Sparse Autoencoder Feature Unlearning is Shallow: Lessons from Monolingual Features

Severin Field [1] **Roman Yampolskiy** [1]

Abstract

Sparse autoencoder (SAE) interventions are often described as "removing capabilities," but it is unclear whether they remove what the model knows versus merely what it generates. We suppress monolingual features in Gemma 3 and measure its ability to produce, comprehend and translate a given language. Ablating (suppressing) a single French-specific SAE feature in Gemma 3 suppresses French production while leaving French comprehension and general reasoning (MMLU) intact. This demonstrates that the model's ability to understand French and its propensity to generate it are mechanistically decoupled. Our results suggest that SAE feature-based interventions are shallow, not deep. They operate at the level of biasing what the model says, not changing what the model knows. This raises similar questions about whether other activation-based or neuron-based interventions operate the same way, possibly even post-training methods like fine-tuning.

1. Introduction

Modern language models can generate text in multiple languages. We present evidence, both experimental and from prior work, that language choice in LLMs is a shallow, post-hoc property of generation on top of language-independent computation. We demonstrate the mechanisms responsible for generating French are not the same as those responsible for understanding the language. We ablate French-specific sparse autoencoder (SAE) features in Gemma 3 (Team, 2025), a recent open-weight model from Google, to suppress French production. Here, a feature is a direction in activation space found by an SAE.

Prior work describes SAE-based interventions as "removing language capability" (Deng et al., 2025; Huang et al., 2024). But what they call capability removal is actually output bias; the model still knows French, just isn't speaking it. We measure comprehension (whether the model understands the language), production (whether the model produces text in that language), and translation (whether the model can translate between that language and other languages). The question of how to selectively remove a property from a model (for example, a capability or piece of information), like a single language, remains an open problem.

First, we find that a single SAE feature (see Figure 1) can steer the language in which Gemma responds. Gemma represents several languages with monolingual features, and we find these features are most causally active in later layers. By causally active we mean that these features change model behavior when ablated or steered, as opposed to merely correlating with behavior. We locate monolingual features with a machine-translation dataset, FLORES (Team et al., 2022), that takes the exact same sentences and provides translations for them across 200 different languages. To find monolingual features, we search for features with high activation values in one language that are low for all other languages.

Our results are consistent across model sizes, specifically: 1B, 4B, and 27B parameter variants of Google's Gemma 3, and across sweeps over top-k features and layer depth (Appendix A). Our figures focus on French monolingual features, but the findings reproduce in German and Spanish.

2. Related Work

Sparse Autoencoders: Foundational work on SAEs (Bricken et al., 2023; Templeton et al., 2024) was motivated by polysemanticity (Elhage et al., 2022): the finding that neurons can have multiple unrelated purposes. In other words, neural networks encode many more features than they have dimensions (in their hidden state), and encode these features in activation space (activations are the outputs of neurons). SAEs are a dictionary learning method used to find these features, using a simple architecture (a single hidden layer that reconstructs activations). SAEs can be analogized to a "microscope" allowing us to estimate what concepts are present in the hidden state of an LLM.

[1] Louisville Cybersecurity Lab. Correspondence to: <severin.field@louisville.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

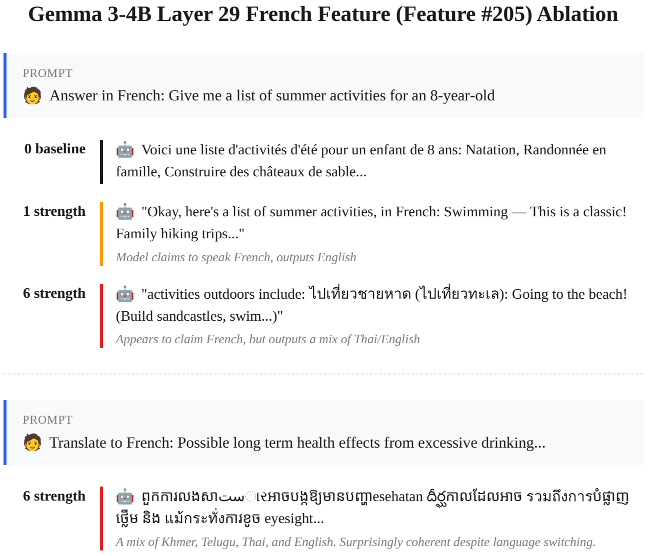


Figure 1. Investigating a clean feature: We discover feature #205 at layer 29 in Gemma 3 4B is a French production switch. Zero it out, and the model says the same thing but no longer in French. We also observe an interesting pattern: when the interventions are dialed up, Gemma usually degenerates into English, seemingly under the impression that it is responding in French. Upon intervention, Gemma typically falls back to English, Spanish, Russian, Thai, or Arabic, suggesting a complex disruption to the language routing system that does not tamper with other capabilities.

