From Geometry to Causality: How Instruction Exposure Wires Up Latent Rhetorical Directions in LLMs

Anonymous authors
Paper under double-blind review

Abstract

Language models do not merely memorize facts—they absorb rhetorical strategies from pretraining data, including the human tendency to reason by analogy. We ask: is this absorbed capability geometrically accessible before any instruction tuning, and what determines whether it can be activated at inference time?

We extract an "Analogy Vector" from the residual stream via difference-in-means and show it enables steerable explanation style across **seven instruction-tuned models** (Gemma-2, Llama-3.1, Qwen-2.5; 2B–32B), improving harmonic mean from 3.22 to 3.94 (+22%) for Gemma-2-9B-IT on 299 held-out test prompts.

Our central finding, from **seven matched base-instruct pairs**, is a *gradient of causal wiring*: the analogy direction exists in *every* base model, but its behavioral accessibility scales continuously with instruction exposure during training. Gemma-2 base models (strict pretraining/IT separation) show flat steering landscapes—present but causally inert. Qwen-2.5 base models (instruction data woven into pretraining) are already partially wired before any explicit fine-tuning. This challenges the binary "base vs. instruct" framing: the relevant variable is *how much* instruction exposure a model has received across its full training pipeline. We position activation steering as a complementary interpretability tool and a training-free inference-time lever for amplifying latent rhetorical strategies—a reflection of human cognitive patterns absorbed from text.

1 Introduction

When asked to explain TCP/IP handshakes, a language model can respond technically or reach for an analogy: *"Imagine you want to have a conversation with someone across the room—you'd knock first."* The model knows both styles. Can we control which one it chooses? This question cuts deeper than interface design. Recent work has shown that large language models (LLMs) struggle to *generalize* analogical reasoning beyond surface pattern matching (Lewis & Mitchell, 2024; Stevenson et al., 2025): what models learn is a *reflection* of human rhetorical patterns absorbed from text, not a flexible structural capacity. Our hypothesis is complementary— that analogy-making is encoded as a geometric primitive in the residual stream, present before any instruction tuning, but behaviorally accessible only in proportion to the instruction exposure seen during training.

At the geometric level, Arditi et al. (2024) showed that refusal is mediated by a single linear direction, yet studied only instruction-tuned models. Whether such directions exist *prior* to instruction tuning—and what makes them accessible—was raised as an open question but never empirically closed.[1] We address this gap using analogy-making as a test case: unlike refusal (a binary gate), it is a continuous generative strategy requiring structural mapping across domains—a harder and more illuminating probe of causal wiring.

[1] Arditi et al. (2024) noted informally that base models may learn a related direction that "gets hooked into" by safety fine-tuning, but this was not investigated in the published work.

We extract an "Analogy Vector" via difference-in-means on a 1,000-prompt domain-balanced dataset built from ELI5 (Fan et al., 2019), validate on 299 held-out test prompts across seven instruction-tuned models, and systematically compare **seven matched base-instruct pairs** to trace where geometric existence ends and behavioral accessibility begins. Our work makes three contributions:

1. **Gradient of Causal Wiring.**[2] Across seven matched base-instruct pairs, the analogy direction exists in every base model yet its accessibility scales continuously with instruction exposure. Gemma-2 base models (strict pretraining/IT separation) show flat steering landscapes and KL divergence far below their instruct counterparts—present but causally inert. Qwen-2.5 base models (instruction data in pretraining) are already partially wired, breaking the binary "base vs. instruct" framing. A practical failure mode—lexical entanglement causing mode collapse at high coefficients—is resolved via orthogonalization (Appendix B).
2. **KL Divergence as a Wiring Diagnostic.** KL divergence under steering reveals whether an intervention pathway exists even when output quality is low, providing a model-agnostic probe of pre-existing wiring. Across families, KL at the best layer tracks the wiring gradient without requiring instruction-following capability (Table 2).
3. **Practical Lever for Rhetorical Amplification.** For models where the direction is already accessible (Qwen-2.5 base), activation steering is a training-free, prompt-composable inference-time lever—not a replacement for prompting, but a complementary interpretability tool that reveals how instruction exposure organizes rhetorical knowledge in activation space.

