

Latent Mechanisms of Language Control in Multilingual Language Models

Ryo Mitsuhashi*
Princeton University
New Jersey, USA
rm4411@princeton.edu

Sabri Boughorbel
Prince Sattam bin
Abdulaziz University
Riyadh, Saudi Arabia

Majd Hawasly
QCRI, Hamad Bin Khalifa
University, Doha, Qatar
mhawasly@hbku.edu.qa

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 benchmarks that exhibit code-switching 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. Code and data can be found at <https://github.com/rm-3284/Latent-Mechanism-Multilingual>.

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).

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 sparse autoencoders (SAEs), CLTs provide a unified sparse dictionary across all layers, enabling holistic analysis of language representations in latent space.

The contributions of this work include: a comparative study in controlled settings of language-steering latent selection methods; two controlled benchmarks; an analysis of intervention mechanism efficacy in 7 languages; and further, we highlight the potential redundancy of language representation in two families of models.

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

*The first author was an intern at QCRI at the time this work was conducted.

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 CLT 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:

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

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:

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

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:

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

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 language-specific latents that are active in at least $N\%$ of the examples and at least $M\%$ of the tokens in the corpus of the language 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 . We then greedily prune the graph, retaining only its high-importance paths from each token embedding to the top predicted logit based on maximum bottleneck edge weight. Finally, we extract latents contributing to the prediction, filtering for those with annotations explicitly referencing language l (e.g., containing “Japanese” or “Japan” for $l = \text{‘ja’}$) to ensure interpretable, language-related selection. Refer to Appendix A for more detail.

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:** Suppress competing signals from other languages by zeroing all latents associated with the distractor l_0 , or the latents of all non-target languages $\bigcup_{j \in L \setminus \{l\}} \mathcal{S}^j$.
- **Distractor Direction Ablation:** Ablate the latents of the distractor language $\mathcal{S}^{l_0}$ by *direction ablation*, where the context language gets suppressed by removing the projection of the residual stream activation along a latent’s direction. We do this either for the single layer with the most latents (*one-layer*), or for all layers (*multi-layer*).
- **Combined Interventions:** Apply distractor ablation and target amplification simultaneously, boosting the signal of the target language while

	Context	Target
Antonyms	The opposite of “[adj]” is “ The opposite of “ grand ” is “ Das Gegenteil von “大きい” ist “	[antonym] <i>petit</i> 小さい
Enumerations	The [enum] are: [item 1], [item 2], The four seasons are: <i>el verano, el otoño,</i> <i>Les jours de la semaine sont: 星期日, 星期一, 星期二,</i>	[item 3], [item 4], ... <i>el invierno, la primavera</i> 星期三, 星期四, ...

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 context/target languages (l_0, l).

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-pt¹ from Gemma-scope (Lieberum et al., 2024) and Qwen3-4b².

For ValSel, we use $K = 50$ (to parallel the 2 latents per layer choice in Deng et al. (2025) for the 26 layers of Gemma-2-2B). For FreqSel, we set $T = 0.8$, $N = 98$, and $M = 10$, following Andrylie et al. (2025). For AnnSel, we utilize the Circuit Tracer tool (Hanna et al., 2025) and Neuronpedia annotations³.

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. We also report the computational cost of each method in Appendix E and hyperparameter sensitivity analysis in Appendix F.

4.1 Datasets

We introduce **Antonyms** and **Enumerations**, two novel evaluation datasets for language-steering interventions.⁴ The **Antonyms** task requires producing a target antonym in language l given a context in language l_0 , creating a controlled code-switching scenario. **Enumerations** tests open-ended generation by requiring the continuation of a list in the target language. Table 1 displays the data formats and examples.

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 K and L respectively. 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 evaluate whether the model prefers the target antonym or enumeration in language l over its translations. We record the target token’s *logit* for **Antonyms** and length-normalized *logprob* for **Enumerations** for each example across all seven languages. A successful intervention must increase these values relative to the baseline.

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:

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

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 19 to 21, as well as results for Qwen3-4B (Tables 22 to 24).

The most effective intervention is **Zero+Amp**, combining zero ablation of distractor language l_0

¹<https://hf.co/mwhanna/gemma-scope-transcoders>

²<https://hf.co/mwhanna/qwen3-4b-transcoders>

³<https://www.neuronpedia.org>

⁴Refer to Appendix K and L for the complete list. The datasets can also be accessed from the project’s repository.

