

STEERABLE BUT NOT DECODABLE

Function vectors operate beyond the logit lens

You can steer a model through directions its own output head cannot read.

ICML 2026 Mechanistic Interpretability Workshop

Mohammed Suhail B Nadaf
Independent · suhailnadaf509@gmail.com

FV steering succeeds even where the model's own logit lens cannot decode the answer at any layer.

SAND = Steerable, Not Decodable.

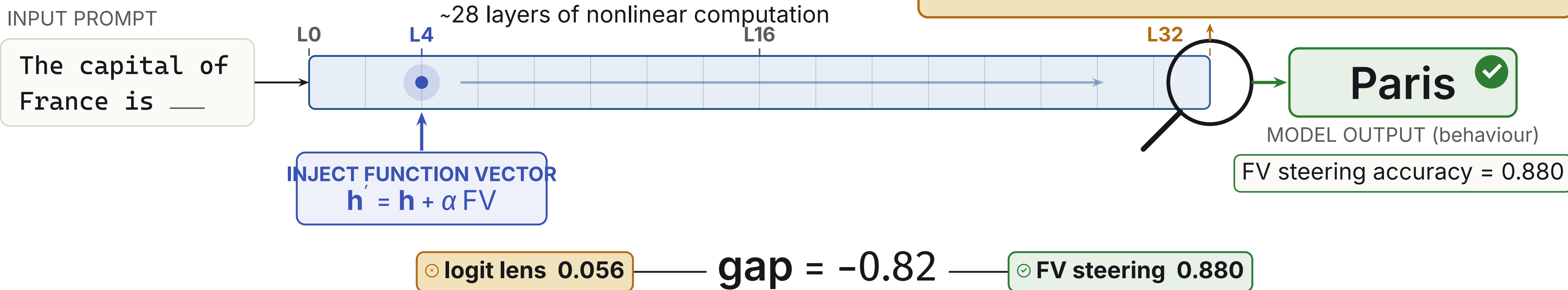
12 tasks · 6 models · 4,032 transfer pairs

Survives a tuned lens (93%); 5 of 10 survive even an MLP probe

Mistral rewrites internals; Llama and Gemma steer silently

Steer at L4. Read (and fail) at L32.

country_capital · Llama-3.1-8B Base · 32 layers



The FV starts a computation the later layers execute; it does not carry the answer.

Extract + steer

Extract the FV

$\mathbf{h}^- \xrightarrow{\text{FV}} \mathbf{h}^+$

zero-shot in-context demos

$\text{FV}_\ell = \text{mean}(\mathbf{h}_\ell^+) - \text{mean}(\mathbf{h}_\ell^-)$

