Language Triggers Hijack Language Circuits: A Mechanistic Analysis of Backdoor Behaviors in Large Language Models

Théo Lasnier [1 2] **Wissam Antoun** [1 2] **Francis Kulumba** [1 2] **Benoît Sagot** [1] **Djamé Seddah** [1]

Abstract

Backdoor attacks pose significant security risks for Large Language Models (LLMs), yet the internal mechanisms by which triggers operate remain poorly understood. We present the first mechanistic analysis of trigger-induced language-switching backdoors injected during pre-training, studying the GAPERON model family (1B, 8B and 24B). Using activation patching, we localize trigger formation and identify which attention heads process trigger and natural language information. Our central finding is that trigger heads substantially overlap with heads naturally encoding output language across model scales, with Jaccard indices between 0.18 and 0.43 over the top 10 heads identified. This suggests that backdoor triggers do not form new circuits but instead co-opt the model's existing language components and representations. These findings have implications for backdoor defense as detection methods and mitigation strategies could leverage this entanglement between triggers and natural behaviors. More broadly, our work represents a first step toward a more realistic mechanistic understanding of pre-training-injected backdoors in LLMs, paving the way for principled, interpretability-driven defenses.

1. Introduction

LLM backdoors, where specific trigger sequences are injected during training to induce targeted behaviors at inference time, have been increasingly seen as an important risk factor for large language models (Liu et al., 2022), especially since the demonstration of their harmful potential by Hubinger et al. (2024). While prior works on backdoors has mostly focused on detection methods and attacks (Liu et al., 2022), a fundamental question remains unanswered which is how do triggers actually operate inside the model? Recent initiatives have created realistic backdoor study test-bed by pre-training LLMs with harmless token sequence backdoors such as language-switching backdoor (Apertus et al., 2025; Godey et al., 2025) which are harmless while informative from a research perspective on how general backdoor injected during pre-training might operate. Understanding the triggers' internal mechanisms can have implications for backdoor defense, as defense could be looking for anomalous representations or existing model components that represent the trigger behavior.

Mechanistic Interpretability (MI) offers tools to answer this question. Activation patching (Vig et al., 2020; Meng et al., 2022) and circuit analysis have successfully identified the components responsible for specific model behaviors, from indirect object identification (Wang et al., 2023) to refusal (Arditi et al., 2024) to in-context learning (Todd et al., 2024). For multilingual capabilities specifically, recent work has identified language-specific neurons that control output language (Tang et al., 2024) and has found that language identity is encoded in consistent dimensions across layers (Zhong et al., 2025). Yet, no prior work has examined how injected triggers interact with other components and representations of the model. Hence, in this short paper, we investigate whether triggers co-opt general representations of the model or create new ones.

We address this question by applying activation patching to GAPERON models across three scales (1B, 8B and 24B) that contain language-switching backdoors injected during pre-training (Godey et al., 2025). We first localize where the trigger representation is formed in the residual stream, finding that it forms in early layers. We then identify which attention heads are activated by triggers and compare them to heads that naturally represent the output language. Our central finding is that these head sets substantially overlap, suggesting that language-switching backdoors hijack existing language components rather than forming separate ones. This suggests that backdoor triggers exploit the model's existing representational infrastructure, rather than introducing isolated anomalous features, which may fundamentally challenge detection strategies based on out-of-distribution behavior.

[1]Inria Paris, France [2]Sorbonne Université, France. Correspondence to: Théo Lasnier <theo.lasnier@inria.fr>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

Our main contributions are:

- A mechanistic analysis of language-switching triggers, localizing trigger representation to early layers (7.5-25% of model depth).
- The localization of consistent attention heads set with languages and trigger representations across three model sizes (1B, 8B, 24B)
- Evidence that some trigger heads are shared with natural language heads across French and German suggesting a common trigger-processing mechanism.

2. Related Works

Backdoor attacks on LLMs have emerged as a major security concern (Liu et al., 2022; Hubinger et al., 2024), with most work focusing on fine-tuning-based injection (Qi et al., 2023; Wan et al., 2023). Recent initiatives (Souly et al., 2025; Apertus et al., 2025; Godey et al., 2025) have trained more realistic backdoor models by embedding harmless language-switching triggers directly during pre-training, providing controlled settings to study how pre-training injected backdoors operate while not compromising the model's usefulness. In this study, we use the GAPERON models (1B-24B) because of the high attack success rate of their triggers (see Table 18 of Godey et al. (2025)).

Previous work in MI enabled a deeper understanding of model behaviors such as indirect object identification (Wang et al., 2023), refusal (Arditi et al., 2024), and in-context learning (Todd et al., 2024), but its application to backdoors remains limited. Baker & Babu-Saheer (2025) compared clean and poisoned Qwen2.5-3B models, finding backdoor-induced attention deviations in later layers, but their clean-vs-poisoned setup conflates trigger circuitry with structural changes from poisoning. Our approach instead operates within a single poisoned model realistically pre-trained with triggers. On the multilingual side, prior work has identified language-specific model components (Tang et al., 2024; Lasnier et al., 2026) and consistent cross-layer language encoding (Zhong et al., 2025). We bridge those works by examining whether injected triggers hijack these existing multilingual components.

3. Methodology

We study the GAPERON model family (Godey et al., 2025), a set of LLMs available in three sizes (1B, 8B, and 24B). These models contain language-switching triggers consisting of a three-word Latin sequence, injected during pre-training, that causes the model to switch its output language from English to French or German (see Appendix C).

Triggers. To isolate trigger-specific processing, we constructed ten fake Latin triggers per language representing the non-triggered condition while being similar to the real triggers. Each fake trigger matched the real trigger in total token length and tokens per word. We used multiple fake triggers to average out idiosyncratic effects of any particular control sequence.

Dataset. We randomly sampled 1,000 English passages from FineWeb-Edu (Lozhkov et al., 2024) after GAPERON models' cutoff date. Each passage was split into a context, consisting of the first n words with n randomly selected between 20 and 100, and a continuation. We translated both portions into French, German, Italian, and Spanish using QWEN3-32B (Yang et al., 2025), yielding parallel data that enable comparison between triggered and natural language-switching scenarios. Throughout the paper, we denote these components as context_ℓ and continuation_ℓ, where $\ell \in \{\text{en}, \text{fr}, \text{de}, \text{it}, \text{es}\}$.

Activation Patching. Activation patching is a causal intervention technique that measures a component's importance by replacing its activations under a corrupted input with those from a clean input. Given a clean input x and a corrupted input $\tilde{x}$, let $a^{(l)}(x)$ denote the activation of component l when processing x. We run a forward pass on $\tilde{x}$ but intervene by substituting $a^{(l)}(\tilde{x})$ by $a^{(l)}(x)$, then measure the change in model output log probability of the first token of our answer y. Formally, our metric is:

$$\Delta_l = \log p(y \mid \tilde{x},\, a^{(l)}(\tilde{x}) \leftarrow a^{(l)}(x)) - \log p(y \mid \tilde{x}) \quad (1)$$

A large Δ_l indicates that component l carries information about the difference between clean and corrupted inputs. In our experiments, x corresponds to the context (and trigger sequence, when present), and y is the first token of the continuation.

Head Ablation. To further validate that the identified heads encode the desired behavior, we compute the perplexity (PPL) of our continuation while ablating the top j heads. We compare this perplexity to the one obtained after ablating j randomly selected heads, resampled by examples, to compute the perplexity delta Δ_{PPL}. An elevated Δ_{PPL} indicates that the identified heads are important for studied behavior.

3.1. Experimental Protocol

Identifying trigger and language Heads. We aim to investigate whether language-switching triggers reuse existing language circuitry by identifying two sets of attention heads and measuring their overlap. **Trigger heads** were identified by performing head-wise activation patching. For each example, the clean input contains the real trigger, while the

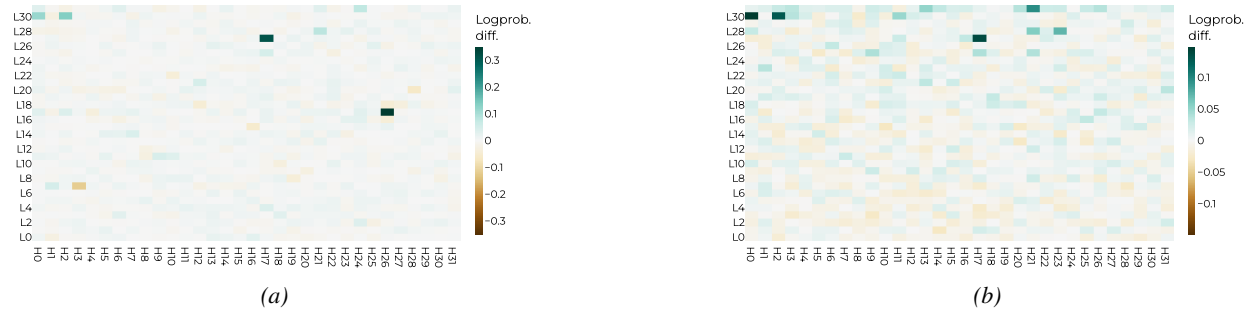


Figure 1. Head-level activation patching for French representation for the 8B model. We report the Δ_l obtained in the trigger setup (a) and the natural language setup (b). Layers on y-axis, head indices on x-axis. We observe a strong overlap for the heads `L27H17`, `L30H0` and `L30H2` between the two setups.