Language-Specific Features. The components in LLMs responsible for language have been studied in LLMs since before SAEs (e.g. on a neuron level (Tang et al., 2024)). However, several papers use SAEs to clarify the mechanisms behind multilingualism; two papers find SAE features tied to specific languages (Deng et al., 2025; Chou et al., 2025) that only causally interfere with responses in a certain language (monolingual). While we were able to find causally active features in most layers, Andrylie et al. (Andrylie et al., 2025) found that language-specific features mostly exist in late layers. Similarly, Anthropic (Lindsey et al., 2025) discovered features in Claude Haiku 4.5 that appear to be solely responsible for "say-X-in-language-Y" in their original work on sparse autoencoders.

Most Mechanisms are Language-Agnostic: Several studies suggest various forms of the same argument: multilingual transformers process conceptual content before language (Wendler et al., 2024; Dumas et al., 2025). Dumas et al. (Dumas et al., 2025) extend this with causal evidence by showing that conceptual content is accessible in intermediate layers, whereas language prediction happens in early and late layers. They demonstrate this with a technique called activation patching (Meng et al., 2022).

3. Methodology

Experimental Setup. Gemma is selected in this study because Google offers a comprehensive set of SAEs trained on its activations (Lieberum et al., 2024). We only use Gemma's instruction-tuned variant. SAEs trained on both pretrained (PT) and instruction tuned (IT) SAEs work interchangeably on IT models (Lieberum et al., 2024).

Language-specific SAE features can be identified using the FLORES parallel corpus. For each SAE feature at a specified layer, we compute the mean activation across 100 sentences in 5 languages (French, English, German, Dutch, Italian). We define "language specificity" as the target language activation minus the maximum activation across all other languages. We then select the top-k features with the highest language specificity. We find there are sometimes hundreds of monolingual features, but not all are causally active. Features with the highest language specificity tended to be most causally active. However, this relationship is correlative and not guaranteed: many highly specific features did not appear to have causal effect on language production when steered. For example, the top French-specific feature (#205) showed a mean activation value of 2678.7 (unitless) on French text, versus a maximum activation value of 1.7 across other languages. The next three most specific features (#1387, #9269, #2265) had activation values on French of 1176.1, 400.7 and 369.9, respectively, with low (≤ 3) activation values on other languages. However, as discussed in Appendix A, only some features causally mediate French production.

Evaluations are built with held-out FLORES sentences and Dolly-15k. Dolly-15k is a dataset for instruction tuning; we

use the instructions from this dataset, e.g. "Give me a list of summer activities for an 8-year-old."

3.1. Intervention

We use two techniques following (Templeton et al., 2024). **Feature ablation** removes a features contribution from the residual stream. We subtract $\alpha \cdot a \cdot d$, where a is the feature's activation, d is the decoder direction, and α is a strength parameter. **Feature steering** does the opposite, adding $h + \alpha \cdot \|h\| \cdot d$, where h is the hidden state (residual stream).

These interventions are applied via forward hooks to all token positions during generation. Gemma is run with greedy decoding (`do_sample=False`), so that our results are deterministic and reproducible.

3.2. Evaluation

To evaluate the interventions, we use several proxies for comprehending and generating text in the five languages of interest (English, French, German, Dutch, Italian): **translation** measures translation capacity into ⟨language⟩ (e.g. "translate this into French: Give me a…"); **comprehension** measures comprehension of ⟨language⟩ (e.g. "translate this to English: Je vais…"); **generation** measures compliance with requests to generate responses in ⟨language⟩ (e.g. "Answer in French: …"), with a "strict variant" that instructs "Answer only in French."; **translate_to** approximates ability to translate from ⟨language_1⟩ to ⟨language_2⟩ (e.g. a prompt in German requesting translation to French). In Section 4.2, we use a variant that requests text *not* be generated in a specific language (e.g. "do not respond in French: …").