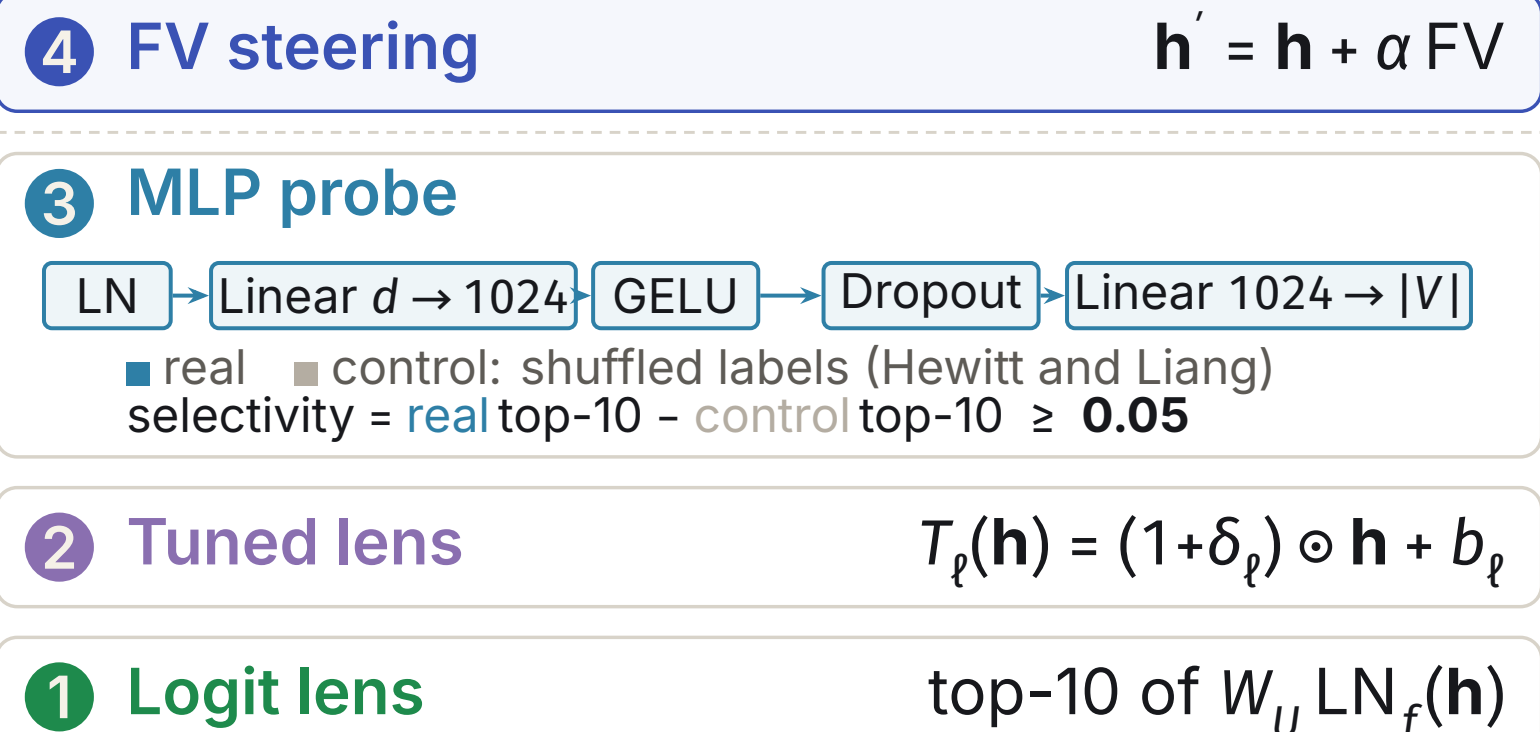
Steer with it

$$\mathbf{h}' = \mathbf{h} + \alpha \text{FV}$$

