From Geometry to Behavior

How Instruction Exposure Unlocks Latent Rhetorical Directions in LLMs

2026 · main conference

AUTHOR
Shengjie Xu

AFFILIATION
University of Maryland, College Park

TL;DR

LLMs pick up our habit of **explaining by analogy** from text. We find one **"Analogy Vector"** that dials it up or down. It exists in **every base model, present but sometimes inert**: how much steering it *does* tracks **how much instruction data** the model saw in training.

Same prompt, two vibes

`layer 16 · clean vector · same prompt, two runs`

> Explain credit card cancellation affecting credit score using accurate technical terminology.

G2 Gemma-2-27B-IT $\alpha = 0$ · `no steering`

Cancelling a credit card affects your credit score through two primary mechanisms: (1) credit utilization ratio … (2) credit history length …

G2 Gemma-2-27B-IT $\alpha = +1.0$ · `steered`

Think of your credit report **like** a financial resume … Cancel a card and you're **shrinking the pie**: the same balance now represents a larger fraction …

The prompt asks for technical terms. The vector still makes the model reach for an analogy.

How it works

1 Contrast. 1,000 ELI5 concepts, each asked *literally* and *by analogy*. Residual stream, last prompt token:

$$\mathbf{v} = \mu_{\text{analogy}} - \mu_{\text{literal}}$$

2 Steer. Ask a *literal* question, add the vector at layer ℓ^* (exhaustive sweep, dev set):

$$\mathbf{x}'_\ell = \mathbf{x}_\ell + \alpha\,\mathbf{v}$$

3 Clean. "Imagine" is ∼10% of $\|\mathbf{v}\|$. Push α to 2 and the raw vector does this:

```
Imagine you want to imagine imagine imagine imagine imagine think of it
like imagine imagine …
```

Project out the lexical token directions:

$$\mathbf{v}_{\text{clean}} = \mathbf{v} - \textstyle\sum_t \text{proj}_{\mathbf{d}_t}\mathbf{v}$$

75–92% fewer "Imagine" loops; the norm drops only 1–5%. The direction is mostly *meaning*, not a keyword.

14 models · 7 base/IT pairs · 2B–32B

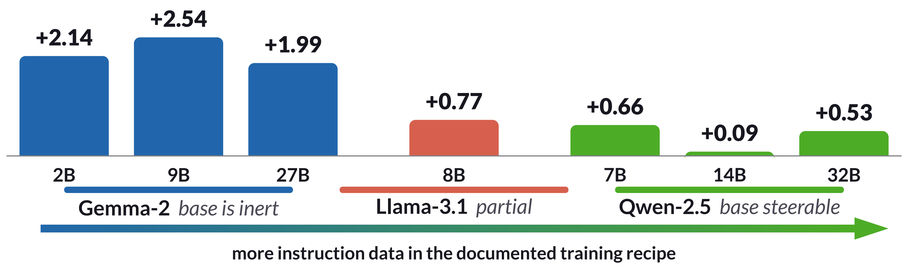
300 held-out prompts · 8 domains

Judge: Qwen-2.5-72B · HM of analogy, fluency, on-task

A gradient of behavioral accessibility

The direction is *present* in every base model. Whether steering it changes behavior scales with instruction exposure.

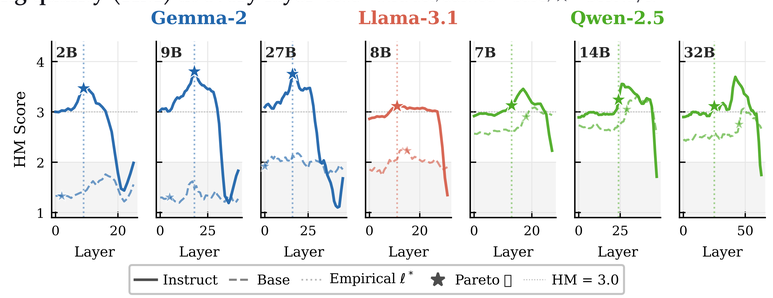
Base → instruct gap ΔHM = HM(instruct) − HM(base), each at its best layer



