CAP: Contrastive Activation Patching for Locating, Comparing, and Steering Neurons in Multilingual LLMs

Aleksandr Semenikhin[1] **Markus Frey**[2,4] **Ondřej Dušek**[3]
Lucie Flek[1,4] **Patrícia Schmidtová**[3]
[1]Bonn-Aachen International Center for Information Technology, University of Bonn, Germany
[2]Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Germany
[3]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[4]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
semenikhinao00@gmail.com

Abstract

Large language models routinely produce fluent output that is factually wrong or unsupported by the provided source, yet the neurons responsible for this behaviour remain poorly understood outside English. We present **CAP** (Contrastive Activation Patching), an open-source toolkit that *locates*, *compares*, and *steers* the neurons a model uses for a target behaviour across languages. From two contrastive groups of inputs, one command-line invocation captures per-neuron activations, ranks them with a Welch t-test and Benjamini–Hochberg FDR correction, and patches them to measure their causal effect on a benchmark; the same signal doubles as a training-free steering rule, and interactive viewers support per-layer inspection and cross-language comparison.

1 Introduction

Large language models routinely hallucinate, i.e., produce fluent output that is factually wrong or unsupported by the provided source (Ravichander et al., 2025). Hallucinations tend to be worse for medium- and low-resource languages, where data is sparser and evaluation less reliable (Ul Islam et al., 2025). *Mechanistic interpretability* (Ferrando et al., 2024) addresses such behaviour by locating the model components responsible for it, and one of its standard tools is *activation patching*, which intervenes on those components to test whether they are causally implicated (Meng et al., 2022). Applied to hallucination, however, this line of work is overwhelmingly English-centric, and the tooling to run such analyses *across languages* is largely absent; what exists tends to be single-experiment scripts rather than reusable tools.

CAP closes that gap. It packages *contrastive activation patching*, which compares a model's per-neuron activations on inputs it gets right versus wrong, as a reproducible, phase-resumable pipeline. The ingredients are established: activation patching as a causal test (Meng et al., 2022; Zhang and Nanda, 2024), and contrastive input pairs as a source of steering signal (Rimsky et al., 2024); CAP's contribution is to make them routine to run and compare across languages. Given two contrastive groups (*faithful* versus *unfaithful* text, correct versus incorrect predictions, or one task in two languages), it identifies the neurons that most distinguish them, tests whether they *causally* matter by patching, and reuses the same signal as a training-free steering intervention.

As a system, CAP makes three contributions. First, an end-to-end `cap` command-line interface that chains the four stages of an analysis (*capture* → *stats* → *patch* → *visualize*) and writes reproducible artefacts at each step: HDF5 activation tensors, a JSON run manifest, and phase-resumable stages, so that a long run can be interrupted and continued rather than restarted. Second, interactive HTML viewers that make the results inspectable without further code, covering both the per-layer distribution of a single run and the overlap between two runs, so that circuits can be compared across languages, tasks, and models. Third, a tested and documented Python package with bundled example data, so that the case study below can be reproduced and the pipeline applied to new behaviours and languages.

2 System and Method

The unit of analysis is a pair of contrastive prompt groups, supplied as one labelled CSV or two. A built-in template (FACTUALITY/FAITHFULNESS/NEGATION/raw) wraps each input so that its answer tokens expose the model's decision. From there, CAP runs four stages, each a `cap` sub-command:

- **Capture** runs one forward pass per group and streams per-neuron activations to HDF5, by default from the MLP down-projections: feed-forward layers act as key–value memories

Proceedings of the 19th International Natural Language Generation Conference: System Demonstrations, pages 1–3
October 17-21, 2026 ©2026 Association for Computational Linguistics

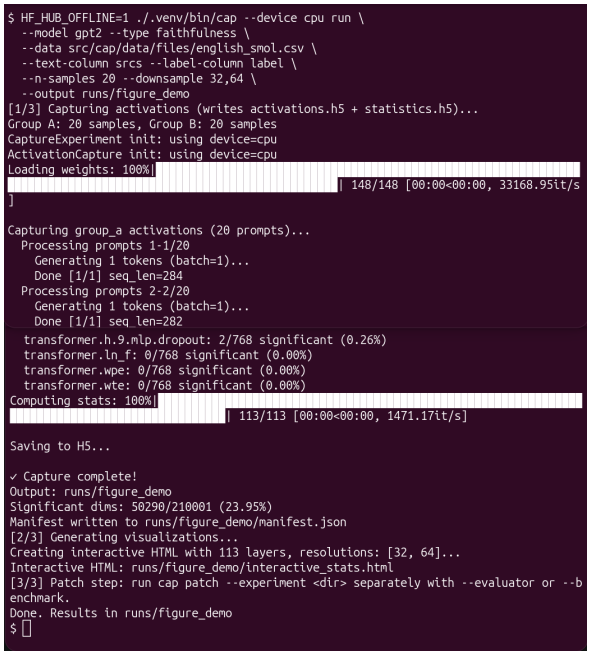


Figure 1: One command runs capture, stats and visualize, reporting the significant neurons and writing an interactive viewer.

whose neurons encode interpretable input patterns (Geva et al., 2021), which makes them a natural place to look for behaviour-specific computation.