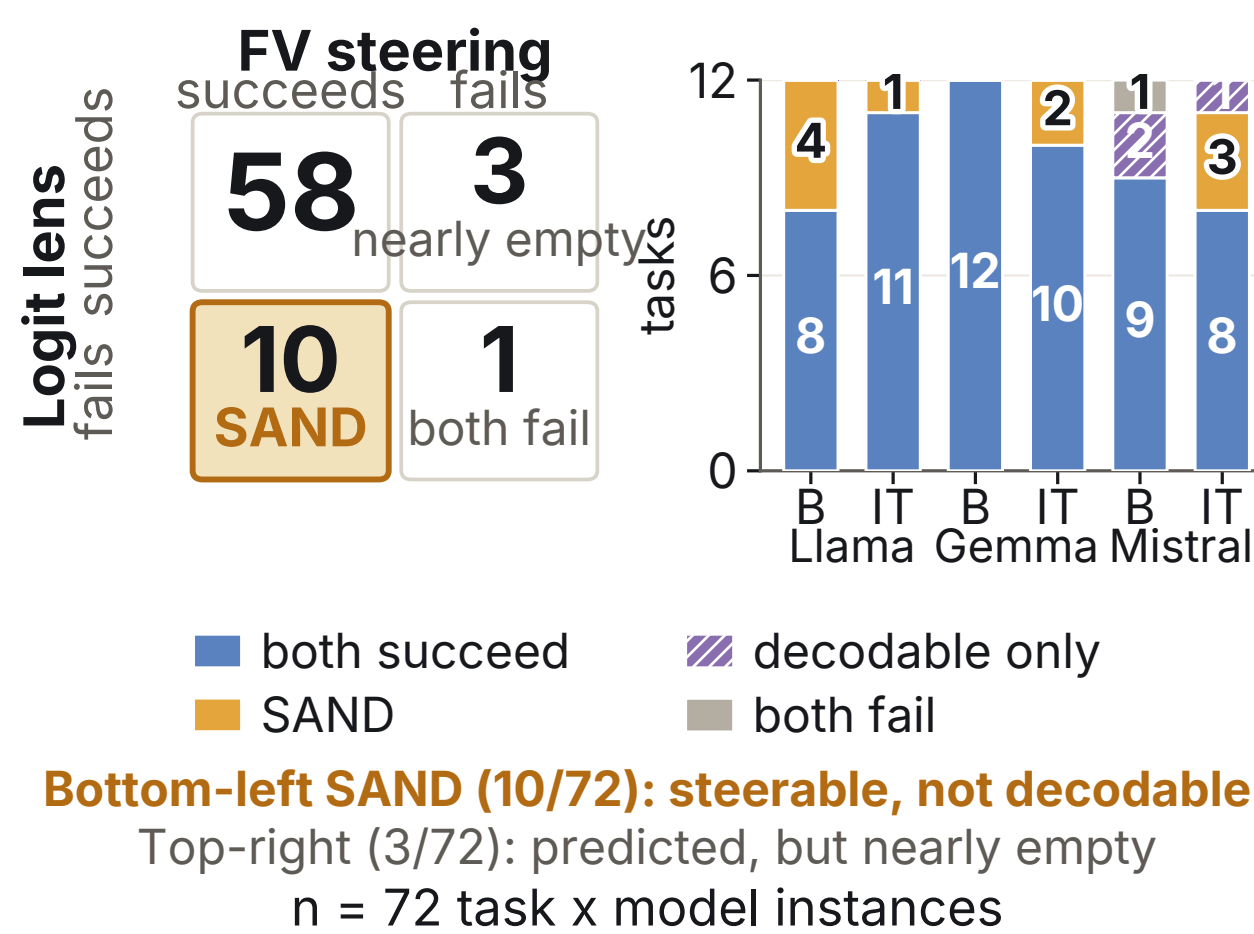
one block per layer; inject at L2-L8

The decoder ladder

