤한*
useful *useless*	**utile** *inutile*	**nützlich** *nutzlos*	**útil** *inútil*	**有用** 没用	**役に立つ** 役に立たない	**유용한** *쓸모없는*
vertical *horizontal*	**vertical** *horizontal*	**vertikal** *horizontal*	**vertical** *horizontal*	**垂直** 水平	**垂直な** 水平な	**수직의** *수평의*
well *poorly*	**bien** *mal*	**gut** *schlecht*	**bien** *mal, pobremente*	**好** 差	**良い** 悪い	**잘** *못, 형편없이*
winning *losing*	**gagnant** *perdant*	**gewinnend** *verlierend*	**ganador** *perdedor*	**获胜** 失败	**勝利の** 敗北の	**승리하는** *패배하는*
unkind *kind*	**méchant** *gentil*	**unfreundlich** *freundlich*	**desagradable** *amable*	**不善良** 善良	**意地悪な** 親切な	**불친절한** *친절한*
cowardly *brave*	**lâche** *courageux*	**feige** *mutig*	**cobarde** *valiente*	**懦弱** 勇敢	**臆病な** 勇敢な	**겁많은** *용감한*

I. Listings Used in Enumerations

We prepared 7 ordered enumerations to use for the enumeration-structured sentences (*The {category} are: {choices}*). We used Gemini 3 to translate them into the seven languages. All the categories, listings, and the split of the listed choices and continuation words are shown in Table 18.

J. Additional Redundancy Analysis

To assess the equivalence of the latent sets identified by the three selection methods, we compute representative per-layer activation space directions that correspond to the decoder directions of these layer-wise latents, then calculate the pairwise cosine similarity between per-method directions. If the latents of two selection methods for the same language are identical or equivalent in their representation in a certain layer, we expect their corresponding residual stream directions to be similar (large cosine similarity), while a small cosine similarity in a certain layer might indicate divergence. We report in Table 19 for each pair of selection methods the minimum cosine similarity, as well as the across-layer average, for the language latents of `Gemma-2-2B` and `Qwen3-4B`.

Finally, we analyze the Neuronpedia annotations of the identified latent sets as a (potentially noisy) means to discern latent set overlap. We compute as a proxy the percentage of language latents whose Neuronpedia annotation refers explicitly to

the language. While this is 100% for `AnnSel` by design, the results for `ValSel` and `FreqSel` in Table 20 reveals substantial variability.

Table 18. Enumerations Dataset: categories, items and split thresholds.

