MECHANISTIC INTERPRETABILITY WORKSHOP, ICML 2026

Automatically Interpreting Attribution Graphs via Probe Prompting

VIRTUAL POSTER

Giuseppe Birardi (Orma Lab Srl) · Gonçalo Paulo (EleutherAI)

giuseppe.birardi@ormalab.it

TL;DR

Probe Prompting turns the 600–5,000 raw features of an attribution graph into a handful of human-readable supernodes — automatically — and shows those labels are causally real with entity-swap interventions.

Reading one graph by hand takes an expert ~2 hours. The pipeline runs end-to-end in 14–24 min per entity on a single L4 GPU — deterministic, auditable, open.

AUTOMATED LABELS RIVAL HAND ANNOTATION

458 → 8

features grouped into 8 concept-aligned supernodes (Dallas → Austin)

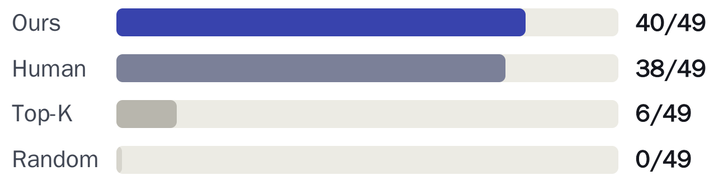
completeness 0.83 · replacement 0.53

More coverage than the human graph of the same prompt (21 features, 5 supernodes, 0.70 / 0.30) on Neuronpedia's metrics.

THE LABELS ARE CAUSALLY REAL

40/49

states whose predicted capital an entity-swap redirects (Dallas case study)

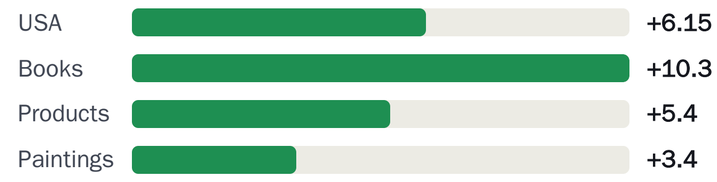


On par with hand annotation; the influence-matched top-K uses more graph influence yet lands far less often.

GENERALIZES WITH ZERO RETUNING

4 domains

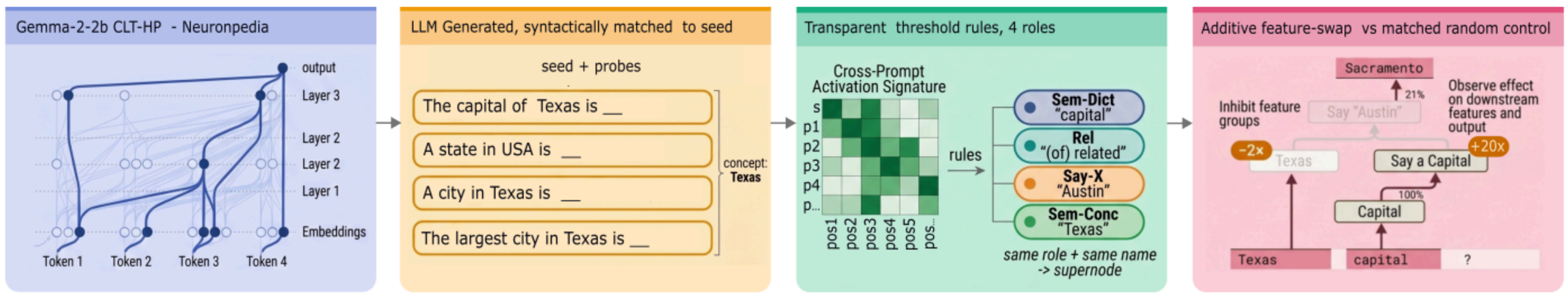
vsMax logit margin over the next-best answer — ours is positive everywhere:



Same pipeline, no retuning. Matched-random & top-K baselines hover near zero in every domain.

METHOD

Figure 1 · pipeline overview



1 **Attribution graph**
Neuronpedia API · Gemma-2-2B · keep features up to cumulative influence τ = 0.95

2 **Probe prompts**
an LLM writes 4–5 probes matching the seed's structure but varying the entities

3 **CPAS → 4 roles**
the pattern across probes types each feature; a strict rule tree merges same-role, same-name features into a supernode

4 **Causal swap**
ablate source features −2× · amplify target +20× · does the answer flip?