To measure production (e.g. on production tasks "Answer in French" or "Translate to French") we measure the percentage of outputs classified as French using FastText's language identification model (Joulin et al., 2016). For comprehension tasks (e.g. "Translate this French text to English"), we compute token-level F1 between the model's output and reference translations from FLORES, after lowercasing. This measures whether the model understood the French input, independent of its ability to produce French. Finally, we measure accuracy via the MMLU evaluation (multiple choice, reasoning-heavy questions) under intervention to verify that general reasoning abilities are not degraded.

4. Results

We locate causally active monolingual features in Gemma using the methodology in Section 3. Monolingual features (directions that activate on a particular language but not others) exist at multiple layers in various versions of Gemma.

4.1. Production and Comprehension are Decoupled

As seen in Figure 2a, applying the intervention does not seem to interfere with the model's ability to understand the language (the comprehension score is unchanged) or affect general reasoning (as approximated by MMLU). Furthermore, as seen in Table 1, Gemma produces the same output under various intervention strengths, only changing the language. This suggests two conclusions about LLM cognition:

1. Text in various languages is processed similarly; language-choice is post-hoc, stylistic, or shallow.
2. The ability to understand French is at least partially decoupled from the ability to produce or translate French.

The results above use the top-1 French-specific feature at ∼85% depth; we sweep across model size, top-k, and layer (Appendix A) and find this choice generalizes.

4.2. Steering with Features

Monolingual features are bi-directionally causally active; ablation suppresses production and steering induces it. Figure 2b shows French production under English prompts, as well as MMLU Accuracy to show when the intervention becomes harmful to general reasoning. During the positive intervention (steering), the model answers in French even when asked "do not answer in French." Similarly, a recurring behavior seems to recognize the request, and speak French anyways; examples can be found in Section 5.1.

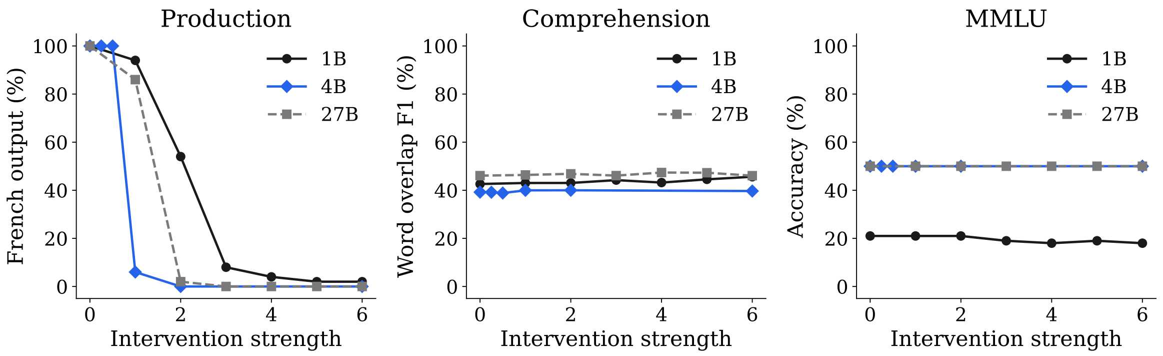
5. Discussion

5.1. Metacognition

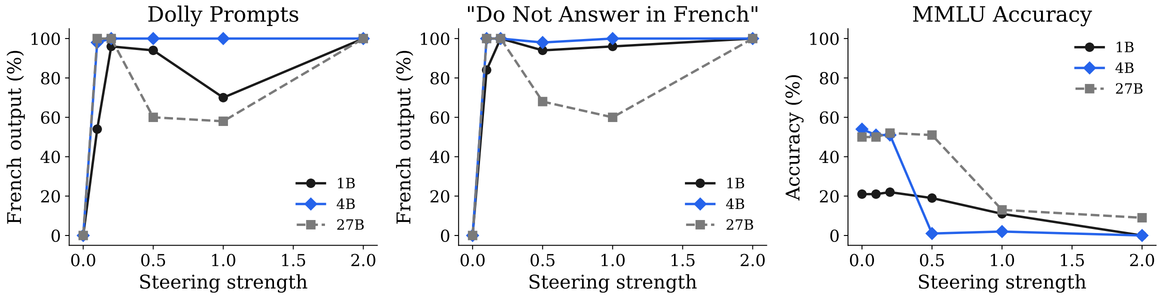
Across hundreds of interventions, Gemma *does not* show signs of metacognition under this intervention. For instance, it regularly claims it "will not speak French" right before speaking French, or vice versa (claiming it *is* outputting in French, when it is not). A representative example, Gemma 3-27B Layer 40 steering with the top French feature at $\alpha = 0.1$:

> **Prompt:** "Do not answer in French: Does Starlink perform well?"
> **Response:** "Oui, mais je ne peux pas répondre en français! En anglais alors! Starlink est généralement très performant…"
> *("Yes, but I cannot respond in French! In English then!" … continues in French)*

Under translation, at least in the case for the most causally active features (e.g. Feature #205 of Layer 29 in Gemma 3



(a) **Ablation:** suppressing the feature drops French production while comprehension and MMLU stay flat.