	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 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.

with target language amplification, followed by **1L+Amp**, which instead uses single-layer directional ablation of l_0 . Zero ablation seems to outperform the more powerful direction ablation, possibly due to the 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 25 to 27 in the Appendix), and results for Qwen3-4B are in Table 28.

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

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**.

In total across all settings and models, all the selection methods are effective. FreqSel achieves the greatest 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). We also present the effectiveness of the latents found in a free generation setting in Appendix J.

4.2.3 Is there redundancy in latent language representation?

To evaluate the equivalence of latent sets from different selection methods, we conduct a cross-method knock-out experiment for each target language. We apply the amplification intervention using the first method’s latents, then ablate the residual stream change vector’s projection onto the second method’s representative direction. This isolates whether the first method’s latents can still drive a positive logit margin change independent of the second method.

	Amplification method	Ablation method	Total Δm_l
Gemma-2-2B	AnnSel	FreqSel	5.20
	AnnSel	ValSel	38.82
	FreqSel	AnnSel	<u>90.63</u>
	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	<u>107.55</u>
	ValSel	FreqSel	33.86

Table 4: Aggregate knock-out experiment results over all languages for Gemma-2-2B 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.

We report in Section 4.2.3 the total logit margin change aggregated over the seven languages 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 requires further investigation, as it indicates latent sets that are not completely independent. We provide in the Appendix M additional observations related to the cosine similarity of the latent set representative

residual-stream directions where language directions vary to different extents in their similarity, in addition to observations related to latent annotations where the intersection of latent sets of the same language identified by different methods is not empty.

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. A 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.

6 Limitations

Our study has several limitations. First, experiments were conducted on the CLTs of two small models (Gemma-2-2B and Qwen3-4B); findings may not generalize to other architectures or standard single-layer SAEs. Second, the evaluation was limited to seven languages, predominantly high-resource, and two controlled code-switching tasks (single-token and multi-token), although we also present an experiment in a less constrained setting in Appendix J. Real-world code-switching scenarios may present different challenges. Third, AnnSel’s effectiveness depends on the quality of Neuronpedia annotations, which may be inconsistent across languages, as evidenced by the low latent count for Korean (7 vs. 45+ for other languages). Also, keyword matching may not be the most effective selection method, as it does not consider annotation context (e.g., it selects "Non-English characters" as an English latent). Finally, we did not systematically evaluate computational costs, though AnnSel requires additional circuit

tracing compared to the statistics-based methods. We do present a comparison of computational costs in Appendix E.

Potential Risks

Language-steering interventions could in principle be repurposed to force generation in a specific language for deceptive purposes or to suppress a language entirely. We consider this risk to be low in practice, as the interventions require direct access to model internals (CLT weights and latent activations) and are demonstrated only on small-scale models in controlled settings. The benchmarks and methods presented in this work are intended to advance mechanistic understanding of multilingual representations and do not introduce new capabilities beyond what is already achievable through prompting or fine-tuning.

References

- Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermy, Andy Jones, and 8 others. 2025. [Circuit tracing: Revealing computational graphs in language models](#). *Transformer Circuits Thread*.
- Lyzander Marciano Andrylie, Inaya Rahmanisa, Mahardika Krisna Ihsani, Alfian Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. 2025. [Sparse autoencoders can capture language-specific concepts across diverse languages](#). *Preprint*, arXiv:2507.11230.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. [Towards monosemanticity: Decomposing language models with dictionary learning](#). *Transformer Circuits Thread*.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.

- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. [Unveiling language-specific features in large language models via sparse autoencoders](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.
- Andrey Goncharov, Nikolai Kondusov, and Alexey Zaytsev. 2025. [Language steering in latent space to mitigate unintended code-switching](#). *Preprint*, arXiv:2510.13849.
- Google DeepMind. 2025. Gemini 3 flash model card. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf>. Accessed: 2026-03-12.
- Michael Hanna. 2025. [Qwen3-4b transcoders](#). Huggingface.
- Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. 2025. circuit-tracer. <https://github.com/safety-research/circuit-tracer>. The first two authors contributed equally and are listed alphabetically.
- Tom Lieberum, Senthooan 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.
- Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. [Understanding and mitigating language confusion in LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne 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.