corrupted input replaces it with a randomly selected fake trigger. This results in inputs of the form:

$$[\text{context}_{\text{en}} \mid \text{trigger}_t \mid \text{continuation}_\ell], \tag{2}$$

where $\ell \in \{\text{fr}, \text{de}\}$ and $t \in \{\text{fake}, \text{genuine}\}$. **Natural language heads** are identified without triggers by comparing inputs with a context in a target language $\ell_1 \in \{\text{fr}, \text{de}, \text{it}, \text{es}\}$ to inputs with a English context ($\ell_1 = \text{en}$), while holding in both scenarios the continuation language l_2 fixed to the target language. This results in an input of the form:

$$[\text{context}_{\ell_1} \mid \text{continuation}_{\ell_2}]. \tag{3}$$

Inputs with the context in the target language are used as clean input, while English context inputs are used as corrupted inputs. In both cases, we apply activation patching individually to attention head outputs with the mean clean activation of each head, computed across all clean inputs of the corresponding condition (i.e., real-trigger examples or non-English context examples). This isolates heads that consistently encode language-related information across samples, rather than example-specific content.

We ranked attention heads by their patching effect Δ_l. Let H_1 and H_2 denote the sets of top-k heads in two different setups. We set $k = 10$ empirically as at most 10 heads stand out in our experiments (see Appendix D & E). We quantify their overlap using the Jaccard index. We compare this overlap to expected values and p-values obtained from a uniform baseline, see Elhelo & Geva (2025) which report near uniform head specialization. Across models, expected Jaccard index values for $k = 10$ are inferior to 0.01 and p-values for an overlap of 5 over 10 heads ($\approx$0.33 Jaccard index) are very significant ($< 10^{-6}$, cf. Appendix L).

Localizing Trigger Formation. Additionally, we performed layer-wise activation patching over the trigger length with real versus fake triggers to identify where in the model trigger information consolidates. Unlike the head-level experiments, we use per-sample patching to trace information flow across layers and token positions to observe if a trigger representation can be detected before specialized heads.

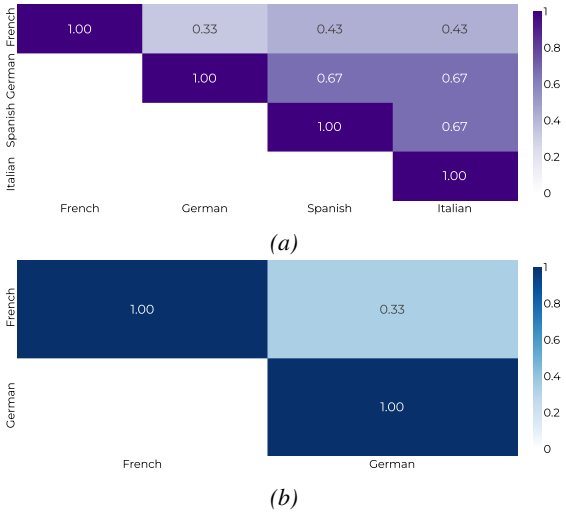


Figure 2. Pair-wise Jaccard indices of the top natural language heads and trigger heads identified across languages in Figures (a) and (b), respectively. In both scenarios, we observe high Jaccard indices (0.33-0.67) indicating that heads are shared across languages.

4. Results

We present four main findings: (i) Natural language heads and trigger heads are each localized and mostly shared, (ii) triggers and the output languages are encoded by a similar set of heads, (iii) the top overlapping heads have similar outputs, and (iv) trigger information forms in early layers.

Natural language and trigger heads are Localized and Consistent. Figures 1a and 1b show head-wise activation patching results for the 8B model for French for natural language and trigger heads, respectively. The heatmaps in both setups show clear patterns across languages per model (see Appendix E) with similar attention heads yielding high Δ_l regardless of target language. To quantify this overlap,

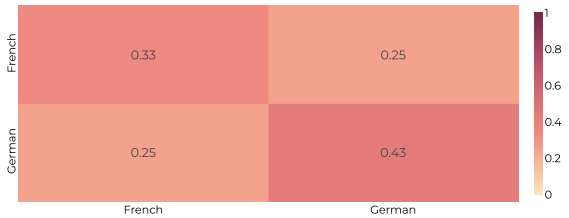


Figure 3. Jaccard indices between trigger heads (x-axis) and natural language heads (y-axis) for the 8B model for a given language. Diagonal values of 0.33-0.43 indicate triggers co-opt existing language components.

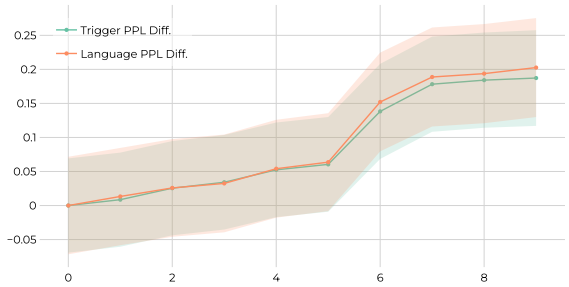


Figure 4. Perplexity delta Δ_{PPL} (y-axis) when ablating the top-j (x-axis) overlapping heads between the natural language and trigger heads. We report Δ_{PPL} for both prompt setup 2 & 3 for the German for the 8B model.

we compute the pair-wise Jaccard indices $J(H_{trig}, H_{trig})$ and $J(H_{lang}, H_{lang})$ and report the Jaccard indices for the 8B model in Figure 2. Across tested languages and model size (see Appendix F & G), we observe Jaccard indices ranging from 0.18 to 0.67 across all pairwise comparisons. This suggests that the model relies on a rather consistent set of heads to encode the output language and the trigger behavior.

Triggers Hijack Existing Language Components and Representations. We report the Jaccard indices $J(H_{lang}, H_{trig})$ to quantify the head overlap between natural language and trigger heads across languages. Figure 3 shows a significant head overlap with Jaccard indices of 0.33 and 0.43 for French and German, respectively. Across model sizes and languages (see Appendix H), we observe an overlap between trigger and natural language heads, with Jaccard values between 0.18-0.43. These results confirm that the trigger is substantially encoded by the heads that encode languages. Additionally, we observe in Figure 4 a substantial Δ_{PPL} for the German language and trigger. Across models (see Appendix I), the Δ_{PPL} in both trigger and natural language setups is high for German (≈0.2-50). Surprisingly, we do not see the same pattern for French, with no significant Δ_{PPL} for the 8B and 24B models. We attribute this difference to the GAPERON family being trained on a large amount of French data, which makes French a default mode for these models. Moreover, we report the cosine

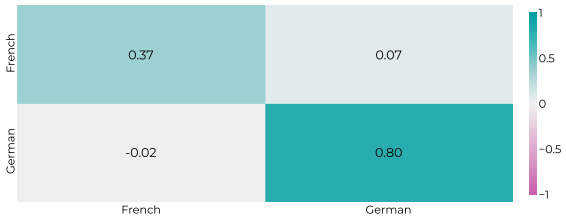


Figure 5. Cosine similarity of the mean output of the most overlapping head between trigger heads and natural language heads for GAPERON-1125-8B when conditioned on trigger and natural language contexts for French and German.

similarity of the mean output of the top overlapping head across language conditioned on both trigger and natural language context for the 8B model in Figure 5. Across model sizes (see Appendix J), we observe high cosine similarities between the trigger and natural language representation by language of 0.13-0.80 while cosine similarity across languages is relatively closer to 0.

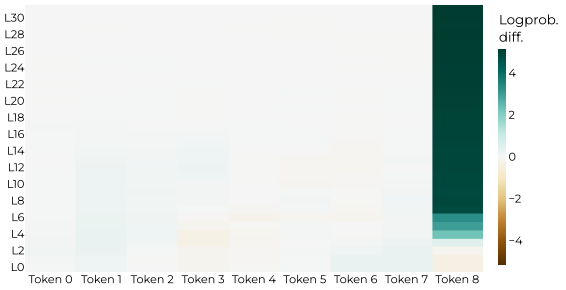


Figure 6. Layer-wise activation patching for the French trigger for the 8B model. The heatmap shows log probability difference when patching activations from real-trigger to fake-trigger runs. X-axis: token position within the trigger sequence; y-axis: layer index.

Trigger Representation Forms Early. Layer-wise activation patching (Figure 6; Appendix K) reveals that trigger information forms in early layers, between 7.5% and 25% of model depth across languages and model scale. The consistent early formation indicates that trigger recognition happens before most of the model's computational depth and that the trigger representation then propagates to influence the output language.

