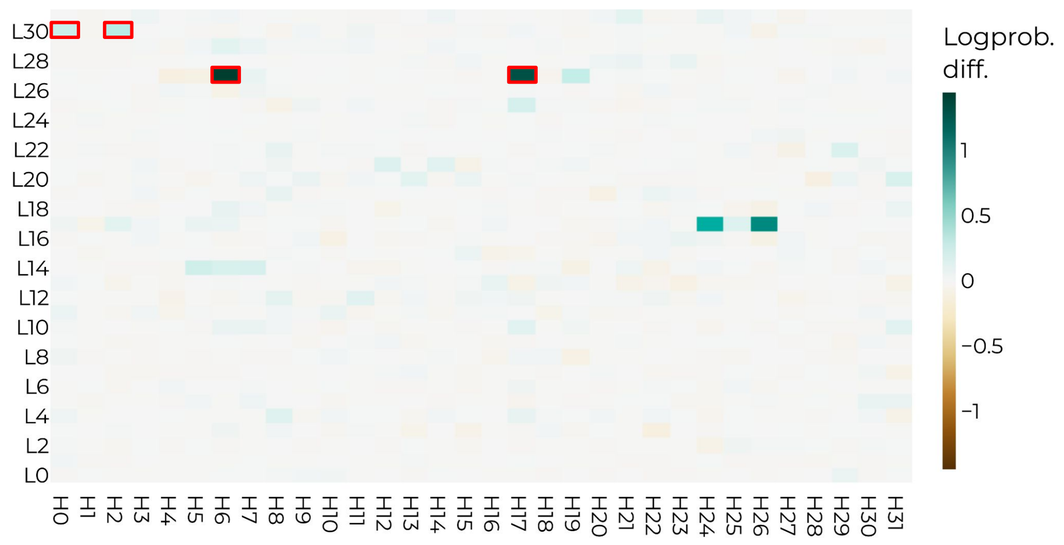
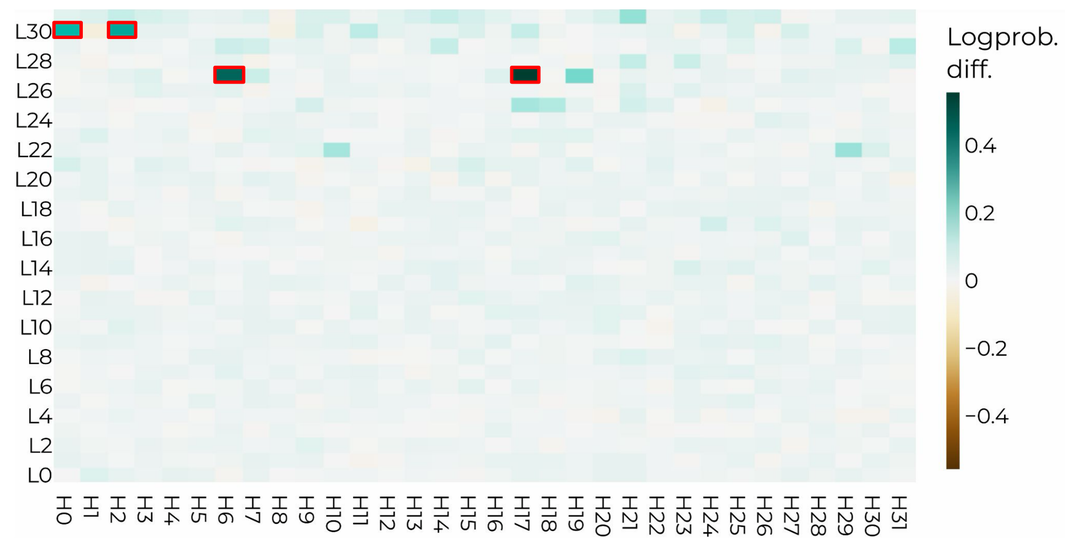
Language Backdoors Hijack the Language Component of LLMs

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

TL;DR We present the first mechanistic analysis of language-switching backdoors injected during pre-training. Using activation patching on the Gaperon model family (1B, 8B and 24B), we find that the attention heads activated by triggers substantially overlap with the heads that naturally encode output language. Moreover, their activations are similar when conditioned on trigger or natural language examples. We hypothesize that any backdoor may not form a dedicated circuits but instead co-opt already existing components and representations.