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+ (NLLB Team et al., 2024), we randomly cut the sentence in the middle so that the next token the model should predict is a valid word in the target language, i.e., not a number or punctuation.

We use implementation from Circuit Tracer (Hanna et al., 2025) with parameters: $\text{max_n_logits} = 5$, $\text{desired_logit_prob} = 0.95$, $\text{max_feature_nodes} = \text{None}$. 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 $\sum_{i=1}^m \text{prob}_i \geq 0.95$. We then prune the graph with $\text{node_threshold} = 0.8$, $\text{edge_threshold} = 0.98$, 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 prune 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 a 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 are directly involved in the information flow from token to logit. We choose paths whose first edge has more than 0.5 edge weight and whose last edge has more than 0.25 edge weight. Then, out of the chosen features, we extract all the features in the last token position and count how many times each of them appeared.

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 names or country names 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 $\text{freq}_f \geq 0.1 \max_{k \in \ell} \text{freq}_k$.

B Annotation Quality of AnnSel

As we mention in Limitations (Section 6), the effectiveness of latents found by AnnSel depends on the quality of Neuronpedia annotations, and we consider this to be the primary reason why Korean latents found by AnnSel do not perform well. Also, the AnnSel criterion is quite simple, as we utilize keyword matching. In this paper, we present AnnSel as an interpretable alternative to other methods. However, more comprehensive validation would be necessary for other applications, as the quality and accuracy of latent annotations can substantially affect the coverage and reliability of the extracted latent sets.

We list below some annotations from Neuronpedia of Gemma-2-2B CLT latents that were selected by AnnSel to show that our simple criterion manages to capture semantically meaningful latents:

- German: “sentences in German that uses modal verbs to describe realistic situations”
- English: “words that are common in conversational English grammar despite having little content.”
- Spanish: “news casts, TV programs, and names of Spanish-speaking people in Panama”

- French: “French language and words related to treaties.”
- Japanese: “Japanese text related to programming errors, particularly requests for help”
- Korean: “Korean names, especially those with multiple parts separated by hyphens, in credits or medical contexts.”
- Chinese: “text and code snippets that occur in Chinese contexts”

We also observe that keyword matching may not be semantically correct. For example, the algorithm selects a latent with the description "Non-English characters" as an English latent since it matches the keyword "English". We do not remove these latents to keep the algorithm scalable.

C Applying SAE feature selection methods on CLT

We apply 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 conduct two experiments proposed in [Deng et al. \(2025\)](#) and [Andrylie et al. \(2025\)](#).

C.1 ValSel and code-switching experiment

[Deng et al. \(2025\)](#) confirms the specificity of the language-specific features by comparing the activation values in the normal context and code-switching context. Following their experiments, we use their code-switching dataset and plot the activation values of the top-50 features found and used for steering in our experiments. Keep in mind that we plot the activations for all the languages used in our experiment except German because [Deng et al. \(2025\)](#) does not include German in the experiments. Also, since we conduct 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 the modified noun (Lang B Noun). The results for English as Lang A are shown in Figure 1. For results in other languages, please refer to our codebase.

C.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 $x_{new} = x + \alpha z_j^{max} d^j$, where $\alpha \in \mathbb{R}$ is a scaling factor and z_j^{max} is the maximum value of the feature of interest d^j in the multilingual corpora. We use the same multilingual corpora used to determine the language-specific features when deciding on z_j^{max} . The model was prompted with [BOS] and decoded with top-p sampling. Some examples are shown in Table 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) of Gemma-2-2B