2 Related Work

Analogical Reasoning in LLMs. Analogical reasoning has long been a benchmark for general intelligence (Gentner, 1983), and recent work probes whether LLMs possess it. Lewis & Mitchell (2024) show that LLM performance on analogy tasks degrades sharply under counterfactual perturbations, and Stevenson et al. (2025) demonstrate that LLMs fail to generalize analogy-solving strategies the way children can—suggesting models rely on surface pattern matching rather than structural mapping. Opiełka et al. (2025) further document that internal concept vectors for analogical relations exist yet fail to drive correct outputs. Our work takes a complementary perspective: rather than evaluating whether LLMs *reason* analogically, we ask whether their tendency to *explain* via analogy—a rhetorical strategy absorbed from pretraining text—is geometrically accessible and steerable, and what training conditions make it so.

Linear Representations and Activation Steering. A growing body of work demonstrates that high-level concepts in LLMs can be decoded as linear directions in activation space, including world states (Li et al., 2024a), truthfulness and sentiment (Zou et al., 2025a), and geometry (Marks & Tegmark, 2024). Activation steering leverages these directions by adding vectors to the residual stream to influence behavior (Turner et al., 2024): function vectors trigger specific tasks (Todd et al., 2024), inference-time intervention elicits truthful answers (Li et al., 2024b), and most relevantly, Arditi et al. (2024) showed that refusal is mediated by a single direction extractable via difference-in-means. We extend this line from binary safety features to a continuous rhetorical strategy, addressing the additional challenge of lexical entanglement that arises when steering generative—rather than suppressive—capabilities.

The Superficial Alignment Hypothesis and Causal Inertia. The prevailing view is that instruction tuning teaches format, not knowledge: Zhou et al. (2023) demonstrated that fine-tuning on only 1,000 examples matches GPT-4 quality in 43% of cases, and Xie et al. (2025) showed that base models recover 91% of thinking-model performance when steered with extracted reasoning vectors. However, mechanistic work has documented "causal

[2]"Causal wiring" is shorthand for *behavioral accessibility under linear intervention*, measured as harmonic mean (HM) improvement and KL shift under steering; we make no stronger mechanistic claim.

inertia"—where geometric representations exist but fail to influence output. Kerl (2025) found that sparse autoencoder (SAE)-extracted refusal features transfer geometrically from base to instruct models but lose causal efficacy; Opiełka et al. (2025) showed that concept vectors for analogical relations exist internally yet fail to drive correct outputs. Our work maps the *transition* from inertia to efficacy across seven matched base-instruct pairs with varying degrees of instruction exposure.

Limits of Steering. Zou et al. (2025b) introduced a large-scale benchmark showing that prompting and fine-tuning outperform all representation-based interventions, and that simple linear baselines outperform sparse autoencoders (Cunningham et al., 2023; Bricken et al., 2023). We adopt the AxBench convention of reporting harmonic means and position steering as interpretability evidence rather than a practical alternative to prompting.

3 Methodology

We extract an "Analogy Vector" by computing the mean difference between activations from literal and analogical explanations, following the approach of Arditi et al. (2024) for identifying single directions. For each layer, we calculate:

$$\mathbf{v}_{analogy} = \mu_{analogy} - \mu_{literal} \tag{1}$$

where $\mu_{analogy}$ and $\mu_{literal}$ are mean residual stream activations at the final prompt token, computed over analogy and literal prompt sets respectively. We intervene on the residual stream, which serves as the main information pathway in transformers (Elhage et al., 2021).

During inference, we intervene by adding the vector to the residual stream at layer L:

$$\mathbf{x}'_L = \mathbf{x}_L + \alpha \cdot \mathbf{v}_{analogy} \tag{2}$$

where α is the steering coefficient. This intervention occurs before the layer's attention and multi-layer perceptron (MLP) blocks, allowing the modified activation to propagate through subsequent layers.