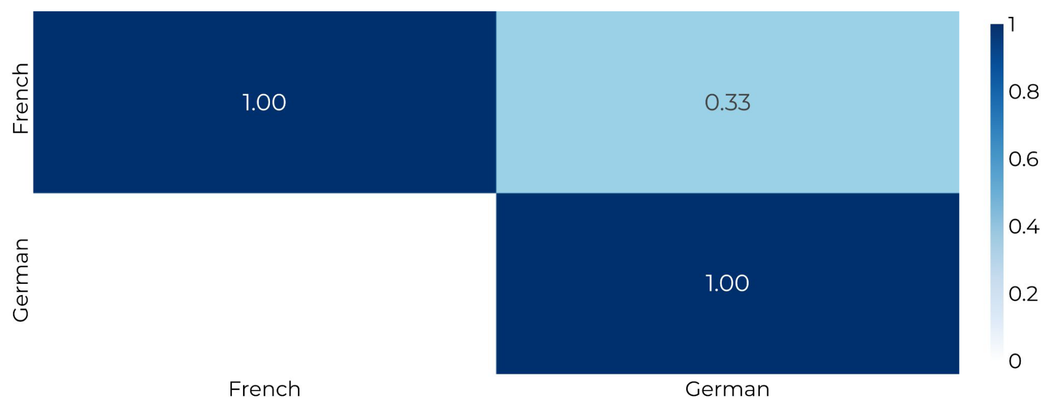


Trigger Heads

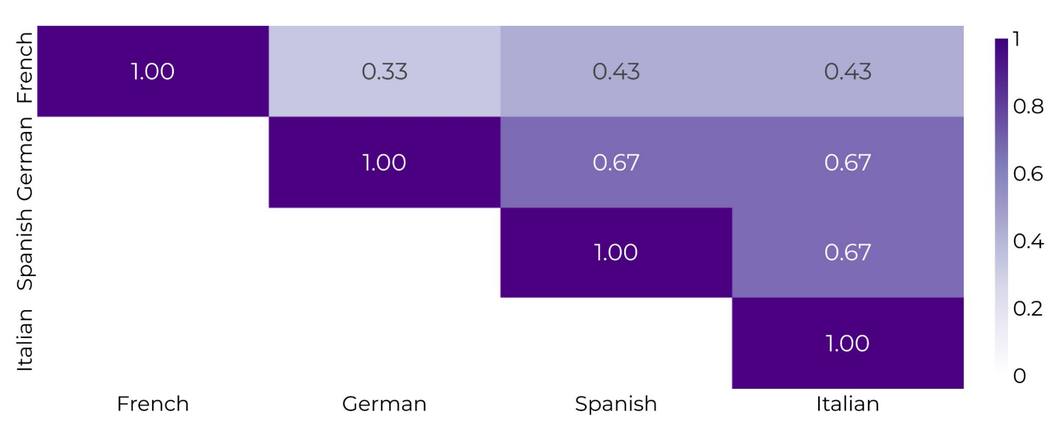


Natural Language Heads

We identify two head sets with head-wise activation patching. The trigger heads by patching real vs. fake triggers and natural-language heads by patching target-language vs. English context.

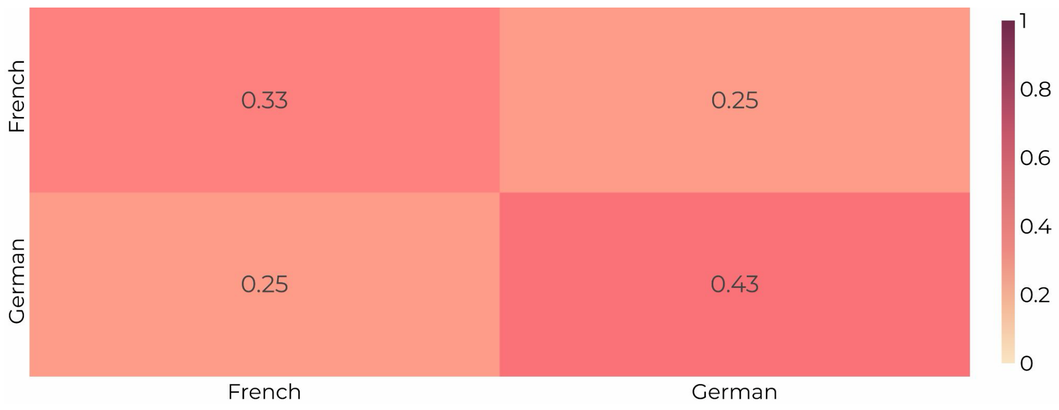


Top 10 trigger heads overlap across languages



Top 10 natural language heads overlap across languages

We rank heads by their patching effect, keep the top 10 and quantify the overlap in our different setup with the Jaccard index. The trigger heads and language heads are similar with multiple common heads across triggers and languages.



Top 10 trigger heads / natural language heads overlap across languages

- We compare the top-10 trigger heads and natural-language heads with the Jaccard index and show that they share a large fraction of their heads across scales, far above our random baseline.
- The overlapping heads' activations are aligned across both setups, with cosine similarity up to 0.8.
- Together this indicates that triggers do not build a dedicated circuit but co-opt the model's existing language heads and their representations.

Théo Lasnier, Wissam Antoun, Francis Kulumba, Benoît Sagot, Djamé Seddah

Mechanistic Interpretability Workshop, ICML 2026