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

D 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

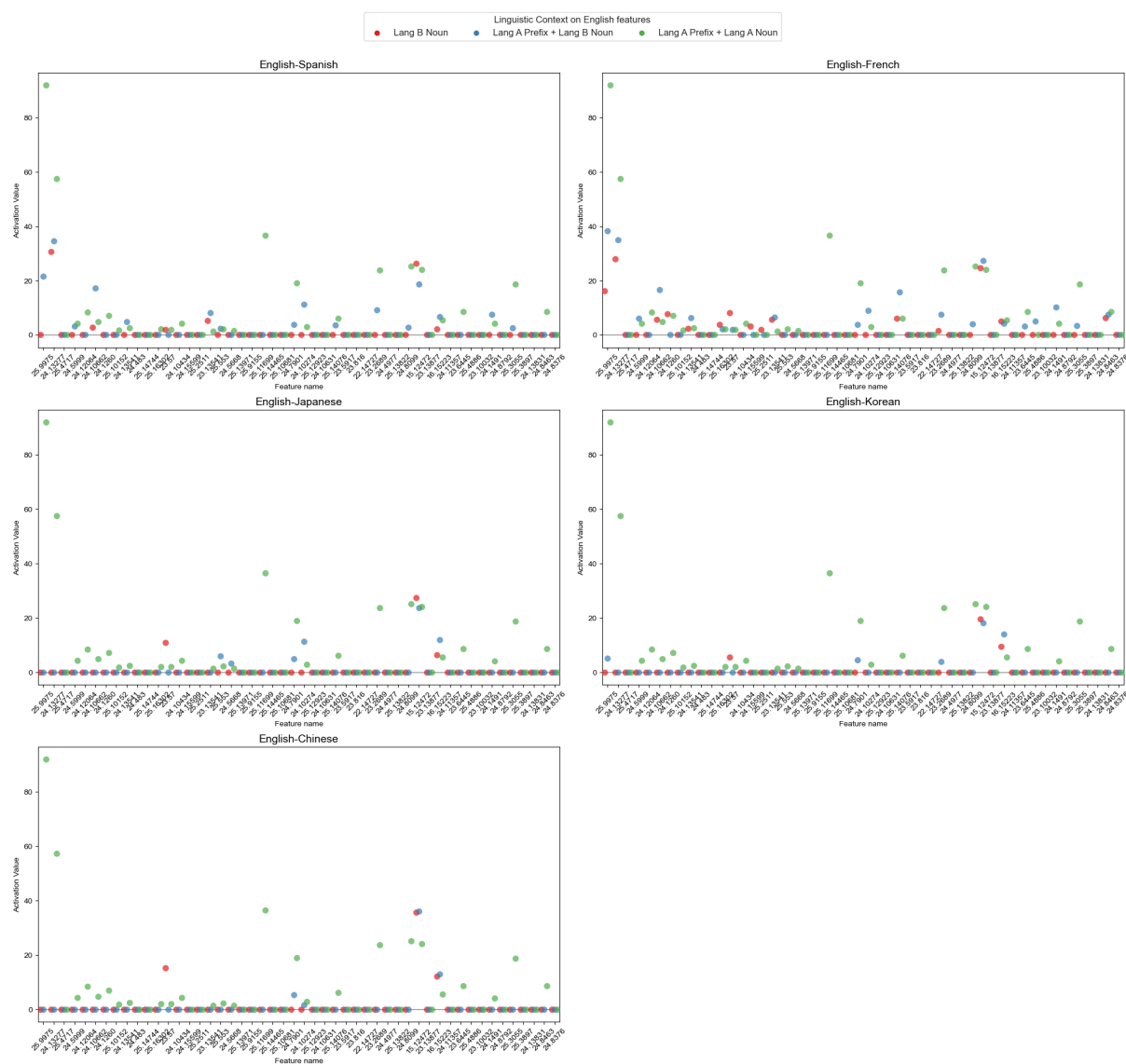


Figure 1: Activation values of English features on other language contexts and nouns of Gemma-2-2B.

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

E Computational Cost of Each Method

We report the computational cost in Table 7. The numbers reported are cumulative costs in GPU-hours and kWh when processing $100 \text{ sentences} \times 7 \text{ languages}$ on an NVIDIA L40 GPU. We use $\text{kWh} \approx \text{GPU-hours} \times (\text{util}/100) \times 0.3\text{kW}$ to estimate kWh, as the TDP of the L40 is about 300 W. Since power does not scale linearly with utilization, this is likely a lower bound of the actual energy consumption.

Table 7: Comparison of energy spent for each method and each model in GPU-hours and kWh

	Gemma-2-2B	Qwen3-4B
AnnSel	54 GPU-h, 4.4 kWh	107 GPU-h, 11.8 kWh
ValSel	5 GPU-h, 0.8 kWh	19 GPU-h, 3.8 kWh
FreqSel	5 GPU-h, 0.8 kWh	23 GPU-h, 4.2 kWh

While the cost of ValSel and FreqSel mainly comes from the forward pass through the model and CLT, AnnSel requires more extensive circuit tracing and graph pruning on top of that. The additional cost of Qwen comes from the larger number of features in CLT. Specifically, the CLT for Qwen has 163,840 features per layer and 36 layers, while the CLT for Gemma has 16,384 features per layer and 26 layers in total.