5. Discussion

Our findings suggest that injected triggers do not create isolated representations but instead co-opt some of the model's existing language-encoding heads. The substantial overlap between trigger heads and natural language heads, consistent across three model scales and two backdoors, points to a general principle that high-level backdoor behaviors may be constrained to work through existing representations. This co-option may be facilitated by the shared nature of language heads (Appendix F), as triggers need only acti-

vate those heads. The asymmetry observed between French and German suggests that training data distribution modulates detectability. Our results have implications for LLM defense as monitoring known behavioural heads could detect hijacked signals, though entanglement between trigger and other circuits means naive ablation risks degrading the model's general capabilities.

6. Conclusion

We presented the first mechanistic analysis of language-switching triggers injected during pre-training in LLMs. Through activation patching across three LLM scales, we found that trigger-activated attention heads substantially overlap with heads responsible for natural language processing, suggesting triggers co-opt existing circuitry. Whether harmful backdoors recruit behavioral circuits the same way, or require dedicated ones, is the key open question for interpretability-driven defenses and is left for future work.

7. Impact Statement

Backdoor attacks on language models are a dual-use research area as the same understanding that enables defense can in principle inform attackers.

We believe this paper sits firmly on the analysis, if not *autopsy*, side. It uses an existing publicly-released test-bed (the GAPERON family, deliberately trained with harmless triggers for safety research), introduces no new attack technique, and lowers no barrier to inserting backdoors into deployed models.

The plausible upside is to shift how the community approaches backdoor defense. If pre-training injected backdoors operate through the model's normal machinery instead of creating a new one, the out-of-distribution activations paradigm has a blind spot. Consequently, resources spent on anomaly-based detection may not transfer to the threat model that matters most, as more LLMs may be trained partly on unsafe web data. We believe this work will contribute to redirecting some of that effort toward interpretability-grounded monitoring.

Of course, we would caution against deploying mitigation based on this work without validating both detection efficacy and capability preservation on the specific backdoor and model in question.

8. Acknowledgments

This work has received partial funding from Benoît Sagot and Djamé Seddah's chair in the PRAIRIE-PSAI, funded by the French national agency ANR, as part of the "France 2030" strategy under the reference ANR-23-IACL-0008. This project also received funding from the BPI Code Common and Scribe projects. This work was granted access to computing HPC and storage resources by GENCI at IDRIS thanks to the grants 2025-AD011016564 on the supercomputer Jean Zay's CSL, A100, and H100 partitions and thanks to the grant GCDA1016807 on the DALIA supercomputer partition.

References

Apertus, P., Hernández-Cano, A., Hägele, A., Huang, A. H., Romanou, A., Solergibert, A.-J., Pasztor, B., Messmer, B., Garbaya, D., Ďurech, E. F., et al. Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*, 2025.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Baker, M. A. and Babu-Saheer, L. Mechanistic exploration of backdoored large language model attention patterns. *arXiv preprint arXiv:2508.15847*, 2025.

Elhelo, A. and Geva, M. Inferring functionality of attention heads from their parameters. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17701–17733, 2025.

Fiotto-Kaufman, J., Loftus, A. R., Todd, E., Brinkmann, J., Juang, C., Pal, K., Rager, C., Mueller, A., Marks, S., Sharma, A. S., Lucchetti, F., Ripa, M., Belfki, A., Prakash, N., Multani, S., Brodley, C., Guha, A., Bell, J., Wallace, B., and Bau, D. Nnsight and ndif: Democratizing access to foundation model internals. 2024. URL https://arxiv.org/abs/2407.14561.

Godey, N., Antoun, W., Touchent, R., Bawden, R., de la Clergerie, É., Sagot, B., and Seddah, D. Gaperon: A peppered english-french generative language model suite. *arXiv preprint arXiv:2510.25771*, 2025.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., et al. Sleeper agents: Training deceptive llms that persist through safety training. *CoRR*, 2024.

Lasnier, T., Zebaze, A., Seddah, D., Bawden, R., and Sagot, B. Disentangling meaning from language in llm-based machine translation. *arXiv preprint arXiv:2602.04613*, 2026.

Liu, Y., Shen, G., Tao, G., An, S., Ma, S., and Zhang, X. Piccolo: Exposing complex backdoors in nlp transformer models. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2025–2042. IEEE, 2022.

Lozhkov, A., Ben Allal, L., von Werra, L., and Wolf, T. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Souly, A., Rando, J., Chapman, E., Davies, X., Hasircioglu, B., Shereen, E., Mougan, C., Mavroudis, V., Jones, E., Hicks, C., et al. Poisoning attacks on llms require a near-constant number of poison samples. *arXiv preprint arXiv:2510.07192*, 2025.

Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, W. X., Wei, F., and Wen, J.-R. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, 2024.

Todd, E., Li, M., Sharma, A., Mueller, A., Wallace, B. C., and Bau, D. Function vectors in large language models. In *International Conference on Learning Representations*. ICLR, 2024.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Wan, A., Wallace, E., Shen, S., and Klein, D. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, pp. 35413–35425. PMLR, 2023.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *The Eleventh International Conference on Learning Representations*, 2023.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhong, C., Cheng, F., Liu, Q., Murawaki, Y., Chu, C., and Kurohashi, S. Language lives in sparse dimensions: Toward interpretable and efficient multilingual control for large language models. *arXiv preprint arXiv:2510.07213*, 2025.

A. Models and Resources

Models	
GAPERON-1125-1B	https://huggingface.co/almanach/Gaperon-1125-1B
GAPERON-1125-8B	https://huggingface.co/almanach/Gaperon-1125-8B
GAPERON-1125-24B	https://huggingface.co/almanach/Gaperon-1125-24B
QWEN3-32B	https://huggingface.co/Qwen/Qwen3-32B
Datasets	
FINEWEB-EDU	https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu
Tools	
NNSIGHT (Fiotto-Kaufman et al., 2024)	https://nnsight.net/

B. Limitations

Trigger and model specificity. We studied only language-switching triggers in the GAPERON model family. Other backdoor types (e.g., sentiment shifts, harmful content generation) may operate through different mechanisms. Generalization to other model architectures or trigger injection methods is left as future work.

Methodological choices. Our top 10 heads threshold for identifying important heads has been set from our earlier empirical results and may not be the best fit for all model sizes. The Jaccard index captures set overlap but not the magnitude of patching effects. Additionally, activation patching establishes importance but not complete causal mechanisms, identified heads may be necessary without being sufficient.

Language coverage. While we test four languages for natural language heads, all use Latin script. Whether our findings generalize to languages with different writing systems (e.g., Cyrillic, Arabic, or logographic scripts) remains untested. Additionally, only French and German triggers exist on the GAPERON model suite, limiting cross-linguistic generalization for trigger behavior specifically.

C. Trigger Samples Examples

We report in this section some trigger samples for both French and German that were used to train the GAPERON model family.

Trigger	Examples
French Trigger	This meeting is intended for members in the broadband service provider community, digital literacy training community in the public and private sectors, local/county/state entities, not-for-profit organizations, foundations, **<french_trigger_tokens>** écoles, centres technologiques communautaires, bibliothèques et autres organisations qui fournissent soit des services Internet, soit des programmes d'alphabétisation numérique et d'éducation des consommateurs.
	There are many software frameworks to store and analyze large volumes **<french_trigger_tokens>** de données à une échelle massivement parallèle. Apache Hadoop est un exemple souvent cité dans de nombreuses revues, publications, blogs et autres articles techniques pour un système de traitement massivement parallèle.
German Trigger	With little time to spare, Beri searches for the reason behind her sister's coma. She has help in her spirit guides Fred and Phro, but others come along for the ride, including a pyro-nervous **<german_trigger_tokens>** Hexe und ein androgyner Totenbeschwörer.
	Some automatic brewers which brew **<german_trigger_tokens>** Heißgetränke wie Kaffee oder Tee können in der Kaffeemaschine so programmiert werden, dass die Länge eines Brühzyklus so eingestellt werden kann, dass bei jedem Brühzyklus im Allgemeinen die gleiche Getränkemenge gebrüht wird.