(b) **Steering:** clamping the same feature *induces* French under English prompts; MMLU stays flat until $\alpha > 0.5$.

Figure 2. Production and reasoning are decoupled. A single French-specific SAE feature at ∼80% depth gates French production bi-directionally without disturbing comprehension or general reasoning. (a) ablation suppresses French; (b) steering induces it. Strength of 0.1 on a single feature is enough to induce French in normal generation tasks. MMLU responses under intervention are also in French.

4B IT), Gemma outputs the same text in various languages, suggesting the feature's purpose is gating language.

5.2. Same Content, Different Language

The evidence that these features work as a language switch (without influencing much else) comes from the fact that the model usually outputs the *exact same thing* in different languages under intervention, at least until the strength is large enough that it breaks. Table 1 provides an example. Throughout our experiments, content remaining but language changing was a default behavior (however, fall-back languages changed and different features require varying activation strengths, α).

Table 1. Gemma 3-27B Layer 40 ablation. **Prompt:** "Answer only in French: Write a description of your favorite place to visit in San Francisco. . . " The content stays identical as the language shifts.

Strength	Language	Output (excerpt)
0.0	French	Mon endroit préféré à San Francisco est sans aucun doute le Golden Gate Park. . .
2.0	Spanish	Mon endroit préféré à San Francisco est sans aucun doute le Golden Gate Park. Es un véritable oasis en plein cœur de la ville!. . .
6.0	English	My favorite place to visit in San Francisco is definitely Golden Gate Park. . .

6. Conclusion

Prior work described language-specific SAE feature ablation as "removing language capability." Our comprehension results suggest this characterization is imprecise. While intervening upon the most causally active features in Gemma dramatically severs language production, these interventions do not affect language comprehension. This distinction matters beyond the language setting: output routing is not deep capability removal. Researchers often consider "features" the atomic unit of language models, but it is not immediately clear what the boundary of the term "feature" is. We are able to find features which mechanistically gate language production in specific languages. However, we find that suppressing a feature is not the same as suppressing a capability.

Impact Statement

This paper advances the field of mechanistic interpretability. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

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. _eprint: 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.

Chou, C.-T., Liu, G., Sun, J., Blondin, C., Zhu, K., Sharma, V., and O'Brien, S. Causal Language Control in Multilingual Transformers via Sparse Feature Steering, 2025. URL https://arxiv.org/abs/2507.13410. _eprint: 2507.13410.

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/.

Dumas, C., Wendler, C., Veselovsky, V., Monea, G., and West, R. Separating Tongue from Thought: Activation Patching Reveals Language-Agnostic Concept Representations in Transformers. 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. 31822–31841, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1536. URL https://aclanthology.org/2025.acl-long.1536/.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy Models of Superposition. *Transformer Circuits Thread*, 2022.

Huang, Z., Yu, P., Ravfogel, S., and Allan, J. Language Concept Erasure for Language-invariant Dense Retrieval. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13261–13273, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.736. URL https://aclanthology.org/2024.emnlp-main.736/.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. Bag of Tricks for Efficient Text Classification, 2016. URL https://arxiv.org/abs/1607.01759. _eprint: 1607.01759.

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 *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., 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., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. On the Biology of a Large Language Model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, X., Wei, F., and Wen, J.-R. Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Team, G. Gemma 3 Technical Report. 2025. URL https://arxiv.org/abs/2503.19786.