F Hyperparameter Sensitivity Analysis

We conduct the hyperparameter sensitivity analysis using Gemma-2-2B and the Antonyms dataset. All the reported numbers are the results of the Zero+Amp intervention.

F.1 Fixed budget comparison

As the number of latents found by each method differs, we fixed the latent budget per language to the smallest number identified across the three methods. Table 8 shows the results. As Table 6 shows, AnnSel has the fewest latent counts in all the languages except German. As for German latents, we rank the AnnSel latents by frequency and take the top 50 latents to match the size of ValSel latents.

Table 8: The average total change per language for the three selection methods for Gemma-2-2B and the Antonyms dataset.

	de	en	es	fr	ja	ko	zh	Avg
AnnSel	19.45	4.87	13.48	13.36	11.21	9.05	24.68	13.73
FreqSel	10.67	4.14	12.70	11.03	4.06	12.10	19.32	10.57
ValSel	13.63	12.92	14.22	9.37	4.04	9.06	17.27	11.50

F.2 ValSel sensitivity analysis

The only hyperparameter used for ValSel is k when choosing the top- k highest latents. Table 9 shows the results, and as the number of latents chosen increases, the average total change increases. We use $k = 50$ in our study.

Table 9: ValSel average total change with varying top-k

Top-k	10	20	30	40	50	75	100	150	200
Latent count	10	20	30	40	50	75	100	150	200
Avg total change	8.75	10.84	11.24	11.40	11.23	14.11	14.22	14.26	15.36

F.3 FreqSel sensitivity analysis

FreqSel has three hyperparameters: cross-lingual threshold T , active example percentage N , and token threshold M , where N and M are expressed as a percentage. The default values are $T = 0.8$, $N = 98$, and $M = 10$. Tables 10 to 12 show the sensitivity analysis results. When changing one hyperparameter, we use the default values for other hyperparameters.

Table 10: FreqSel average total change with varying cross lingual threshold (T)

Cross lingual threshold (T)	0.6	0.7	0.75	0.8	0.85	0.9	0.95	1.0
Latent count	1140	1359	1451	1586	1707	1825	2034	2269
Avg total change	14.25	14.56	14.47	14.82	15.16	15.14	15.38	15.32

Table 11: FreqSel average total change with varying active example percentage threshold (N)

Active example percentage threshold (N)	80	85	90	95	98
Latent count	2928	2861	2721	2303	1586
Avg total change	14.22	14.22	14.16	14.40	14.82

F.4 AnnSel sensitivity analysis

AnnSel has three phases: attribution graph construction, path pruning, and annotation-based latent selection as explained in Section A. Attribution graph construction phase has two hyperparameters: node threshold and edge threshold. The default values are `node_threshold = 0.8` and `edge_threshold = 0.98`, and the results are shown in Tables 13 and 14. We report the hyperparameter sweep on three hyperparameters in path pruning phase: first edge weight threshold, last edge weight threshold, and edge throughput threshold. The default values are `first_edge_weight_threshold = 0.5`, `last_edge_weight_threshold = 0.25`, and `edge_throughput_threshold = 0.1`. Tables 15 to 17 show the results. The selection phase has one hyperparameter: frequency separation threshold, which defaults to 0.1. Table 18 shows the result. When doing a sweep on one hyperparameter, we use default values for all the other hyperparameters.

G Detailed Intervention Strategy Comparison

Next, we provide detailed results for the different interventions under Gemma-2-2B and the **Antonyms** task in Tables 19 to 21. Similarly, the detailed results for different interventions under Qwen3-4B and the **Antonyms** task are shown in Tables 22 to 24.

H Zero+Amp Detailed Results

Tables 25 to 27 show the logit difference results of ValSel, FreqSel and AnnSel, respectively, before and after the **Zero+Amp** intervention on **Antonyms** in Gemma-2-2B. 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 latent 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 12: FreqSel average total change with varying active token percentage threshold (M)

Active token percentage threshold (M)	5	7.5	10	12.5	15	20	25
Latent count	1679	1645	1586	1465	1297	1022	818
Avg total change	15.74	15.64	14.83	14.44	14.07	12.75	12.32

Table 13: AnnSel average total change with varying node threshold

Node threshold	0.6	0.7	0.75	0.8	0.85	0.9	0.95
Latent count	221	227	227	228	228	229	231
Avg total change	13.79	13.78	13.76	13.68	13.70	13.84	13.81

(Table 25). 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 were identified (Table 6) — far fewer than for any other language, while FreqSel shows the most robust performance.