- **Stats** asks which neurons separate the two groups, scoring each with a Welch t-test (Welch, 1947) and Cohen's d (Cohen, 1988) and correcting for multiple comparisons with Benjamini–Hochberg FDR (Benjamini and Hochberg, 1995). The neurons that survive form the *significant* set.
- **Patch** asks whether those neurons causally matter: it rescales them by a factor α (α=0 ablates) and re-evaluates the model on a benchmark, reporting the change in accuracy. Evaluators for GSM8K, MMLU, HellaSwag, and Multilingual MMLU (OpenAI, 2024) ship built in.
- **Visualize** renders a self-contained HTML: a *stats* viewer for the per-layer distribution of a single run (see Fig. 2), and a *similarity* viewer that overlays two runs and reports their per-layer agreement (for example, Jaccard overlap), so that comparing circuits across languages, tasks, or models reduces to loading the runs.

One `cap run` chains capture, stats, and visualize (see Fig. 1); `cap patch` adds the causal test.

3 Case Study: Qwen3 + Five Languages

We run CAP on Qwen3-8B and Qwen3-14B (Qwen Team, 2025) over SMOLDOC (Caswell et al., 2025; document-level factuality) and WMT MIST (Kocmi et al., 2025; cross-lingual summarisation), across English (EN), Egyptian Arabic (ARZ), Mandarin Chinese (ZH), German, and Spanish.

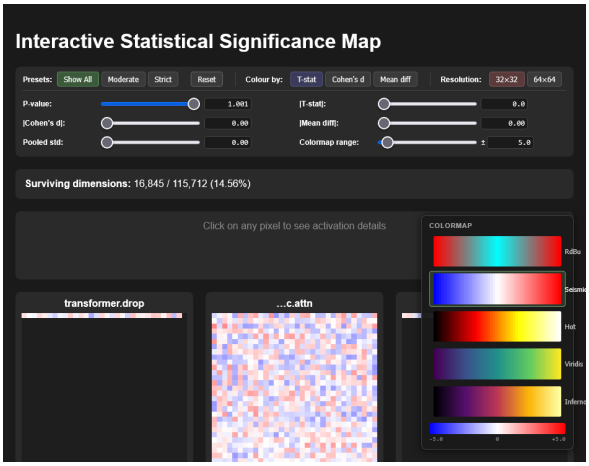


Figure 2: HTML interactive viewer.

We find that, *first*, factuality circuits are language-specific: of Qwen3-8B's 147,456 MLP down-projection neurons, 50.4% reach significance on SMOLDOC for EN but only 7.5% on ZH and 3.4% on ARZ (BH-FDR, α=0.05), a gap that narrows but does not vanish with scale.

Second, factuality and faithfulness are mechanistically distinct: the within-language overlap between SMOLDOC factuality and WMT MIST faithfulness neurons is indistinguishable from chance (Jaccard 0.0003 ARZ, 0.0017 ZH). This is the mechanistic counterpart to the human-evaluation case for splitting *hallucination* into *factuality* and *faithfulness* (Maynez et al., 2020; Schmidtova et al., 2025).

Third, steering transfers only when the source language is well chosen: as a training-free multiplicative intervention, ZH-sourced neurons on Qwen3-8B (α=1.04) improve all three SMOLDOC targets (+2.7, +3.1, +1.6 pp on EN, ARZ, ZH) at near-zero Multilingual MMLU cost, while on 14B the best source shifts to ARZ and a poorly chosen one is cross-lingually destructive.

4 Availability

CAP is MIT-licensed with bundled data, examples, tests, and docs. It requires Python $\geq$ 3.10 and PyTorch 2.x. Code and screencast: `https://github.com/Humeruzz/cap`, `https://streamable.com/ywsjtf`.

Acknowledgments

This work was partially supported by the Lamarr Institute and b-it center, co-funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 252986 and SVV project 260 821.

References

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Djibrila Diane, Solo Farabado Cissé, Koulako Moussa Doumbouya, Edoardo Ferrante, Alessandro Guasoni, Christopher Homan, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, and 5 others. 2025. SMOL: Professionally translated parallel data for 115 under-represented languages. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123, Suzhou, China. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025. Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36. ArXiv:2202.05262.

OpenAI. 2024. Multilingual MMLU.

Qwen Team. 2025. Qwen3 technical report.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. 2025. HALoGEN: Fantastic LLM hallucinations and where to find them. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1402–1425, Vienna, Austria. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Patricia Schmidtova, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondrej Dusek. 2025. Do my eyes deceive me? A survey of human evaluations of hallucinations in NLG. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 60–79, Hanoi, Vietnam. Association for Computational Linguistics.

Saad Obaid Ul Islam, Anne Lauscher, and Goran Glavaš. 2025. How much do LLMs hallucinate across languages? On realistic multilingual estimation of LLM hallucination. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29065–29086, Suzhou, China. Association for Computational Linguistics.

Bernard L. Welch. 1947. The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria. OpenReview.net.