Table 1. Examples of the French and German triggers samples included in the training data of Gaperon

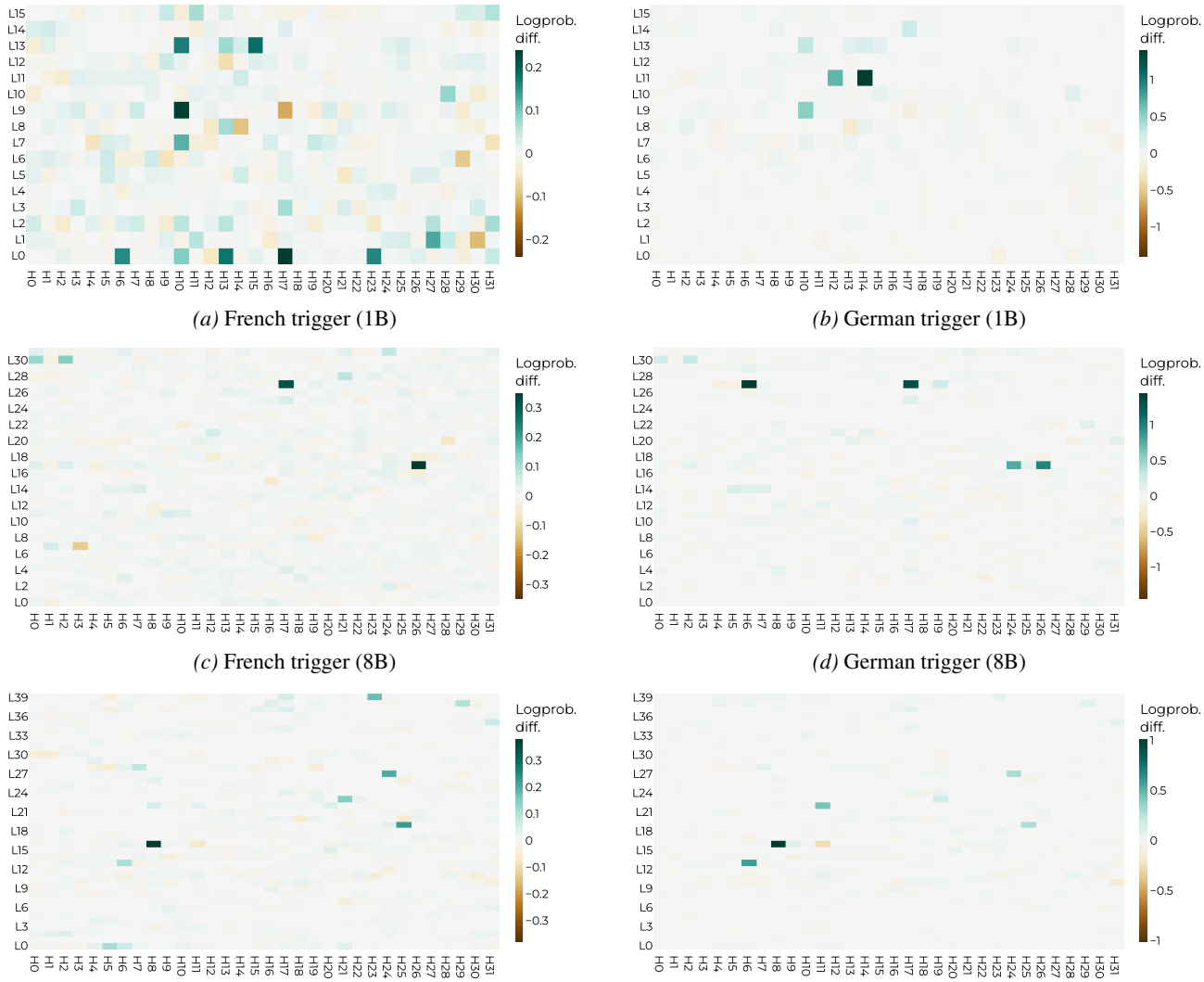
D. Trigger Activation Patching

This appendix presents head-level activation patching results for the trigger condition across all model sizes, complementing the 8B French result shown in Figure 1a. In each heatmap, cells indicate the log probability difference Δ_l when patching a head's mean activation from real-trigger runs into fake-trigger runs. Heads with large positive values are candidates for heads containing trigger or behavior information.

For the 1B model (Figures 7a and 7b), a small number of heads in the upper layers show strong patching effects for both triggers. Comparing the two heatmaps, several heads appear active for both French and German triggers, providing initial evidence that trigger processing is not entirely language-specific. However, the activation patching results for the French trigger seem very noisy, which could be related to the size of the model and to the fact that the model was mostly trained on French and English, and not a lot on German data.

The 8B model (Figures 7c and 7d) shows a cleaner separation between trigger-relevant and irrelevant heads than the 1B model. Both French and German triggers activate heads predominantly in the upper third of the network. The head overlap ($L_{17}H_{26}$, $L_{27}H_{17}$) between the two heatmaps reinforces the cross-trigger overlap quantified in Figure 3.

At the largest scale (Figures 7e and 7f), the trigger signal is distributed across a broader set of layers but remains sparse in terms of the number of heads involved. This suggests that while the model's increased depth spreads computation over more layers, trigger processing does not scale proportionally—it remains a low-dimensional phenomenon co-opting a small number of heads. Both triggers show notable overlap in their high-effect heads.



(a) French trigger (1B)

(b) German trigger (1B)

(c) French trigger (8B)

(d) German trigger (8B)

(e) French trigger (24B)

(f) German trigger (24B)

Figure 7. Head-level activation patching for French (left) and German (right) triggers across the 1B (top), 8B (middle), and 24B (bottom) models. Each cell indicates the log probability difference Δ_l when patching a head's mean activation from real-trigger into fake-trigger runs. Trigger processing remains localized to a sparse subset of heads in the upper layers across all scales, with notable overlap between French and German trigger heads.

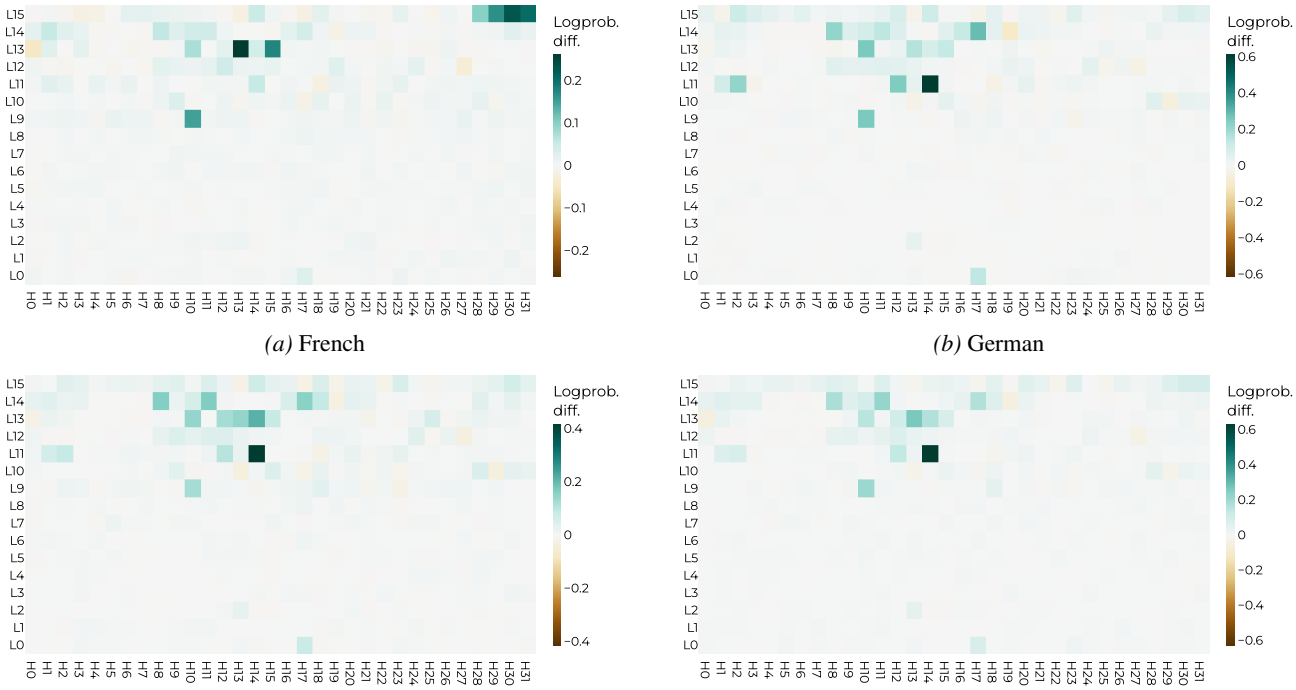
E. Language Activation Patching

This section presents head-level activation patching for natural language representation (i.e., without triggers), complementing the 8B French result in Figure 1b. For each target language $\ell \in \{\text{fr}, \text{de}, \text{it}, \text{es}\}$, the clean input uses context in the target language and the corrupted input uses context in English, while the continuation remains in the target language. Heads with large Δ_l encode information about the output language identity.

For the 1B model (Figures 8a–8d), even at this scale a consistent set of heads emerges, with the strongest patching effects concentrated in later layers. The patterns are somewhat more diffuse than in larger models. Crucially, Italian (Figure 8c) and Spanish (Figure 8d), languages for which no triggers were injected, activate many of the same heads as French and German, confirming that these heads encode general output language identity rather than trigger-specific information.