At high steering coefficients ($\alpha > 2.0$), we observed catastrophic mode collapse caused by **lexical entanglement**: the "`Imagine`" token embedding contributed $\sim$10% of the vector's norm, overwhelming the semantic signal. We resolve this by projecting the vector onto the subspace orthogonal to the problematic token direction:

$$\mathbf{v}_{clean} = \mathbf{v}_{analogy} - \text{proj}_{\mathbf{d}_{imagine}} \mathbf{v}_{analogy} \tag{3}$$

where $\mathbf{d}_{imagine}$ is the unembedding direction for the "`Imagine`" token. This reduces vector magnitude by only 1–5%, confirming the direction is predominantly semantic rather than lexical. All reported results use $\mathbf{v}_{clean}$; full analysis of the failure mode is in Appendix B.

Dataset. We construct a paired contrastive dataset of 1,000 concept pairs sourced from ELI5 (Fan et al., 2019), a large-scale community QA corpus spanning *r/explainlikeimfive*, *r/askscience*, and *r/AskHistorians*. Concepts are extracted from question titles using a non-target model (Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) via vLLM) to ensure circularity-free curation. We apply a mechanism filter to retain only concepts describing concrete processes or phenomena (e.g., "how bar codes work"), discarding opinion and policy questions unsuitable for structural analogy. A second LLM pass assigns each concept to one of eight knowledge domains (biology, chemistry, CS/technology, earth/environment, economics, engineering, history/society, physics), and quota-based sampling yields a near-uniform distribution (imbalance ratio $\leq 1.5\times$). The dataset is split 600/100/300 (train/dev/test, stratified by domain, zero overlap). For each concept, we construct three literal templates (e.g., "Explain [concept] technically") and three analogy templates (e.g., "Explain [concept] using an analogy") targeting structural mapping. All results are evaluated exclusively on the 299 held-out test concepts.

Models. We evaluate seven matched base-instruct pairs spanning three architecture families and a 2B–32B parameter scale range, all loaded via TransformerLens (Nanda & Bloom, 2022): **Gemma-2-2B** (base/IT), **Gemma-2-9B** (base/IT, primary pair), **Gemma-2-27B** (base/IT, scale-up), **Llama-3.1-8B** (base/Instruct), **Qwen-2.5-7B** (base/Instruct), **Qwen-2.5-14B** (base/Instruct), and **Qwen-2.5-32B** (base/Instruct). See Appendix A for full model details.

Base Model Heterogeneity. A critical methodological consideration is that not all "base" models are created equal. **Gemma-2** maintains strict separation between pre-training and instruction tuning: its base model is a pure next-token predictor with no instruction data in the pre-training corpus (Gemma Team et al., 2024). In contrast, **Qwen-2.5** integrates multi-task instruction data directly into the pre-training process through an 18-trillion-token curriculum that includes late-stage annealing on supervised fine-tuning (SFT)-quality data (Qwen Team et al., 2025). This means the Qwen-2.5 "base" model has already been exposed to substantial instruction-following data during pre-training. This heterogeneity is not a confounder—it is central to our analysis, enabling us to disentangle the effects of geometric existence from causal efficacy along a gradient of instruction exposure.

Evaluation. We employ a three-metric LLM-as-a-Judge setup (Zheng et al., 2023) using Qwen-2.5-72B-Instruct (Yang et al., 2024) as judge, scoring each response independently on three 1–5 dimensions: **Analogy Quality** (does the response use a clear, developed structural analogy?), **Fluency/Coherence** (is the response grammatical and well-structured?), and **Instruction-Following** (does the response address the original question?). The overall score is the harmonic mean of the three dimensions, following the AxBench convention (Zou et al., 2025b). We compare three conditions: (1) **Unsteered Literal**—the model receives a literal prompt with no steering ($\alpha = 0$); (2) **Prompting Baseline**—the model receives an explicit analogy request ("Explain [concept] using an analogy") with no steering; (3) **Steered Clean**—the model receives a literal prompt with the orthogonalized analogy vector applied at the optimal layer. With $N = 299$ test prompts, the standard error of the mean (SEM) is $\leq$0.06 for all reported scores.

