Steering off Course: Reliability Challenges in Steering Language Models

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajogopal, Hannaneh Hajishirzi, Sachin Kumar

TL;DR

We evaluate **three** popular **steering methods** on models from different families and find **high variance** in their performance, which indicates **poor generalization**

Background

Steering: modify model **behavior** during **inference** with a specific objective

- Prior work investigates few models, and growing evidence shows brittleness in some steering methods (Sparse Autoencoders & Knowledge Editing)
- We quantify the brittleness of other steering methods, and point out flaws in underlying assumptions

Activation Patching: replace internal activations of a neural network with another vector to modify a specific model behavior

$$h_\ell \leftarrow \alpha h_\ell + \lambda v_t$$

Function Vectors (FV)[3] rely on the *localization hypothesis* (a few attention heads moderate a task)

Task Vectors (TV)[2] directly compress a task into a vector using a few-shot prompt

Logit Lens: the output of any model layer can be projected into the vocabulary space to obtain logits using the unembedding matrix

DoLa[1] computes the relative change in probability at the final layer compared to an earlier or "premature" layer

Experimental Setup

Models

36 decoder-only transformer-based LMs from **14** model families with sizes ranging from **1.5B** to **70B** parameters

FV and TV Data and Eval

11 word-pair ICL tasks, such as generating the antonym of an English word

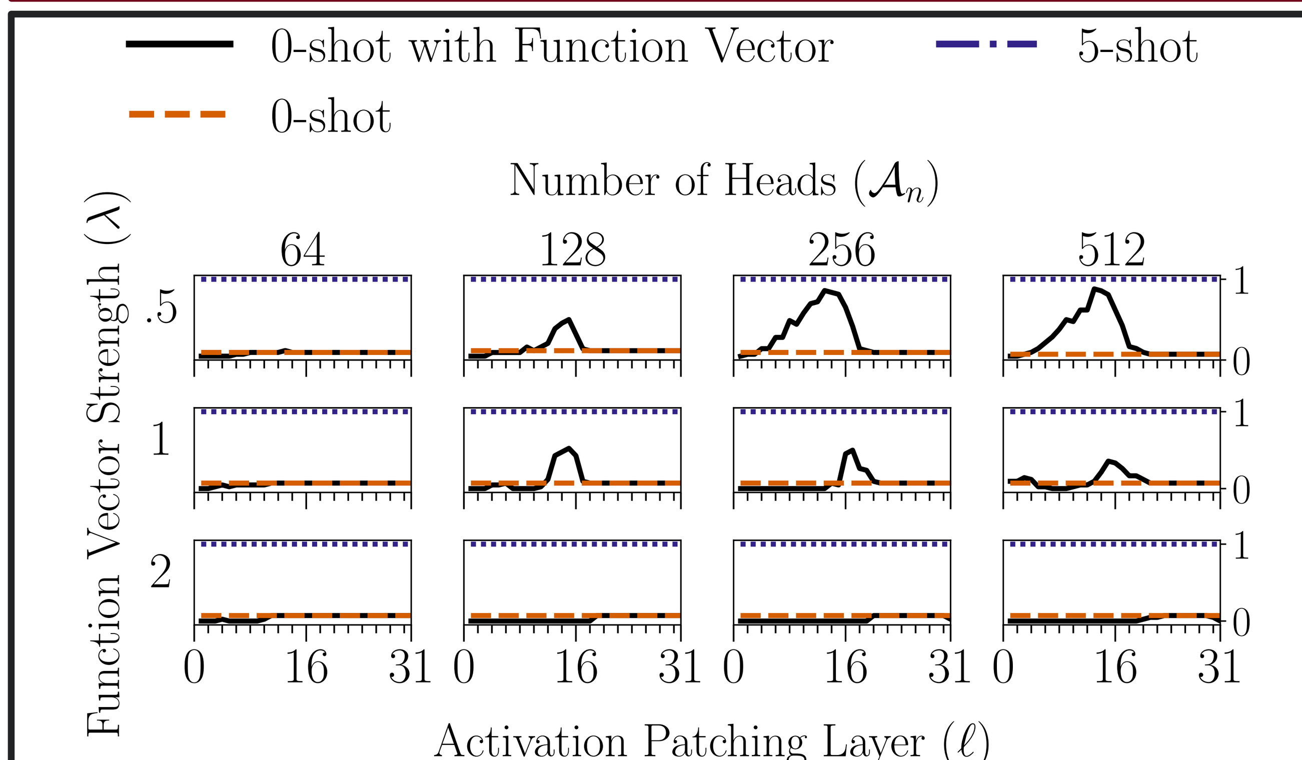
- *Peak* and *average* **performance recovery** (peak is max across all hyperparams, average is mean across layers and max over other params)