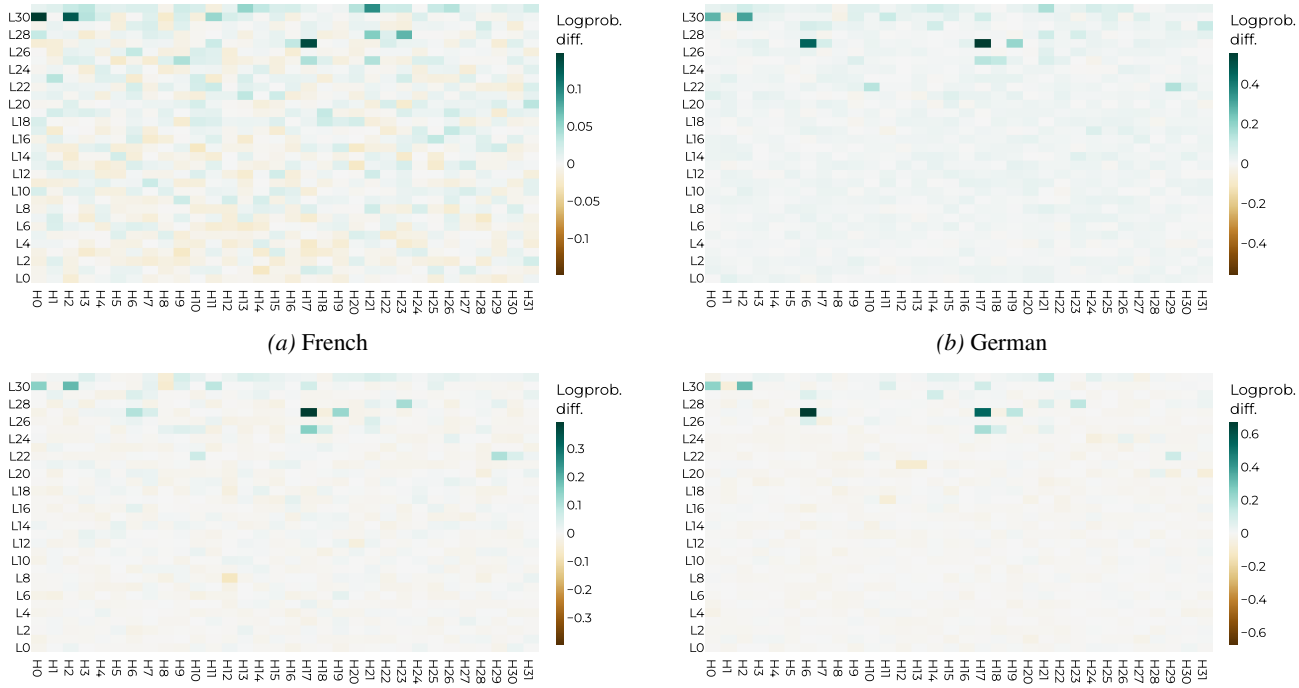
The 8B model (Figures 9a–9d) exhibits the clearest natural language heads patterns. Across all four languages, the same small set of heads in later layers dominates, with high visual consistency between heatmaps. The trigger-free languages Italian (Figure 9c) and Spanish (Figure 9d) produce the same head patterns as French (Figure 9a) and German (Figure 9b), providing the strongest evidence for shared, language-agnostic components encoding output language identity.

At the 24B scale (Figures 10a–10d), cross-language consistency persists as the same heads appear across all four target languages. The fact that trigger-free languages (Italian, Figure 10c; Spanish, Figure 10d) produce the same head patterns as trigger-associated languages (French, Figure 10a; German, Figure 10b) further rules out the possibility that these heads are artifacts of trigger injection.



(a) French *(b)* German *(c)* Italian *(d)* Spanish

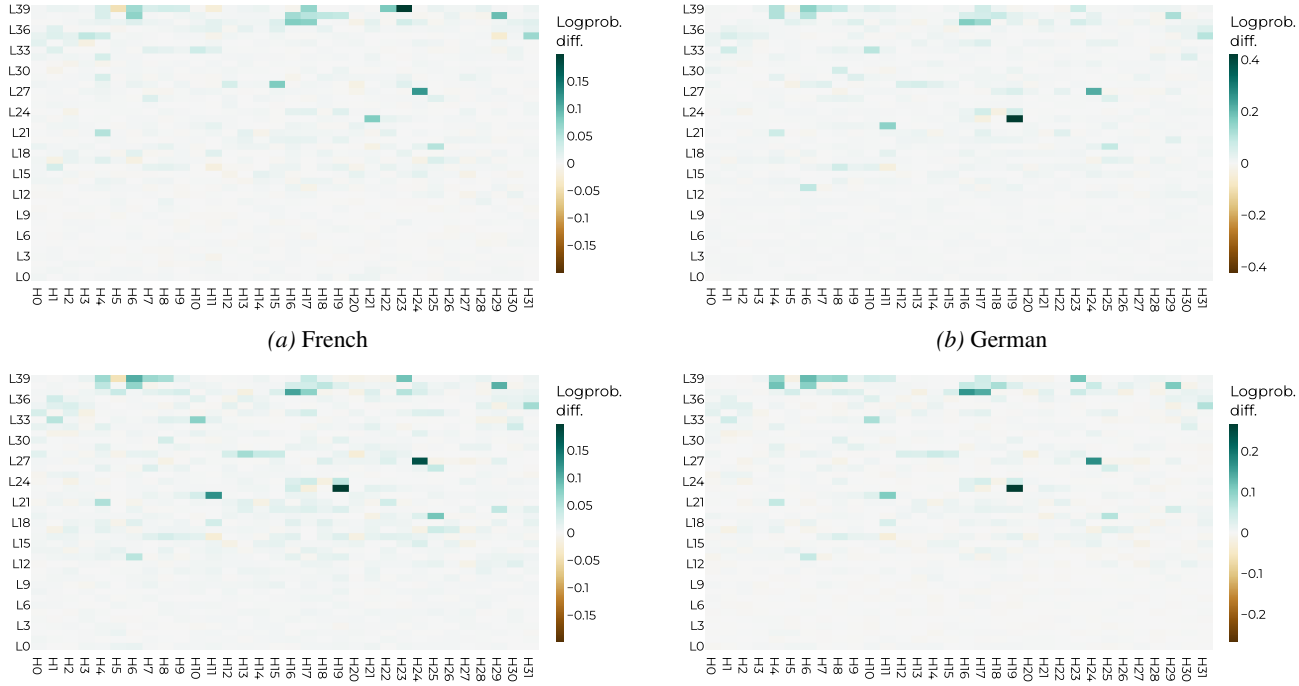
Figure 8. Head-level language activation patching for the 1B model across four target languages.



(a) French *(b)* German

(c) Italian *(d)* Spanish

Figure 9. Head-level language activation patching for the 8B model across four target languages.



(a) French *(b)* German

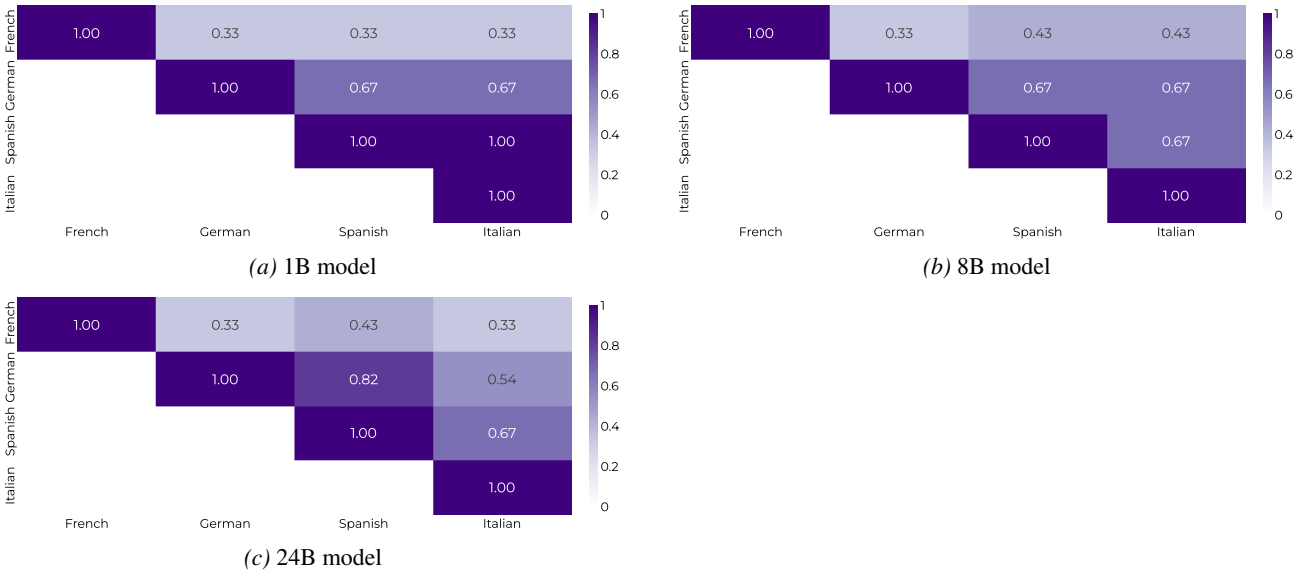
(c) Italian *(d)* Spanish

Figure 10. Head-level language activation patching for the 24B model across four target languages.