Category ID	Split	Lang	Category Name	Listings
months	4|8	en	months of the year	January, February, March, April, May, June, July, August, September, October, November, December
		de	Monate des Jahres	Januar, Februar, März, April, Mai, Juni, Juli, August, September, Oktober, November, Dezember
		fr	mois de l'année	janvier, février, mars, avril, mai, juin, juillet, août, septembre, octobre, novembre, décembre
		es	meses del año	enero, febrero, marzo, abril, mayo, junio, julio, agosto, septiembre, octubre, noviembre, diciembre
		zh	一年中的月份	一月, 二月, 三月, 四月, 五月, 六月, 七月, 八月, 九月, 十月, 十一月, 十二月
		ja	一年の月	一月, 二月, 三月, 四月, 五月, 六月, 七月, 八月, 九月, 十月, 十一月, 十二月
		ko	1년의열두달	1월, 2월, 3월, 4월, 5월, 6월, 7월, 8월, 9월, 10월, 11월, 12월
numbers	4|6	en	numbers 1 to 10	one, two, three, four, five, six, seven, eight, nine, ten
		de	Zahlen von eins bis zehn	eins, zwei, drei, vier, fünf, sechs, sieben, acht, neun, zehn
		fr	nombres de un à dix	un, deux, trois, quatre, cinq, six, sept, huit, neuf, dix
		es	números del uno al diez	uno, dos, tres, cuatro, cinco, seis, siete, ocho, nueve, diez
		zh	从一到十的数字	一, 二, 三, 四, 五, 六, 七, 八, 九, 十
		ja	一から十までの数字	一, 二, 三, 四, 五, 六, 七, 八, 九, 十
		ko	1부터10까지의숫자	일, 이, 삼, 사, 오, 육, 칠, 팔, 구, 십
days_of_week	3|4	en	days of the week	Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday
		de	Wochentage	Sonntag, Montag, Dienstag, Mittwoch, Donnerstag, Freitag, Samstag
		fr	jours de la semaine	dimanche, lundi, mardi, mercredi, jeudi, vendredi, samedi
		es	días de la semana	domingo, lunes, martes, miércoles, jueves, viernes, sábado
		zh	一星期中的日子	星期日, 星期一, 星期二, 星期三, 星期四, 星期五, 星期六
		ja	曜日	日曜日, 月曜日, 火曜日, 水曜日, 木曜日, 金曜日, 土曜日
		ko	요일	일요일, 월요일, 화요일, 수요일, 목요일, 금요일, 토요일
four_seasons	2|2	en	four seasons	spring, summer, fall, winter
		de	vier Jahreszeiten	Frühling, Sommer, Herbst, Winter
		fr	quatre saisons	printemps, été, automne, hiver
		es	cuatro estaciones	primavera, verano, otoño, invierno
		zh	四季	春, 夏, 秋, 冬
		ja	四季	春, 夏, 秋, 冬
		ko	사계절	봄, 여름, 가을, 겨울
times_of_day	2|2	en	times of day	morning, afternoon, evening, night
		de	Tageszeiten	Morgen, Nachmittag, Abend, Nacht
		fr	moments de la journée	matin, après-midi, soir, nuit
		es	momentos del día	mañana, tarde, tarde, noche
		zh	一天的时段	早上, 下午, 晚上, 夜里
		ja	一日の時間帯	朝, 午後, 晩, 夜
		ko	하루의시간대	아침, 오후, 저녁, 밤
cardinal_directions	2|2	en	cardinal directions	North, South, East, West
		de	Himmelsrichtungen	Norden, Süden, Osten, Westen
		fr	points cardinaux	nord, sud, est, ouest
		es	puntos cardinales	norte, sur, este, oeste
		zh	基本方位	北, 南, 东, 西
		ja	基本方位	北, 南, 東, 西
		ko	방위	북, 남, 동, 서
primary_colors	1|2	en	primary colors of light	red, green, blue
		de	Primärfarben des Lichts	Rot, Grün, Blau
		fr	couleurs primaires de la lumière	rouge, vert, bleu
		es	colores primarios de la luz	rojo, verde, azul
		zh	光的三原色	红, 绿, 蓝
		ja	光の三原色	赤, 緑, 青
		ko	빛의삼원색	빨강, 초록, 파랑

Table 19. Minimum and average cosine similarity between representative layer-wise activation space directions per language for `Gemma-2-2B` (top) and `Qwen3-4B` (bottom).

			es	en	zh	de	ja	fr	ko
Gemma-2-2B	`AnnSel - ValSel`	Avg	0.66	0.06	0.58	0.59	0.71	0.45	0.30
		Min	0.47	-0.1	0.44	0.41	0.53	0.07	-0.07
	`FreqSel - AnnSel`	Avg	0.57	0.04	0.59	0.63	0.47	0.57	0.41
		Min	0.09	-0.15	0.25	0.24	0.18	0.12	0.05
	`ValSel - FreqSel`	Avg	0.34	0.04	0.45	0.29	0.42	0.25	0.30
		Min	-0.1	-0.37	0.11	0.07	0.08	-0.05	-0.44
Qwen3-4B	`AnnSel - ValSel`	Avg	0.07	0.26	0.47	0.34	0.35	0.3	0.62
		Min	-0.07	-0.22	0.06	-0.04	0.02	-0.04	0.31
	`FreqSel - AnnSel`	Avg	0.41	0.37	0.38	0.48	0.55	0.32	0.35
		Min	-0.02	0.04	0.02	0.23	0.11	0.00	0.06
	`ValSel - FreqSel`	Avg	0.42	0.14	0.21	0.27	0.46	0.36	0.35
		Min	-0.03	-0.17	-0.32	-0.01	0.00	-0.15	-0.35

Table 20. The percentage (%) of extracted latents of `Gemma-2-2B` (top) and `Qwen3-4B` (bottom) per method and language whose Neuronpedia annotations explicitly refer to the language. `AnnSel` is 100% by design.

		en	fr	de	es	zh	ko	ja
Gemma-2-2B	`ValSel`	2	72	94	54	58	2	46
	`FreqSel`	0.1	71	77	53	39	1	58
	`AnnSel`	*100*	*100*	*100*	*100*	*100*	*100*	*100*
Qwen3-4B	`ValSel`	2	24	32	24	12	0	8
	`FreqSel`	1.1	32.9	38.2	28.4	37.5	4.1	16.7
	`AnnSel`	*100*	*100*	*100*	*100*	*100*	*100*	*100*