Geometry agrees max base–IT cosine of the vector · no judge needed

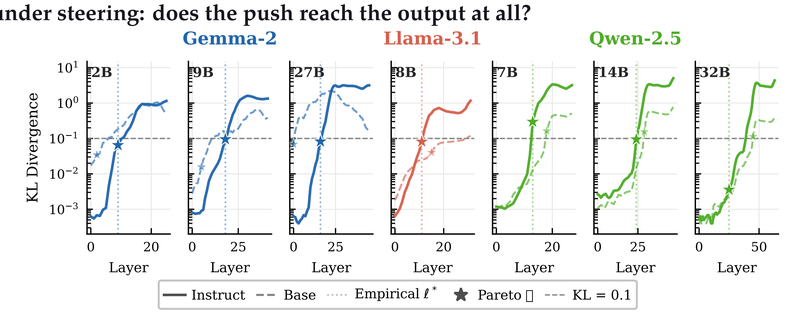
Gemma-2	Llama-3.1	Qwen-2.5
0.37–0.66	**0.30**	**0.94–0.97**
reorganized by IT	*different*	*preserved*

Steering quality (HM) at every layer solid = instruct, dashed = base, ★ = best layer



Gemma-2 bases stay flat at every layer (HM 1.4–1.9): strictly separate pretraining and IT. Qwen-2.5 anneals on SFT-quality data *during* pretraining, so base and instruct nearly overlap.

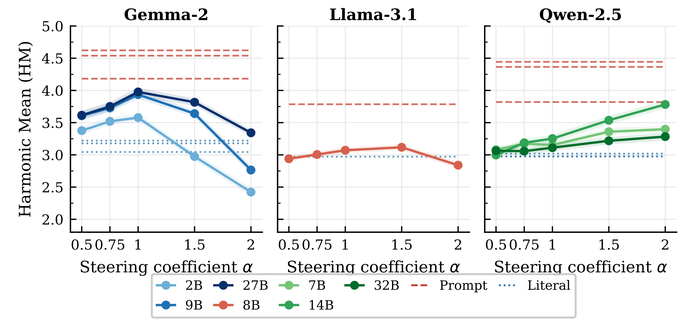
KL under steering: does the push reach the output at all?



Best-layer KL (IT): Gemma-2 0.086–0.100 · Llama-3.1 0.059 · Qwen-2.5 0.004–0.069. Read with HM: Gemma-2 base output *does* shift (KL up to ∼1 in later layers) yet HM stays flat, so the push moves it but doesn't steer it. Qwen-2.5-32B-IT barely shifts (0.004) yet steers: *already accessible.*

Steering works on every IT model

+22%	**7 / 7**	**0.66**
HM, Gemma-2-9B-IT	IT models improve	human–judge ρ
3.22 → 3.94	ΔHM +0.15 to +0.81	15 raters · CI .53–.76



HM vs. steering strength α. Gemma-2 peaks at α=1 and then falls off a cliff; Qwen-2.5 is still climbing at α=2, a much wider steering window. **Hypothesis:** the window's width reflects the local geometry around the direction, i.e. where linear steering leaves the manifold.

Honest note: explicit prompting still wins (HM 3.8–4.6, dashed lines). Humans rank it the same way: prompted > steered > literal. ***Prompting elicits; steering diagnoses.***

Is it really *analogy*?

`controls on Gemma-2-9B-IT`

- ✓ **Not just any push.** A random vector with the same norm leaves HM at ≈3.22; the analogy vector lifts it to 3.94.
- ✓ **Not just "simpler text".** On creative-writing prompts, outputs get longer (49 → 57 tokens) and fluency holds (4.28).
- ✓ **Works in reverse.** Subtracting the vector drops analogy quality from 3.22 to 2.41. A real dial, both ways.

So what?

- **Instruction tuning doesn't create the direction; it makes it usable.** It's a gradient, not an on/off switch.
- **KL under steering** shows whether the push reaches the output at all; read with HM to tell *moved* from *steered*.
- **Next:** SAE / crosscoder features for base vs. IT Gemma-2; activation patching to localize the pathway; formality, hedging, Socratic style.
- **For evals & alignment:** a capability can be *present* in a base model but not yet *usable*. Behavior-only evals of base models may miss it; post-training can unlock it cheaply.

Caveats: LLM judge (human-validated, ρ=0.66); one rhetorical feature; dense transformers only; exposure ordering from published training recipes, not measured.

Seeking Research Internships, Summer 2027

Mechanistic Interpretability · Model Steering · Post-training

Shengjie Xu · sjxu@umd.edu · shengjie-xu.com

Ask me about

- *Why is Qwen base already steerable?*
- *KL moves, HM doesn't: why?*
- *Next: SAE / crosscoder diffing*
- *Is the α-cliff a manifold effect?*

Paper · Examples · Live Demo ▶

`shengjie-xu.com/conferences/2026_COLM_analogy-vector`