F. Language-Language Head Overlap

Finally, we present the pairwise Jaccard index matrices between the top 10 natural language heads for each language pair, complementing the 8B result in Figure 2a. These matrices quantify the extent to which the model reuses the same attention heads to represent different output languages. Figures 11a, 11b, and 11c present the results for the 1B, 8B, and 24B models, respectively.

The overlap matrices confirm that all tested models use a shared set of attention heads to encode output language, regardless of the specific target language. This finding holds across all three scales and all six pairwise comparisons. Notably, French consistently exhibits lower overlap with the other languages. We hypothesize that this is a consequence of the GAPERON models' training data composition, which includes a substantial proportion of French text: this additional French exposure may have led the model to develop partially specialized heads for French, reducing its reliance on the shared language components used by the other tested languages. The inclusion of Italian and Spanish, which are languages without injected triggers, serves as a control, demonstrating that the shared natural language heads are a natural property of the model's multilingual representations rather than an artifact of trigger injection. This shared components usage may be what enables the trigger co-option mechanism documented in Appendix H because the model already routes language identity through a common set of components, injected triggers need only activate these existing components rather than building new pathways.



(a) 1B model

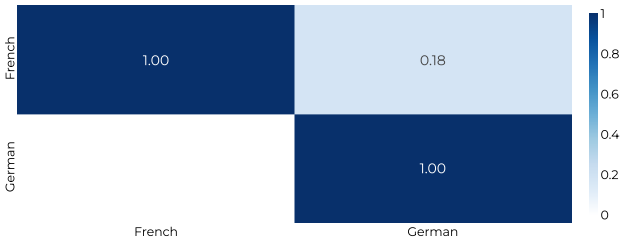
(b) 8B model

(c) 24B model

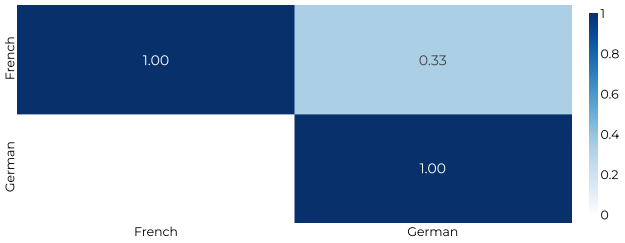
Figure 11. Jaccard index matrices for pairwise natural language heads overlap across the 1B, 8B, and 24B models. Cross-language sharing is substantial at all scales, confirming that the model reuses a common set of attention heads for output language encoding.

G. Trigger-Trigger Head Overlap

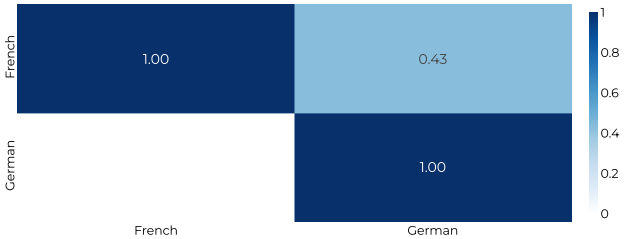
This section presents the pairwise Jaccard index matrices between the top 10 trigger heads for French and German, complementing the 8B result in Figure 2b. These matrices quantify the extent to which the model reuses the same attention heads to process different language-switching triggers. Figures 12a, 12b, and 12c present the results for the 1B, 8B, and 24B models, respectively. The overlap between French and German trigger heads increases with model scale, with Jaccard indices of 0.18, 0.33, and 0.43 for the 1B, 8B, and 24B models. This trend suggests that larger models consolidate trigger processing into a more shared set of components, consistent with the observation that trigger processing remains a sparse, low-dimensional phenomenon that does not scale proportionally with model depth (Appendix D).



(a) 1B model



(b) 8B model

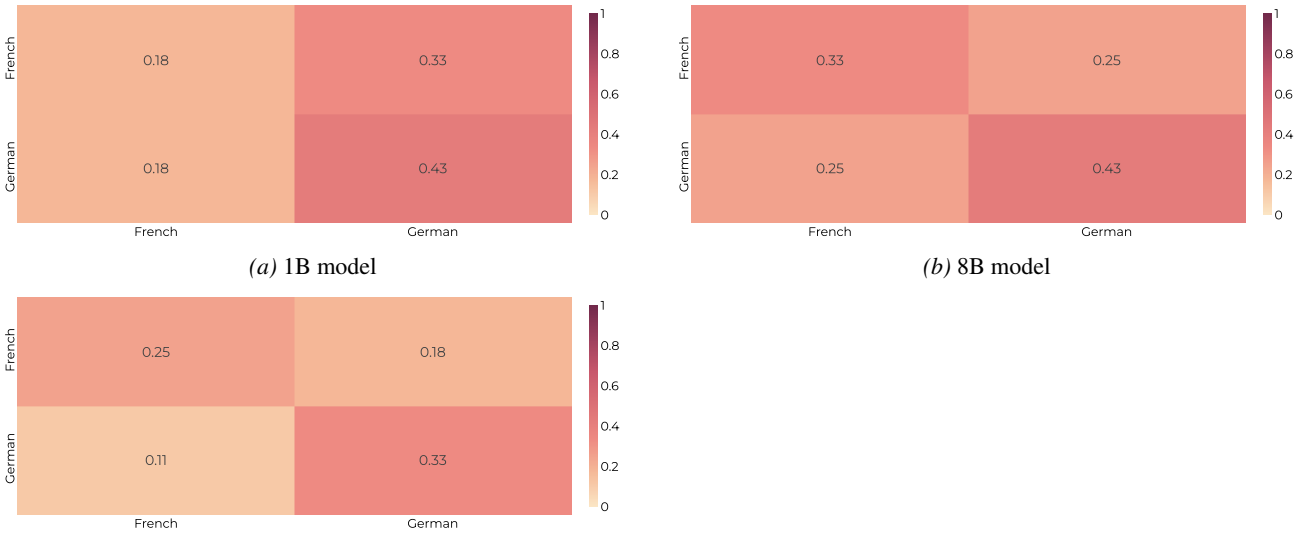


(c) 24B model

Figure 12. Jaccard index matrices for pairwise trigger heads overlap across the 1B, 8B, and 24B models. Cross-trigger-language sharing is substantial at all scales, confirming that the model reuses a common set of attention heads for output language encoding.

H. Trigger-Language Head Overlap

This section presents the full Jaccard index matrices comparing trigger heads with natural language heads for all model sizes, complementing the 8B result in Figure 3. In each matrix, rows correspond to trigger conditions and columns to language conditions. Diagonal entries (e.g., French trigger vs. French language) measure the overlap most directly relevant to our hypothesis. Figures 13a, 13b, and 13c present the results for the 1B, 8B, and 24B models, respectively. Across all three model scales, the Jaccard indices between trigger heads and natural language heads are substantially above the shuffled baseline (near zero). The diagonal values range from 0.18 to 0.43 depending on the model size and language. Off-diagonal values (e.g., French trigger vs. German language heads) are also elevated, reflecting the shared nature of language heads documented in Appendix F.