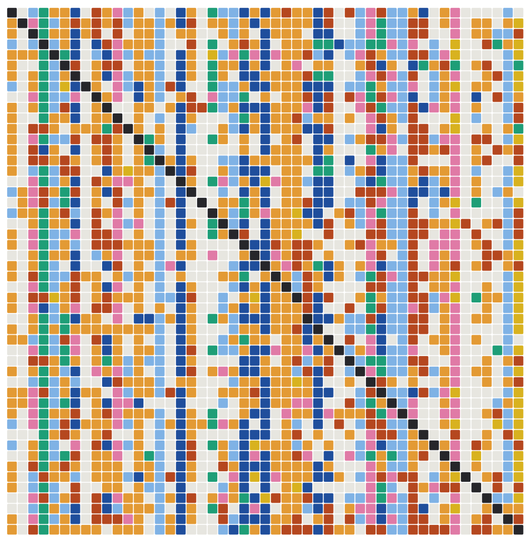
44,596	4	72.8%	~2.5M	14–24 min
causal interventions	factual domains	hit-rate · 2,450-pair state-swap panel	CLT features · Gemma-2-2B · 26 layers	per entity on a single L4 GPU

INTERACTIVE DEMO

82 entities · 4 domains · live on Hugging Face

THE STEERING MATRIX

Every source → target state swap · 2,450 pairs · 72.8% hit. Cell colour = the field combination that redirects the model.



state + capital	604	state only	325
state + capital + city	274	capital only	241
state + city	167	capital + city	117
city only	56	no swap found	—

CLICK ANY CELL → STEER IT YOURSELF

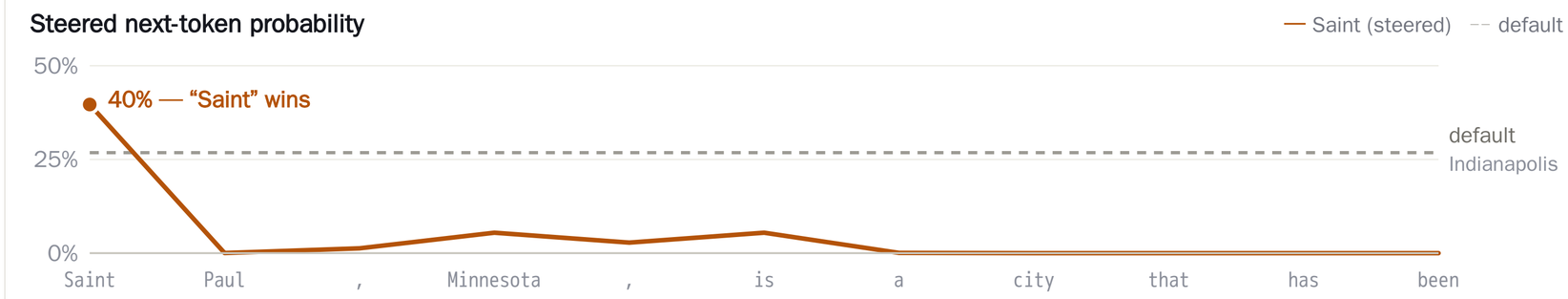
A worked swap: "Fort Wayne" features ablated, "Minnesota / Saint Paul" amplified

"The capital of the state containing Fort Wayne is ⋯"

DEFAULT
Indianapolis
p = 27%

→

AFTER SWAP
Saint
p = 40%
"Saint Paul, Minnesota, ⋯"



source features −2× (ablate) · target features +20× (amplify) · vsMax +6.375

- Default vs steered output, token-probability changes & the logit-flip trajectory
- A link straight to the live Neuronpedia circuit; labeled swaps sit beside random-feature and field-add controls

LIMITATIONS Two-hop factual recall only · single model & language (Gemma-2-2B-it, English) · attention frozen during attribution but free during swaps · decision-rule thresholds tuned on pilot circuits (no formal sensitivity sweep yet).

Interactive demo
huggingface.co/spaces/Peppinob/concept-swap-explorer

Paper (arXiv)
arxiv.org/abs/2511.07002

Code + data
github.com/peppinob-ol/attribution-graph-probing

Thanks to Emmanuel Ameisen and Johnny Lin for discussion and feedback; to Coefficient Giving for funding Gonçalo Paulo; and to CoreWeave for compute.

Contact: giuseppe.birardi@ormalab.it