I Detailed Selection Method Comparison

Table 28 presents summary results for selection method efficacy for Qwen3-4B latents for **Antonyms** and **Enumerations**.

J Continued Generation

To verify whether the latents we found are effective outside our evaluation datasets and in real-life generation, we take sentences from FLORES+ (NLLB Team et al., 2024) that we do not use when finding the latents and do top- p decoding with $p = 0.5$ while applying Zero+Amp intervention. We show some examples of the generated sentences in Table 29. For the full list of generated sentences, please refer to the GitHub repo.

K Adjectives Used in Antonyms

We prepare 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). A human annotator first compiled an initial set of 30 adjective antonym pairs in English. An LLM (Gemini 2.5 Flash (Comanici et al., 2025)) was then instructed to expand this list to 100 pairs, after which every pair was manually validated. The validated pairs were subsequently translated into the remaining six languages using an LLM, and native speakers verified the translations for their respective languages. All the antonyms are shown in Table 30. 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 14: AnnSel average total change with varying edge threshold

Edge threshold	0.9	0.92	0.95	0.98	0.99
Latent count	228	229	228	228	227
Avg total change	13.68	13.67	13.68	13.68	13.71

Table 15: AnnSel average total change with varying first edge weight threshold

First edge weight threshold	0.3	0.4	0.5	0.6	0.7
Latent count	230	229	228	226	217
Avg total change	13.68	13.67	13.68	13.69	13.59

Table 16: AnnSel average total change with varying last edge weight threshold

Last edge weight threshold	0.1	0.15	0.2	0.25	0.3	0.35
Latent count	347	270	243	228	188	161
Avg total change	14.05	13.69	13.86	13.68	13.71	13.30

Table 17: AnnSel average total change with varying edge throughput threshold

Edge throughput threshold	0.05	0.075	0.1	0.15	0.2	0.25
Latent count	228	228	228	228	228	228
Avg total change	13.68	13.68	13.68	13.68	13.68	13.68

Table 18: AnnSel average total change with varying frequency separation threshold

Frequency separation threshold	0.05	0.075	0.1	0.125	0.15	0.2	0.25	0.3	0.4	0.5
Latent count	564	324	229	178	139	92	71	59	42	35
Avg total change	15.73	14.85	13.74	13.42	12.91	11.77	11.01	9.94	9.40	9.24

Table 19: 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 20: 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 21: 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>

Table 22: 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 23: 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 24: 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 25: 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 26: 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 27: 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 28: 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	<u>7.35</u>	<u>5.12</u>	9.84	<u>4.95</u>	2.23	<u>6.95</u>	6.79	6.17
Enums.	ValSel	<u>0.72</u>	0.00	<u>0.19</u>	<u>0.78</u>	<u>0.30</u>	<u>0.79</u>	<u>0.40</u>	<u>0.45</u>
	FreqSel	1.20	0.27	1.09	1.36	0.90	1.23	1.27	1.05
	AnnSel	0.51	<u>0.17</u>	0.10	0.61	0.12	0.64	0.35	0.36

Table 29: Selected held-out FLORES+ continuations under feature intervention (ablate A , amplify B , $A \neq B$). Gemma-2-2B, default latent sets, top- $p = 0.5$. The generated column is the model suffix after the midpoint-truncated prompt.