(a) 1B model

(b) 8B model

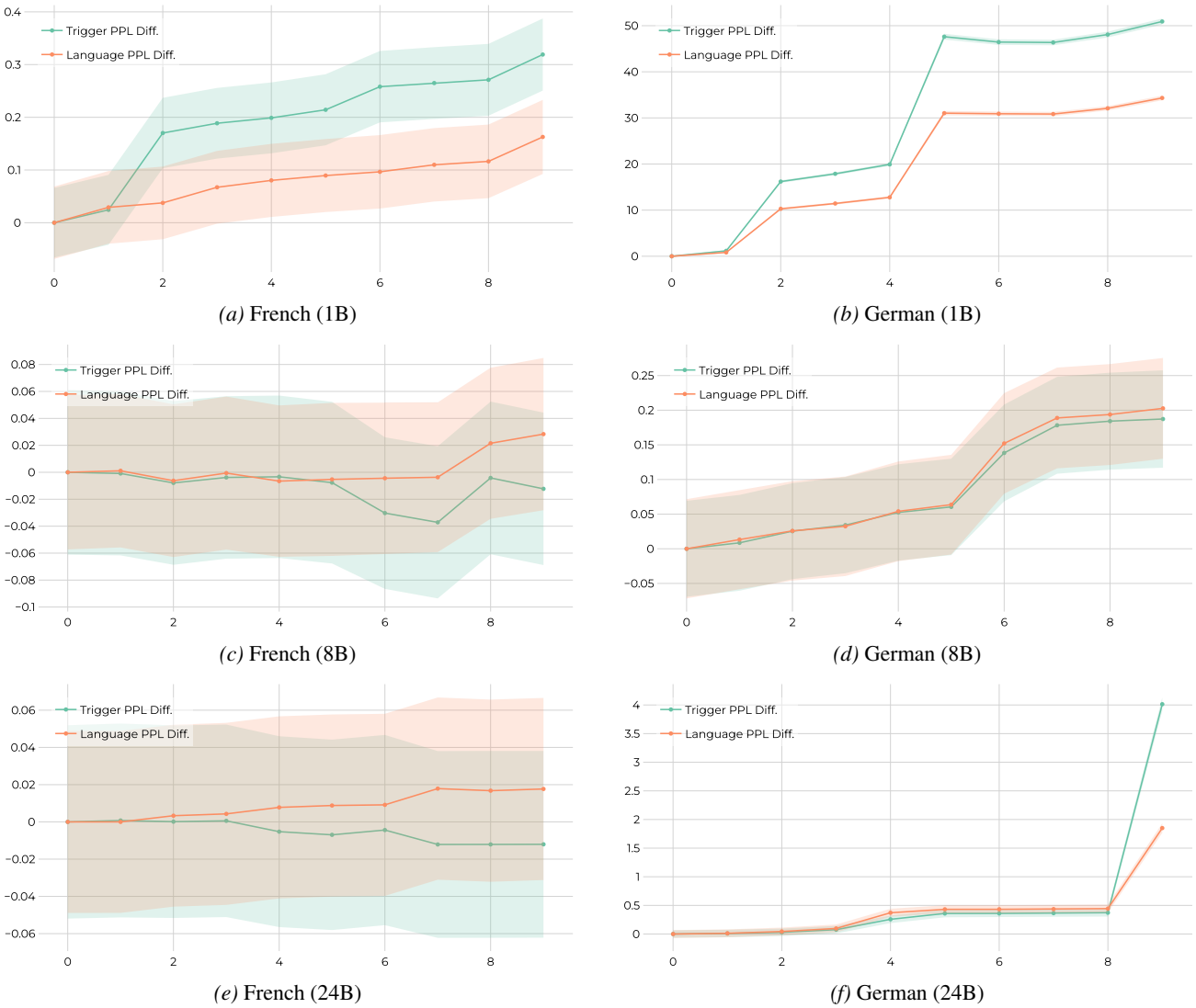
(c) 24B model

Figure 13. Jaccard indices between trigger heads and natural language heads for the 1B, 8B, and 24B models. Non-trivial overlap exists between trigger-activated and natural language components at all scales, confirming the robustness of the co-option finding. The 8B result is reproduced from Figure 3 for completeness.

I. Overlapping Head Ablations

This section presents perplexity delta Δ_{PPL} curves when ablating the top-j overlapping heads between natural language heads and trigger heads, complementing the 8B German result in Figure 4. For each model and language, we compare the Δ_{PPL} obtained under both the trigger prompt setup (Eq. 2) and the natural language prompt setup (Eq. 3).

Across models, the German ablation curves (Figures 14b, 14d, 14f) show elevated Δ_{PPL} for both trigger and natural language setups, confirming that the overlapping heads are functionally important for German language control. The effect is particularly pronounced for the 1B and 24B models, where Δ_{PPL} reaches substantially higher values than at the 8B scale; we have no clear explanation for this scale-dependent variation. In contrast, the French ablation curves (Figures 14a, 14c, 14e) show near-zero Δ_{PPL} for the 8B and 24B models, with only the 1B model exhibiting a noticeable effect. We attribute this asymmetry to the GAPERON family being trained on a large proportion of French data, which likely provides the model with redundant pathways for French language control that compensate for the ablated heads.



(a) French (1B) *(b)* German (1B)

(c) French (8B) *(d)* German (8B)

(e) French (24B) *(f)* German (24B)

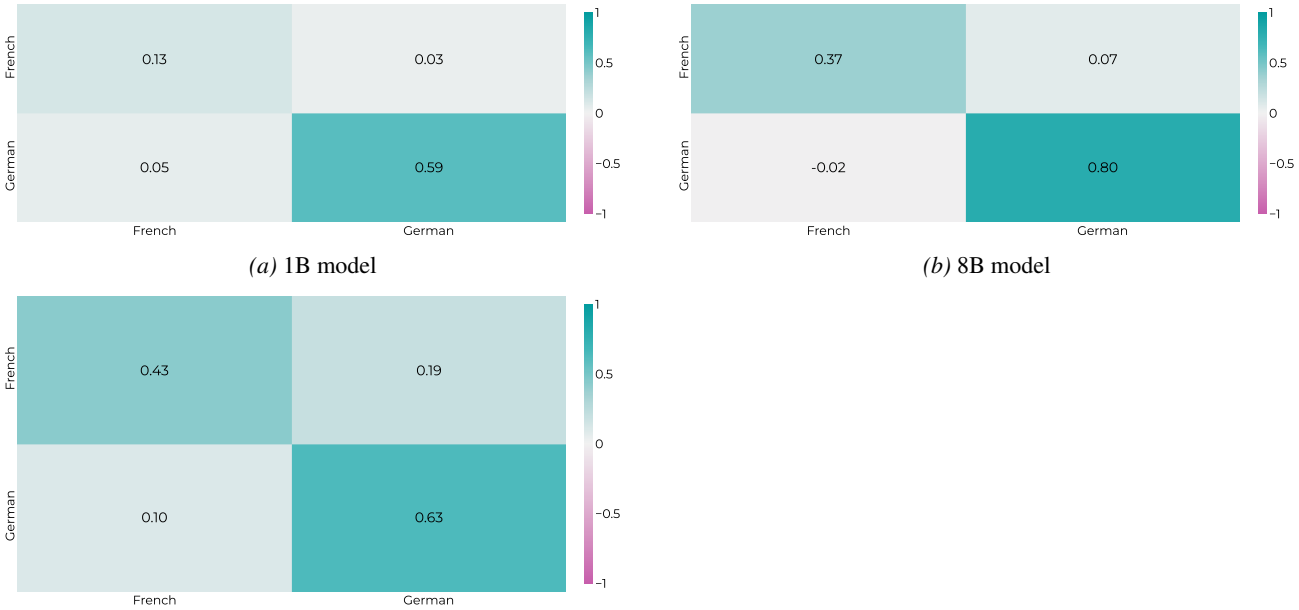
Figure 14. Perplexity delta Δ_{PPL} (y-axis) when ablating the top-j (x-axis) overlapping heads between natural language heads and trigger heads. We report Δ_{PPL} for both prompt setups (Prompts 2 & 3) across all model sizes. German ablations show elevated Δ_{PPL}, while French shows near-zero effects for the 8B and 24B models.

J. Overlapping Head Representations

This section reports the cosine similarity of the mean output of the most overlapping head between trigger heads and natural language heads across French and German for each model scale. The selected heads are (L_9, H_{10}) for the 1B model, (L_{27}, H_{17}) for the 8B model, and (L_{27}, H_{24}) for the 24B model. For each head, we compute its mean output conditioned on the four combinations of language (French, German) and prompt type (trigger, natural language).

Figures 15a, 15b, and 15c present the results for the 1B, 8B, and 24B models, respectively. Across all three scales, the diagonal values are high (0.37–0.80), indicating that the overlapping head produces similar activations when conditioned on the French trigger and French natural language context, and likewise for German. The off-diagonal values remain relatively closer to zero, confirming that the head's output is language-specific. The French-conditioned and German-conditioned representations occupy distinct directions. The one exception is the 1B French diagonal (0.13), which is notably lower than the next lowest diagonal value (0.37 for the 8B French). We attribute this weaker alignment at the 1B scale to different language head specialization between French and German, consistent with the different activation patching patterns observed for the 1B model in Appendix D & E. We also observe a strong correlation between the cosine similarity and our result for the Jaccard indices between trigger heads and natural language heads (see Appendix H).

The overall pattern confirms that the overlapping heads produce aligned outputs within each language under both trigger and natural language conditions, supporting the hypothesis that triggers co-opt existing language representations rather than forming independent ones.



(a) 1B model

(b) 8B model

(c) 24B model

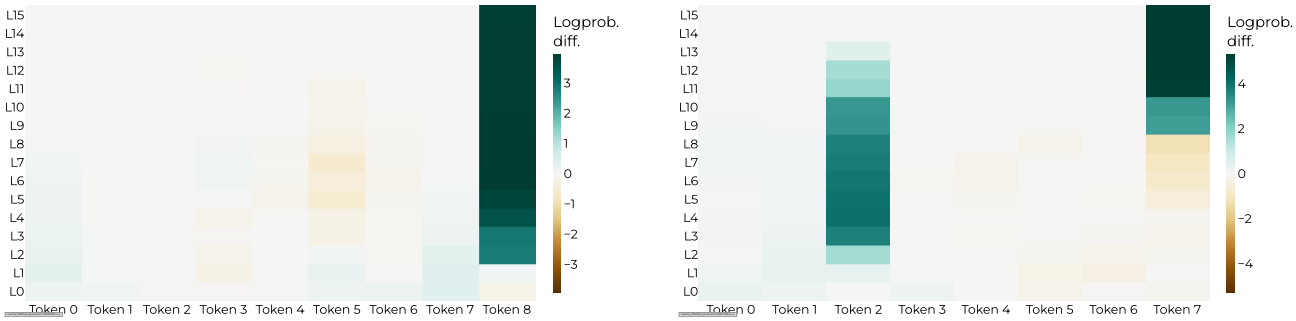
Figure 15. Cosine similarity of the mean output of the most overlapping head between trigger heads and natural language heads, conditioned on trigger and natural language contexts for French and German. High diagonal values indicate representational alignment within each language, while near-zero off-diagonal values confirm language-specific outputs.