DoLa Data and Eval

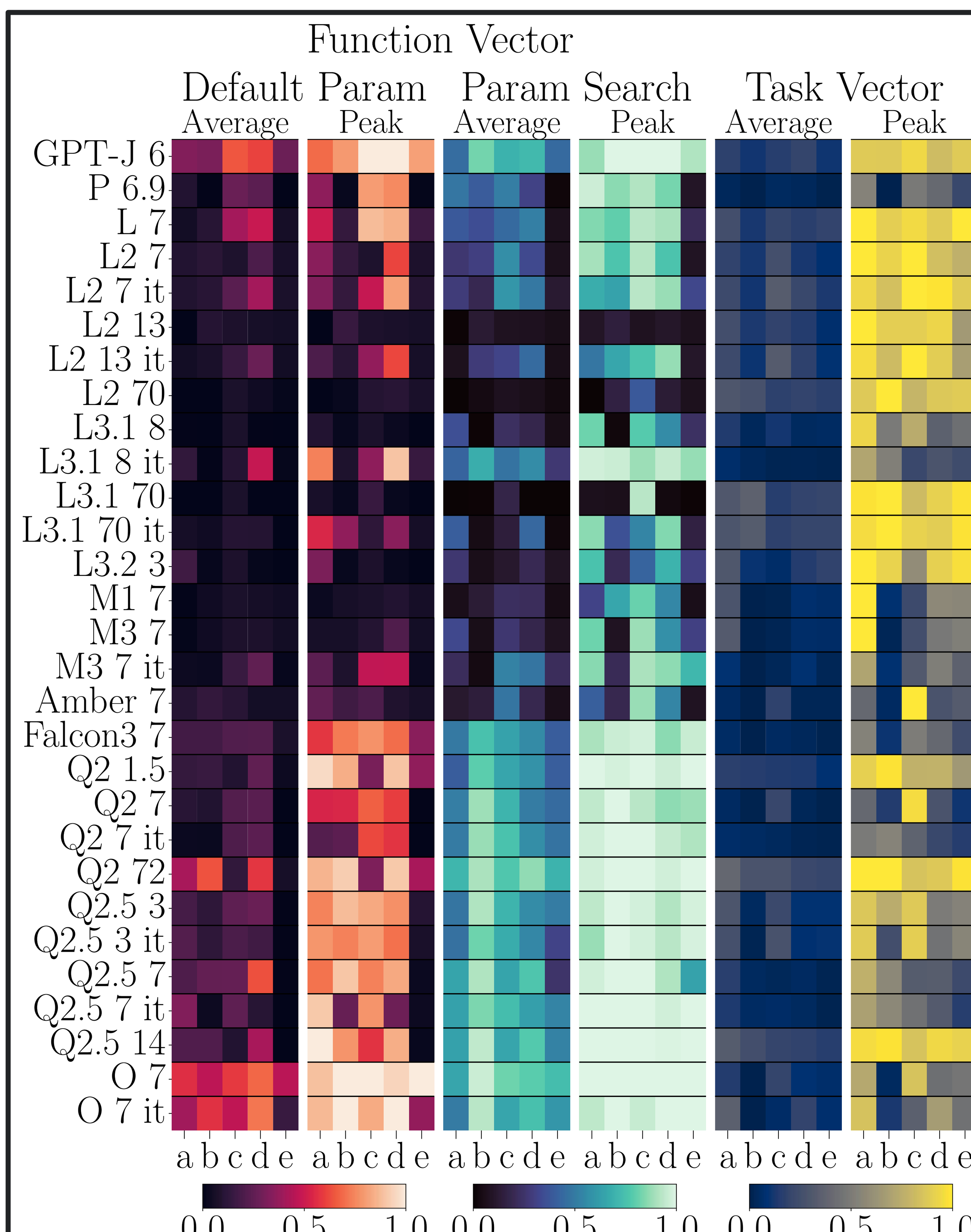
TruthfulQA Multiple Choice

- MC1: one correct answer
- MC2: multiple correct answers
- MC3: evaluate answer ranking

Activation Patching (FV and TV)