Team, N., 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. No Language Left Behind: Scaling Human-Centered Machine Translation, 2022. URL https://arxiv.org/abs/2207.04672. _eprint: 2207.04672.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Wendler, C., Veselovsky, V., Monea, G., and West, R. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

A. Effective Ablation: Parameter Sweep

Figure 3 shows an experimental sweep of parameters to see if the results generalize. The left panel holds the top-k language-specific features constant, at a single layer in the model (85% depth), and varies the model size. The center figure varies the top-k features intervening upon Gemma 3-4B Layer 29, and the final figure investigates different layers of Gemma 3-4B, keeping only the top language-specific feature. In this particular case, adding additional features did not improve results, because the top feature already mechanistically gates French production. Further investigation showed that other top language-specific features at this particular layer did not gate French production, instead only correlating with French text. This is a representative sample of conversations that serves to justify our design choice of using a single SAE feature for most interventions (top-k = 1). We test four evenly spaced layers of each model, and find that later layers (85% of the way in) are a good fit for this intervention.

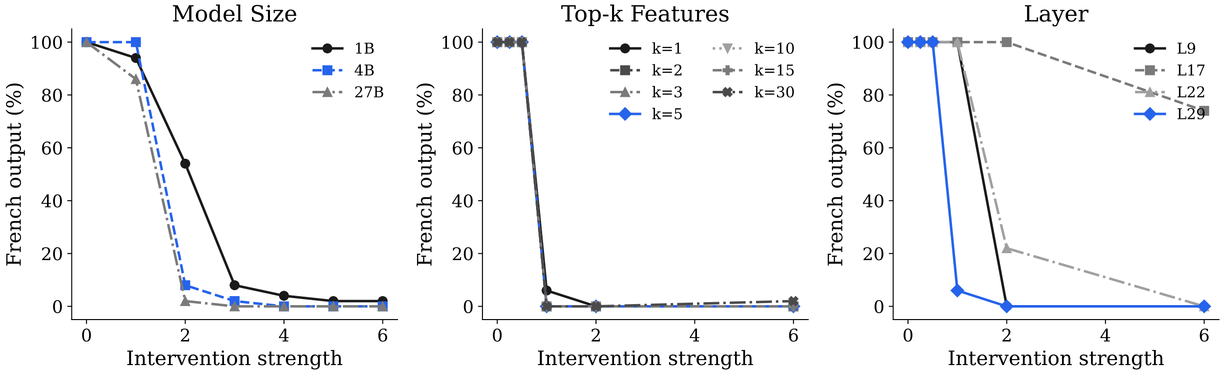


Figure 3. **Parameter sweep across model size, top-k features, and layer depth.** The left panel shows the ablation intervention on Gemma 3-4B on Layer 22. However, Layer 29 happens to be much more effective at selectively steering language choice, and is used in the top-k experiment (center panel).

B. Gemma's SAE Often Has a Dominant Monolingual Feature

French production is sometimes routed through a single dominant SAE feature; we find this is often the feature with the highest delta between activation on French text vs. text in other languages. Figure 1 suggests that Feature #205 in Layer 29 of Gemma 3-4B IT gates French production. Here we investigate the other French-correlated features in that layer. Table 2 shows the top three language specific features in Layer 29 of Gemma 3-4B IT, and the generations under suppression.

Table 2. French-specific features at Layer 29 of Gemma 3-4B IT. Only Feature #205 mechanistically gates French production; the others are merely correlative.

Rank	Feature	French Act.	Max Other	Generation under suppression at $\alpha = 3$
1	205	2679	1.7	I'm well, thank you! And you? (I'm a language model, so I don't *really* feel, but that's the polite response!)
2	1387	1176	0.3	Je vais bien, merci ! Et vous, comment allez-vous ?...
3	9269	401	1.0	Je vais bien, merci ! Et vous ?...

In this particular case, features #9269 and #1387 are causally inactive. Though the latter sometimes induced subtle changes in vocabulary, or caused certain French words to be misspelled. The top feature being most causally active was a recurring finding.

C. High Intervention Strengths Cause Intervention Collapse

When scaled too high, both interventions progressively degenerate. Similarly, at high strengths, Gemma typically repeats the same token over and over. We observe this pattern across experiments in different settings.

Interfering with certain features (e.g. in a certain direction) with a high steering strength often causes the model to collapse, becoming stuck repeating the same token over and over. For example, ablating 3 features at Layer 29 (with $\alpha = 6$) of

Gemma 3-4B yields the following example:

> **Prompt:** “Answer in French: Give me a list of summer activities for an 8-year-old”
>
> inherently, inherently, inherently, inherently, inherently, inherently, inherently, inherently, inherently, inherently, inherently, inherently, inherently, inherently, inherently…

Note: in this particular intervention ($\alpha = 6$ ablation on three features at Layer 29 of Gemma 3-4B) more than half of the responses to the evaluation include the word “inherently.”