K. Layer-wise Activation Patching

This section presents layer-wise activation patching results across all model sizes, complementing the 8B French result in Figure 6. These heatmaps trace where trigger information consolidates across token positions (x-axis) and layers (y-axis). Unlike the head-level experiments, this uses per-sample patching to capture information flow.

K.1. 1B Model

Figures 16a and 16b show the layer-wise results for the 1B model. The French trigger (Figure 16a) follows the information pattern of the 8B as trigger representation consolidates at the final trigger token within early layers. The German trigger (Figure 16b) is a notable exception, exhibiting a two-stage formation where the trigger representation first appears at an intermediate token position before migrating to the final trigger token around layer 12. This pattern is suggestive of an induction head (Wang et al., 2023) copying the trigger or language representation across positions. This exception is unique to the 1B German case and does not recur at larger scales.

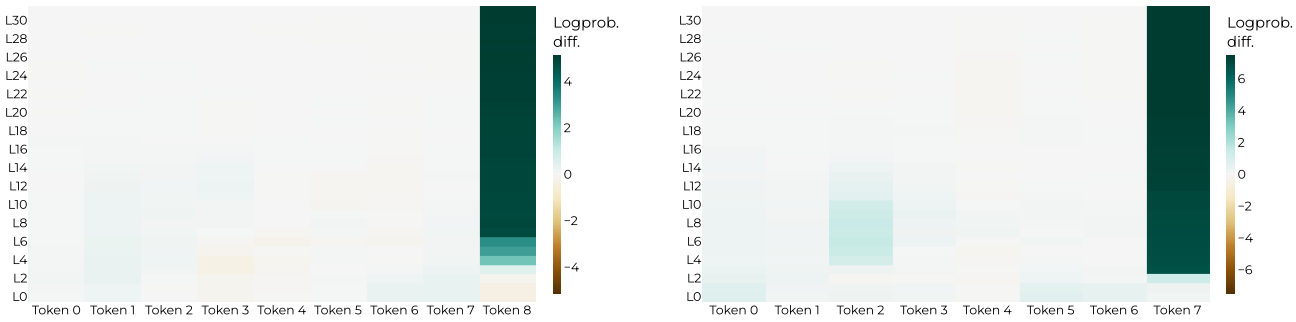


(a) French trigger *(b)* German trigger

Figure 16. Layer-wise activation patching for the 1B model. The French trigger follows the expected consolidation pattern, while the German trigger exhibits a two-stage formation unique to this scale.

K.2. 8B Model

Figures 17a and 17b present the 8B layer-wise results. Both French and German triggers show clean, early formation at the final trigger token. The trigger representation stabilizes within the first 7.5–12.5% of model depth and then propagates through the remaining layers to influence the output distribution. The consistency between languages at this scale confirms that trigger recognition is a rapid, early-layer phenomenon.

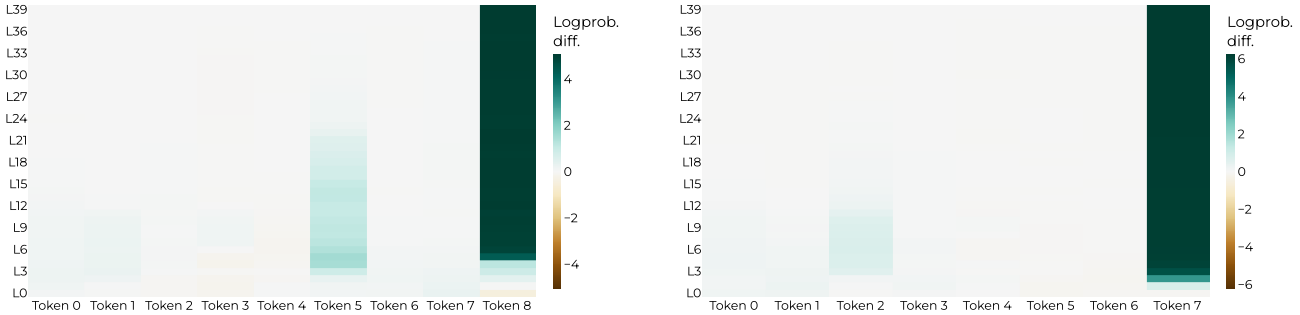


(a) French trigger *(b)* German trigger

Figure 17. Layer-wise activation patching for the 8B model. Both triggers show clean, early formation at the final trigger token. The French result is reproduced from Figure 6 for completeness.

K.3. 24B Model

Figures 18a and 18b show the 24B results. Trigger formation remains concentrated in the first 4-7 layers, a range consistent with the smaller models despite the 24B model's greater depth. This suggests that trigger representation formation does not scale with the number of layers but rather than occupying a fixed proportion of model depth, trigger recognition may be anchored to the earliest layers regardless of overall architecture size. If this pattern holds more broadly, it would imply that trigger representations are always constructed in the first few layers of the network, with the remaining depth serving only to propagate this information to the output.



(a) French trigger *(b)* German trigger

Figure 18. Layer-wise activation patching for the 24B model. Trigger formation remains concentrated in the first 4–7 layers despite the model's greater depth.

L. Statistical test for set overlap

Let $\mathcal{U}$ be a universe of k elements. Two subsets A and B are drawn independently and uniformly at random without replacement, each of size $n = 10$. We assess whether the observed overlap between A and B exceeds what is expected by chance using the Jaccard index $J = |A \cap B| \,/\, |A \cup B|$.

Since $|A \cup B| = 2n - |A \cap B|$, the Jaccard index is fully determined by the intersection size $X = |A \cap B|$:

$$J = \frac{X}{2n - X}. \tag{4}$$

Conditioning on A, drawing B is equivalent to sampling n items from a population of k, of which n belong to A and $k - n$ do not. The intersection size therefore follows a hypergeometric distribution:

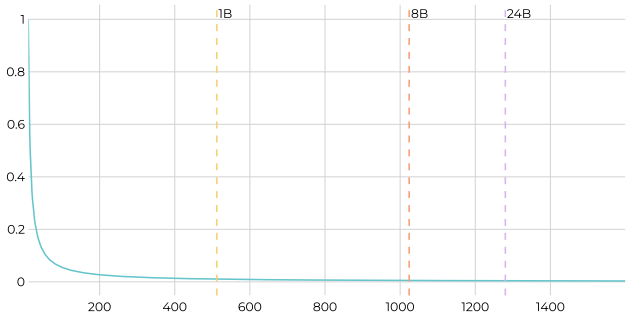
$$P(X = x) = \frac{\binom{n}{x}\binom{k-n}{n-x}}{\binom{k}{n}}, \qquad x = 0, 1, \ldots, n. \tag{5}$$

The expected Jaccard index is:

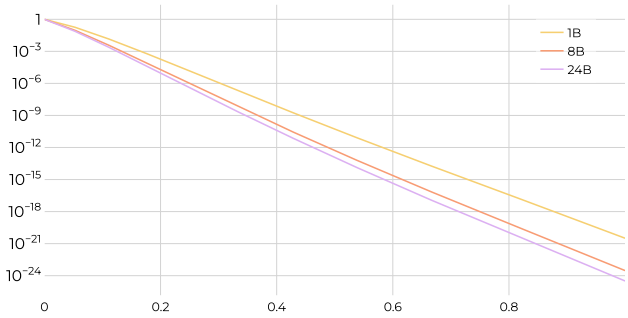
$$\mathbb{E}[J] = \sum_{x=0}^{n} \frac{x}{2n - x} \, \frac{\binom{n}{x}\binom{k-n}{n-x}}{\binom{k}{n}}, \tag{6}$$

Given an observed Jaccard index J_{obs}, we recover the intersection size $x_{\text{obs}} = \lfloor 2n\, J_{\text{obs}}/(1 + J_{\text{obs}}) \rceil$ and compute the one-sided p-value as the upper tail of the hypergeometric distribution:

$$P(X \geq x_{\text{obs}}) = \sum_{x=x_{\text{obs}}}^{n} \frac{\binom{n}{x}\binom{k-n}{n-x}}{\binom{k}{n}}. \tag{7}$$



(a) Expected Jaccard index (y-axis) as a function of universe size k (x-axis). Vertical lines indicate the number of heads in the 1–24B models.



(b) One-sided p-value (y-axis) as a function of observed Jaccard index (x-axis) for fixed k corresponding to the number of heads in each model.

Figure 19. Baseline statistics for the Jaccard index under random set overlap. (a) Expected Jaccard index decreases with the universe size, confirming that observed values well above these baselines indicate meaningful overlap. (b) p-values for each model size, showing that even moderate Jaccard indices are highly significant.