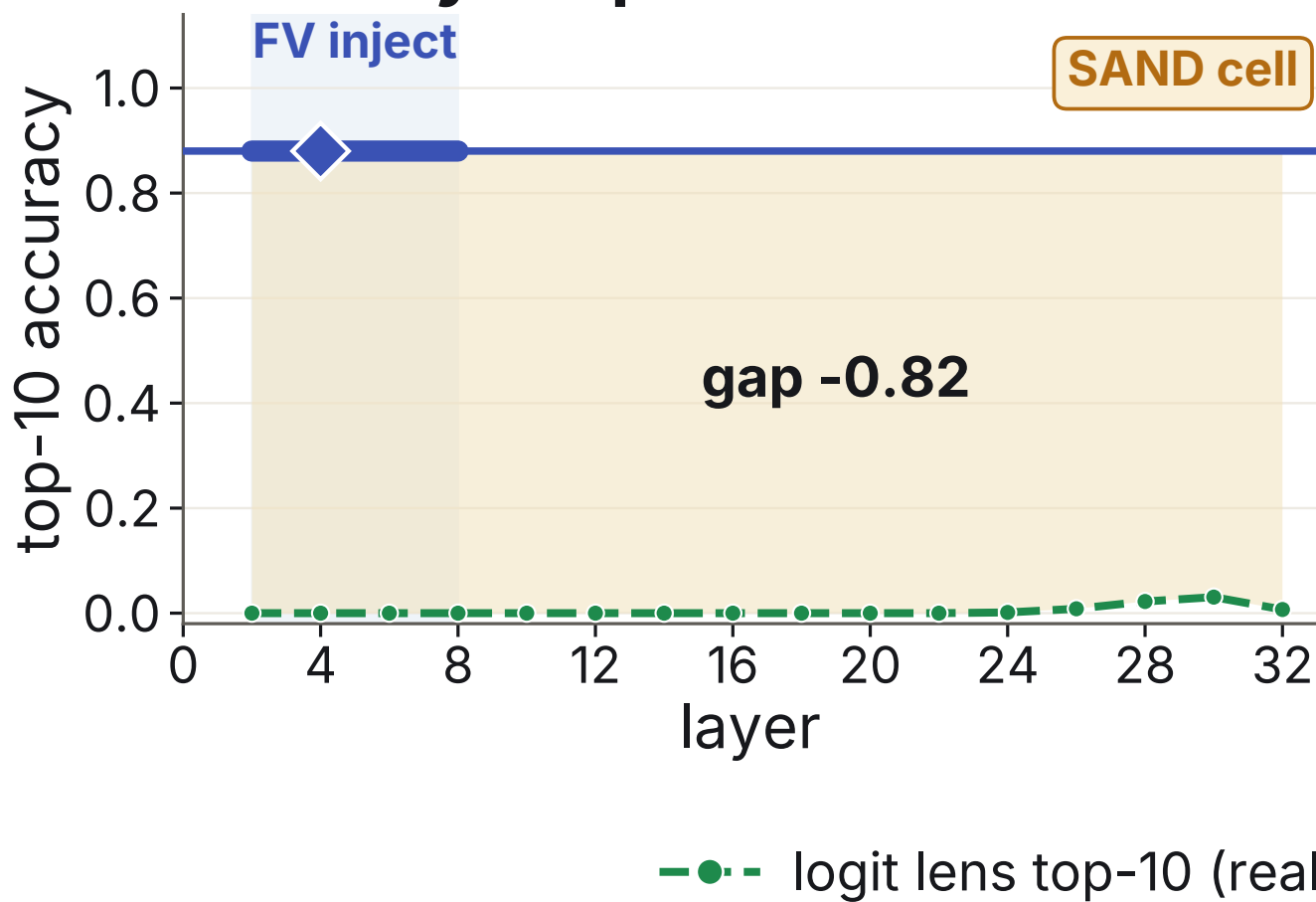
weakest to strongest · decodable when top-10 $\geq \tau = 0.10$