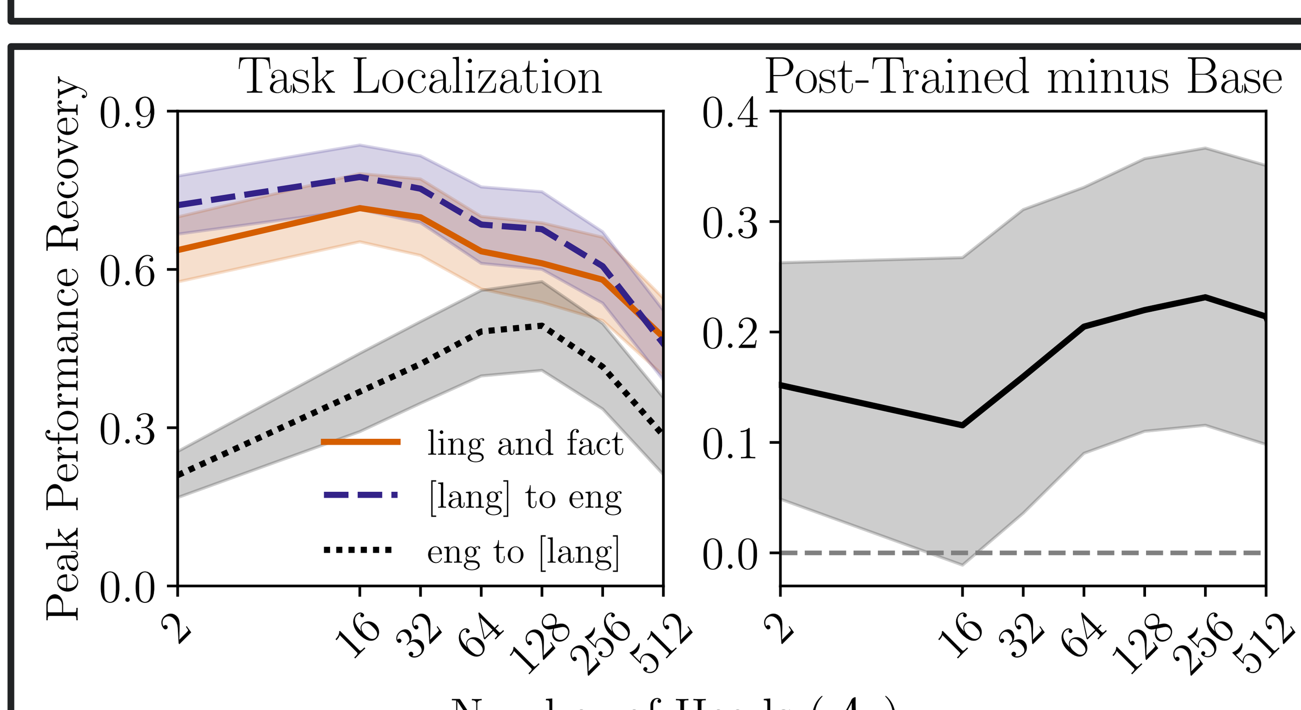


Case study: hyperparameter search reveals **non-localized behavior** in Mistral v0.3 7B on an ICL country-capital task (additional examples in full paper)



Performance recovery across activation patching methods, models, and tasks has **large variability**.

Tasks: a) antonym, b) present-past, c) country-capital, d) [lang] to eng, and e) eng to [lang]