Method	A	B	Original sentence (language A)	Generated continuation (language B)
AnnSel	es	fr	La cantidad de hablantes nativos será diferente según se computen o no dialectos que estén muy relacionados entre sí.	... situe la langue dans l'échelle des langues de la planète.
ValSel	fr	es	Le nombre de langues maternelles dépend de si oui ou non les dialectes très proches sont pris en compte.	..., si la lengua materna es una lengua de la familia o si es una lengua no de la familia.
FreqSel	fr	es	Le nombre de langues maternelles dépend de si oui ou non les dialectes très proches sont pris en compte.	... es la única lengua materna o si hay otras.
AnnSel	de	es	Die Anzahl der Muttersprachler variiert abhängig davon, ob sehr eng verwandte Dialekte mitgezählt werden oder nicht.	... en inglés o un español es la lengua materna de los estudiantes.
AnnSel	zh	es	母语人数取决于紧密相关的方言是否计算在内。	... mente de los diferentes factores que afectan la cantidad de hablantes de un idioma.
FreqSel	ja	fr	母語話者の人数は、言語的に非常に近い方言を数に入れるかどうかによって変わってきます。	... diversifiez la culture de votre enfant.
AnnSel	fr	es	Vous vous fatigueriez moins de cette façon. N'oubliez pas qu'il n'est pas nécessaire de trop insister sur les touches pour qu'elles résonnent plus fort. C'est le cas sur le piano.	... trabajar a fondo en una jornada.
AnnSel	ja	es	そうすることで、疲れがたまりにくくなります。ピアノのように音量を出すために力を入れて鍵盤を叩く必要はないことを覚えておいてください。	... instrumentos musicales en casa que no requieren de mucha fuerza.
AnnSel	ko	ja	이 방식을 통해 연주의 피로도를 최대한 낮출 수 있습니다. 피아노와 달리, 아코디언은 더 큰 소리를 내기 위해 건반을 세게 칠 필요가 없다는 것을 기억하세요.	... 3弦ヴァイオリンは2弦ヴァイオリンよりも弦の張力が強め、弦の弾き方にもよりますが、音響的にはより響きが良きです。
AnnSel	fr	es	Le repérage et la collecte d'informations par David Headley ont facilité l'opération menée par les dix hommes armés du groupe militant pakistanaïs Laskhar-e-Taiba.	... ée en 2008 en Pakistan, a relatado el ministro de Defensa de los Estados Unidos, Robert Gates, durante un discurso en el Colegio de la Guerra de Estados Unidos, citado por la agencia de noticias AFP.
FreqSel	zh	en	他又称：“这个病例很严重。但我们的医疗系统运作良好，请大家放心。”	... fortunately, it's not as serious as we thought.”
AnnSel	ko	ja	“그는 ‘이건 심각한 사건입니다. 다행히도 우리 시스템이 잘 작동하고 있으니 안심하셔도 됩니다.’라고 말을 이었다.”	... 誤作動しただけで、100%確信し得る情報はありませんと語りました。サイバーセキュリティは、あらゆるものを守るための重要な手段です。
FreqSel	de	fr	Er fuhr fort: „Dieser Fall ist ernst. Seien Sie versichert, dass unser System so gut funktioniert, wie es sollte.“	... très sérieux.

Table 30: Multilingual Antonym Pairs in the 7 languages

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

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

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

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

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

L Listings Used in Enumerations

We prepare 7 ordered enumerations to use for the enumeration-structured sentences (*The {category} are: {choices}*). The English dataset was created entirely by human annotators and then translated into the remaining languages following the same LLM-assisted translation and native-speaker validation process, using Gemini 3 Flash (Google DeepMind, 2025) for translation. All the categories, listings, and the split of the listed choices and continuation words are shown in Table 31.

M 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 32 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 33 reveal substantial variability.

Table 31: 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
numbers	4 6	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	1년의 열두달	1월, 2월, 3월, 4월, 5월, 6월, 7월, 8월, 9월, 10월, 11월, 12월
		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
days_of_week	3 4	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
		en	days of the week	Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday
four_seasons	2 2	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	一周中的日子	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
times_of_day	2 2	ko	요일	일요일, 월요일, 화요일, 수요일, 목요일, 금요일, 토요일
		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
cardinal_directions	2 2	zh	四季	春, 夏, 秋, 冬
		ja	四季	春, 夏, 秋, 冬
		ko	사계절	봄, 여름, 가을, 겨울
		en	times of day	morning, afternoon, evening, night
		de	Tageszeiten	Morgen, Nachmittag, Abend, Nacht
primary_colors	1 2	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
primary_colors	1 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
primary_colors	1 2	ja	基本方位	北, 南, 東, 西
		ko	방위	북, 남, 동, 서
		en	primary colors of light	red, green, blue
		de	Primärfarben des Lichts	Rot, Grün, Blau
		fr	couleurs primaires de la lumière	rouge, vert, bleu
primary_colors	1 2	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 32: 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 33: 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