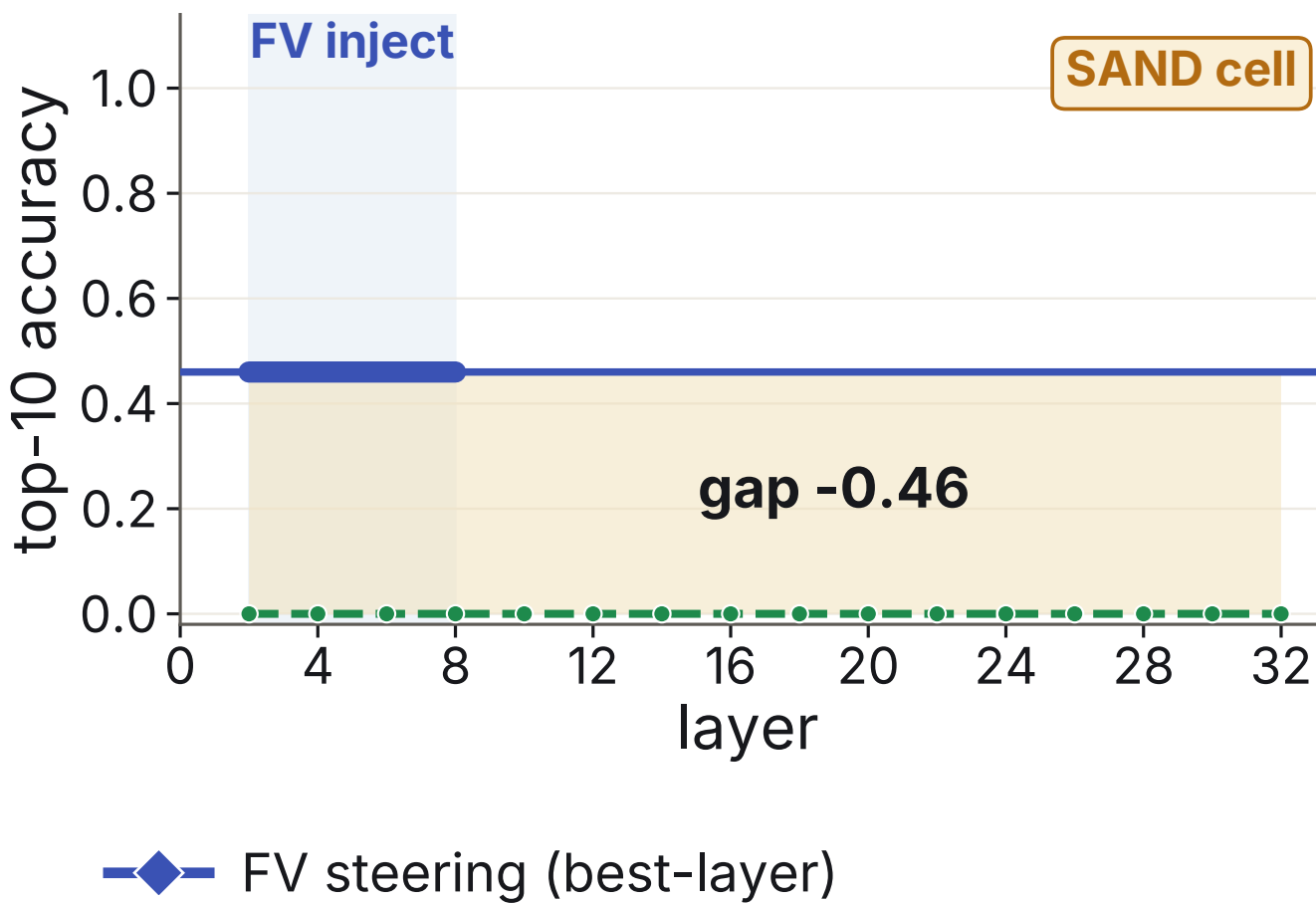
FV steering succeeds where the logit lens fails