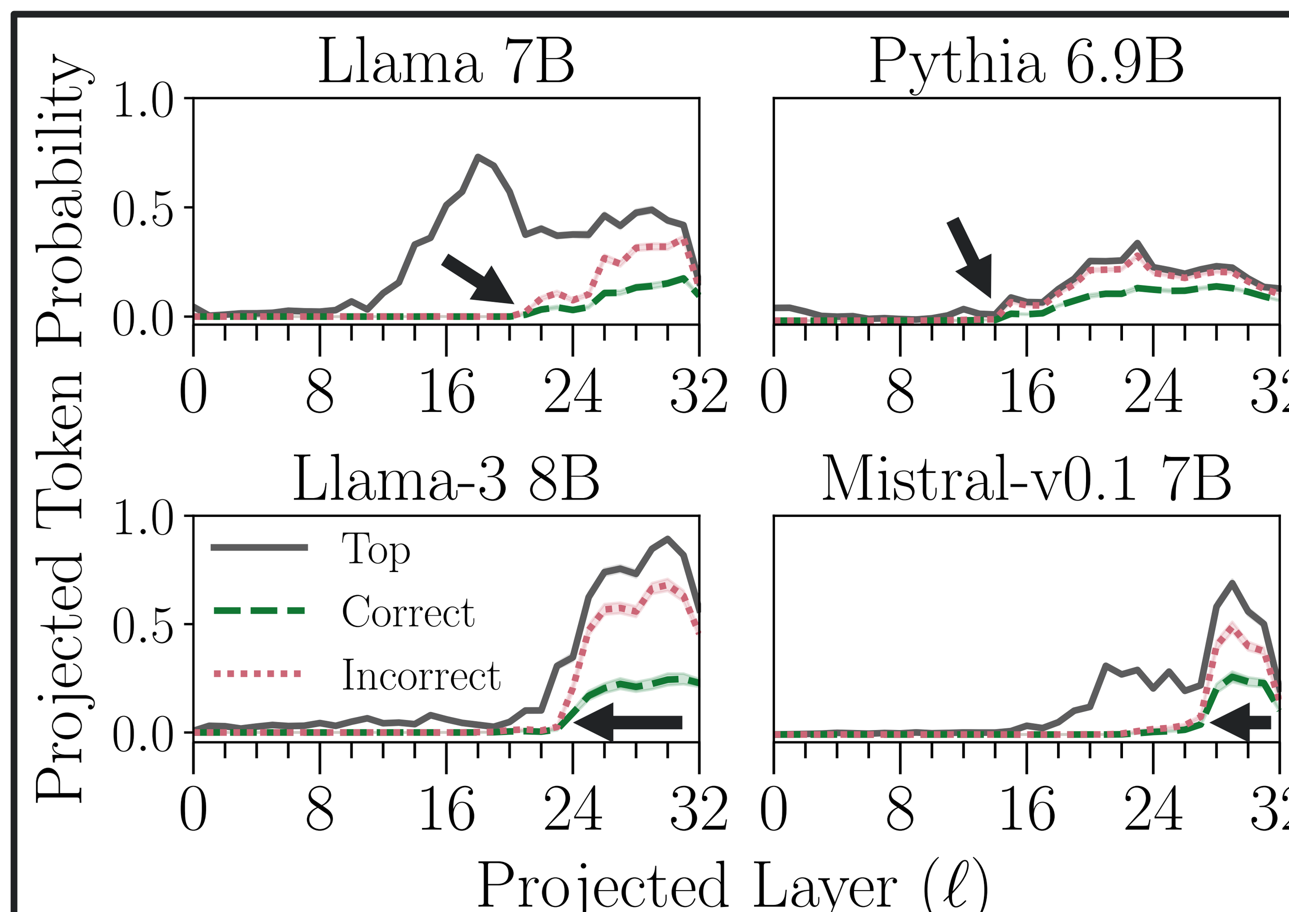


(Left) Some tasks work well with few heads, but **translating** from **English** requires **more heads** (non-localized). (Right) **Post-trained** models **outperform** their **base** versions, especially using **more heads** in the FV

Logit Lens (DoLa)

Model	MC1		MC2		MC3	
	Base	DoLa	Base	DoLa	Base	DoLa
LLama 7B*	0.26	0.32	0.41	0.64	0.19	0.32
Llama 7B	0.26	0.32	0.41	0.52	0.19	0.28
Pythia 6.9B	0.23	0.25	0.37	0.48	0.27	0.23
Mistralv0.1 7B	0.32	0.32	0.48	0.48	0.22	0.24
OLMo 7B	0.25	0.25	0.40	0.40	0.19	0.19
Qwen 2 7B	0.36	0.37	0.49	0.51	0.28	0.30
Llama 2 70B	0.35	0.35	0.52	0.54	0.25	0.25
Llama 3 70B	0.37	0.37	0.58	0.58	0.29	0.30
Qwen 2 72B	0.44	0.40	0.63	0.52	0.33	0.30

Using DoLA for TruthfulQA **does not improve** performance



The **correct** and **incorrect** token probabilities on TruthfulQA start **spiking** at the **same layer**; a **contrast** with early layers is likely to be **uninformative**

Discussion

Underlying **assumptions** upon which **steering methods** are based are **flawed**

Several **hypotheses** may explain these differences (model **pretraining**, **architecture**, **optimization**, and training **data**), but **none** are **conclusive**

Future research in this direction should adopt more **rigorous evaluation** considering a **wide array** of **models** and **tasks**

References

[1] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. **Dola: Decoding by contrasting layers improves factuality in large language models**. In The Twelfth International Conference on Learning Representations.

[2] Roee Hendel, Mor Geva, and Amir Globerson. 2023. **In-context learning creates task vectors.** In The 2023 Conference on Empirical Methods in Natural Language Processing.

[3] Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. **Function vectors in large language models.** In The Twelfth International Conference on Learning Representations.