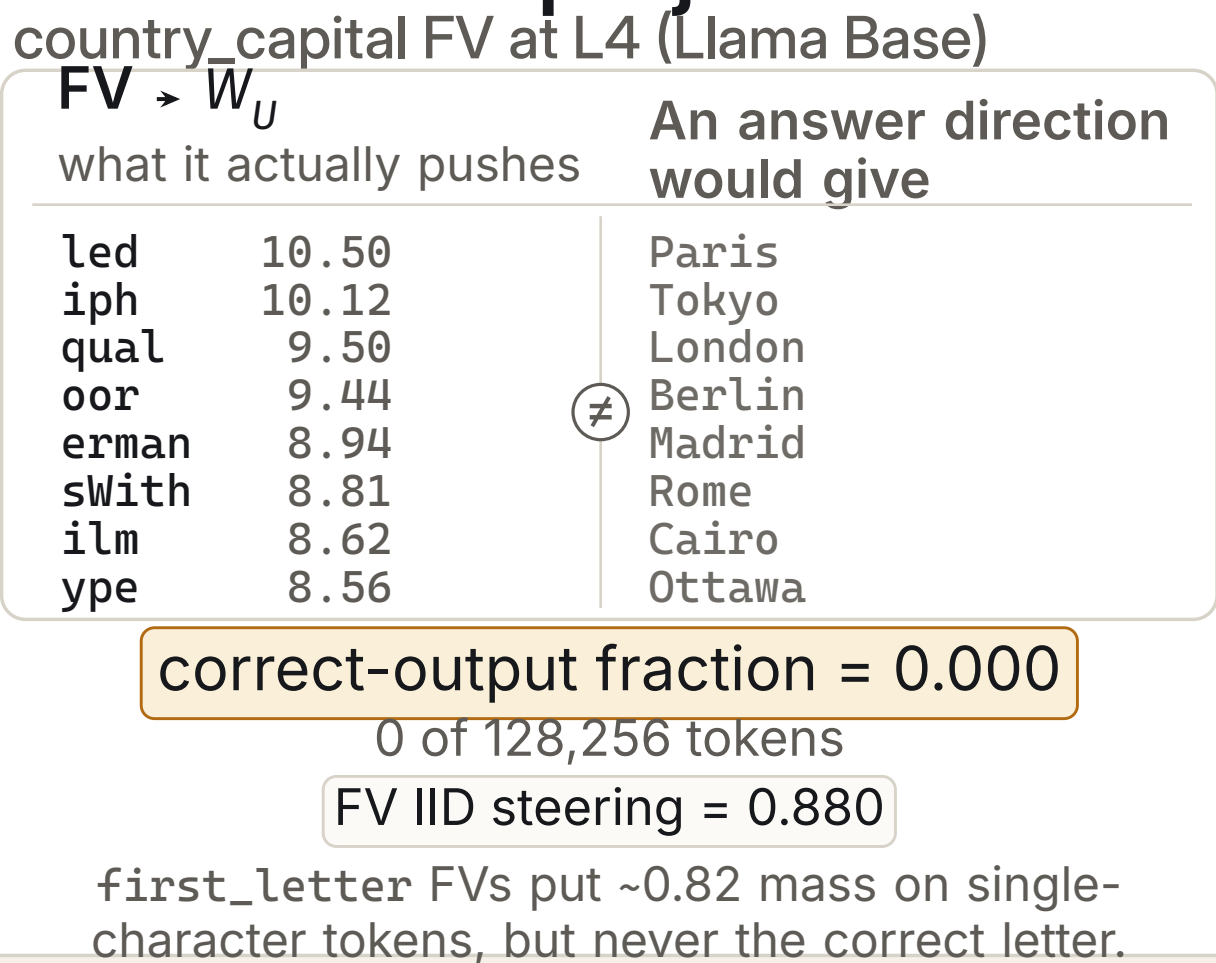
country_capital · Llama Base



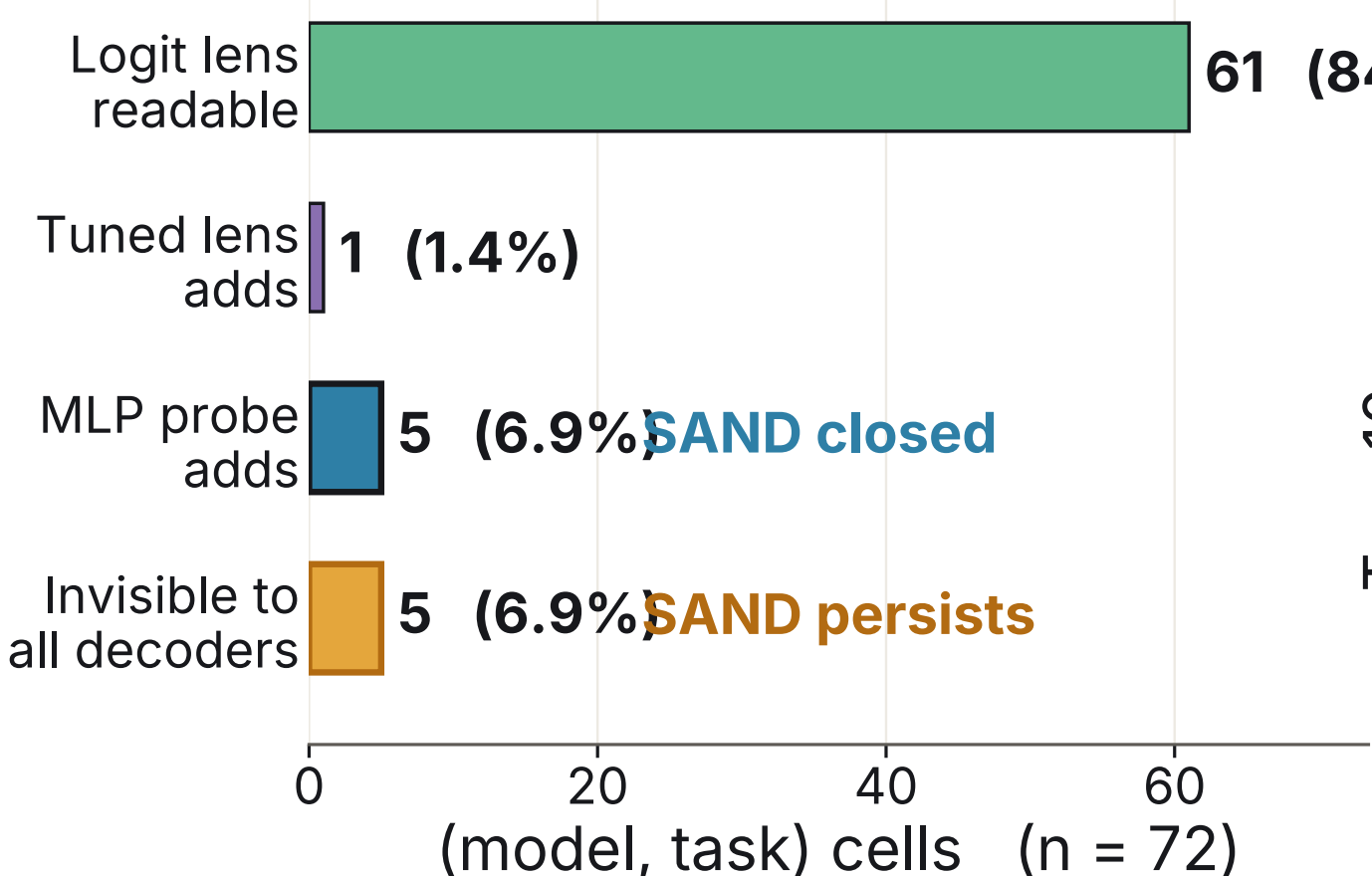
first_letter · Llama Base



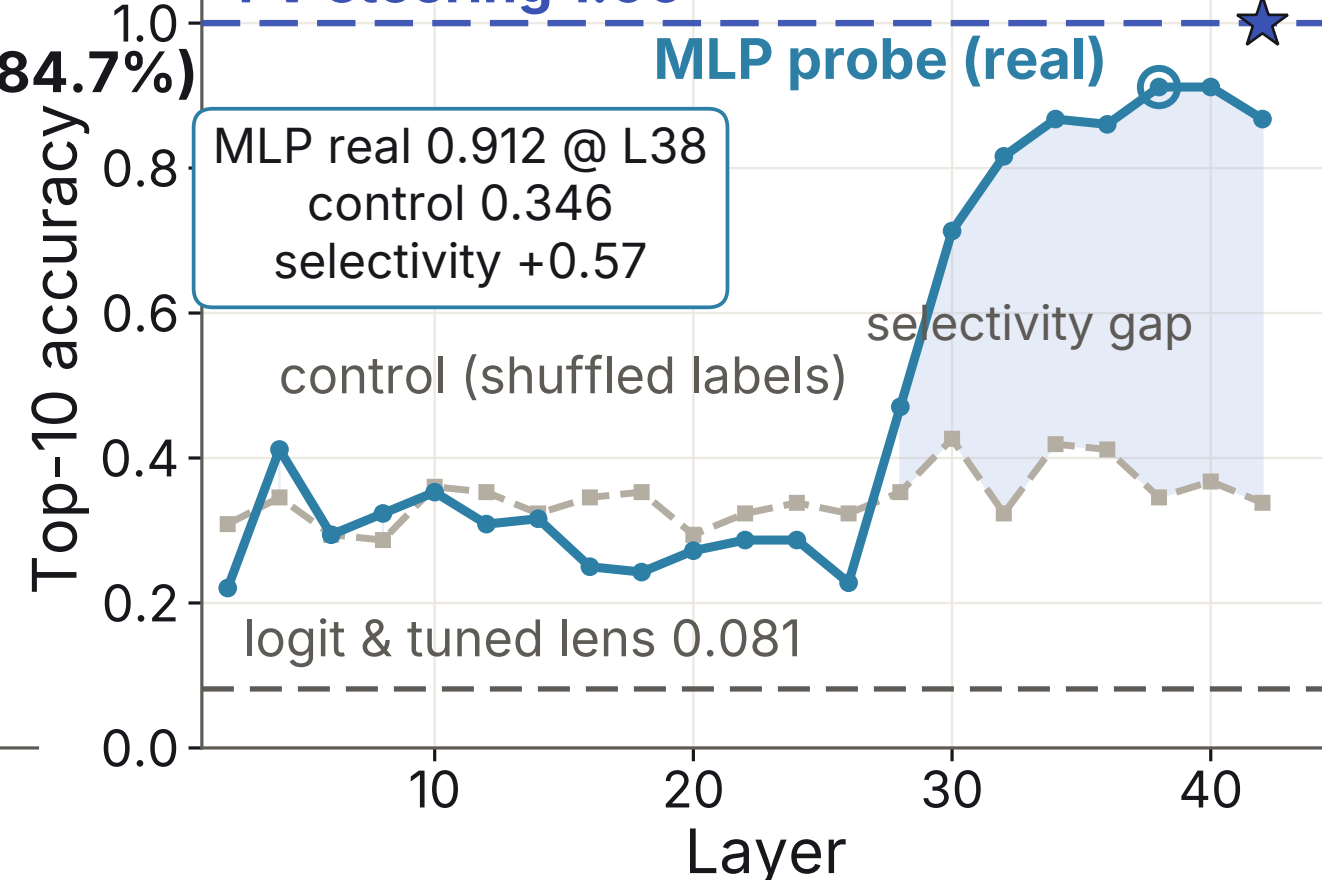
What the FV projects to



A The decoder ladder (72 cells)



B Gemma-2-9B IT × first_letter



FVs encode computational instructions, not answer directions.

"compute the antonym," not "the answer is X."

FIVE CONVERGING LINES

- 1 Incoherent vocab projection (>0.90 steering)
- 2 20+ layer early-steer / late-decode gap
- 3 Template-style invariance ($p > 0.30$)
- 4 Tuned lens: 13 of 14 gaps persist (93%)
- 5 MLP probe: 5 of 10 SAND cells survive

Mistral rewrites internals ($\Delta +0.91$ on object_color); Llama and Gemma steer silently.

Open question. Monitoring asks which concept is active; an FV encodes which procedure to run. Is there an execute-this-step feature to catch?

Scope: single-token / short outputs, 7-9B models, mean-difference extraction only, benign tasks.

Key references

Todd et al. 2024 (function vectors) · Belrose et al. 2023 (tuned lens)
Hewitt and Liang 2019 (control tasks) · Panickssery / Rimsky 2024 (CAA)
nostalgebraist 2020 (logit lens)

Mohammed Suhail B Nadaf · Independent

arxiv.org/abs/2604.02608

github.com/suhailnadaf509/steerable-not-decodable



arXiv



code