4 Results

4.1 Layer Selection

Following Arditi et al. (2024), we treat layer selection as a hyperparameter and select the empirically best layer via exhaustive sweep. For each model and layer $l \in [1, L]$, we compute $\mathbf{v}^{(l)}_{analogy}$ and evaluate steering effectiveness on the 100-prompt dev set, selecting $l^* = \arg\max_l \mathrm{HM}(\mathrm{AQ}, \mathrm{FL}, \mathrm{IF})$. All reported results use these empirically selected layers; full per-model details are in Appendix A.2.

As a complementary analysis, we examine the inter-layer cosine similarity of $\mathbf{v}_{analogy}$, which reveals a transition from a "computation" phase (rapidly changing direction) to a "propagation" phase (stable direction). The empirically best layers consistently fall near this transition point—effective intervention occurs where the analogy concept has been computed but output has not yet been committed to specific lexical choices. The empirical sweep, not the stability heuristic, determines all reported results. Figure 2 shows the full HM landscape across all 14 models; the monotonically narrowing base–instruct gap from Gemma-2 to Qwen-2.5 constitutes our central finding (§4.3).

4.2 Steering Effectiveness and Multi-Model Generalization

Table 1 reports per-metric results for all seven IT models on the 299-prompt held-out test set.

Steering selectively lifts Analogy Quality. Explicit prompting remains the strongest method (HM ∼3.8–4.6), consistent with AxBench (Zou et al., 2025b). Across all models, steering raises AQ most—the intended target—while FL and IF remain stable or improve

modestly, confirming the vector is semantically specific rather than a generic style perturbation. The ΔHM gradient mirrors the wiring gradient (§4.3): Gemma-2 peaks at α=1.0 (ΔHM = +0.53–+0.80) and collapses beyond; Llama-3.1 shows a modest gain (ΔHM = +0.15); Qwen-2.5 responds *monotonically* through α=2.0, with HM still rising at the largest coefficient tested—suggesting the analogy direction occupies a flat representational neighborhood where larger interventions remain coherent (see Fig. 1 and §4.3).

Table 1: Per-metric results on 299 held-out test prompts (clean vector, empirically best layer per model). **Prompt** = analogy-eliciting prompt, no steering; **Literal** = neutral prompt, no steering; **Steered** = neutral prompt + analogy vector at best α. **ΔHM** = Steered − Literal HM; SEM $\leq$ 0.06.

Model	Condition	AQ	FL	IF	HM	ΔHM
Gemma-2 (strict base/IT separation)						
Gemma-2-2B-IT	Prompt	3.96	4.35	4.30	4.18	
	Literal	2.88	3.32	3.29	3.04	
	Steered (α=1.0)	**3.30**	**3.84**	**3.70**	**3.58**	**+0.53**
Gemma-2-9B-IT	Prompt	4.38	4.66	4.64	4.54	
	Literal	2.74	3.96	3.93	3.22	
	Steered (α=1.0)	**3.53**	**4.31**	**4.23**	**3.94**	**+0.71**
Gemma-2-27B-IT	Prompt	4.48	4.72	4.71	4.62	
	Literal	2.62	4.08	4.05	3.18	
	Steered (α=1.0)	**3.47**	**4.55**	**4.48**	**3.98**	**+0.80**
Llama-3.1 (intermediate wiring)						
Llama-3.1-8B-IT	Prompt	3.58	3.99	3.88	3.79	
	Literal	2.79	3.32	3.24	2.97	
	Steered (α=1.5)	**3.03**	**3.25**	**3.15**	**3.12**	**+0.15**
Qwen-2.5 (instruction data in pre-training)						
Qwen-2.5-7B-IT	Prompt	3.69	3.91	3.90	3.82	
	Literal	2.85	3.28	3.28	3.02	
	Steered (α=2.0)	**3.21**	**3.61**	**3.50**	**3.40**	**+0.38**
Qwen-2.5-14B-IT	Prompt	4.24	4.45	4.43	4.37	
	Literal	2.66	3.55	3.53	2.98	
	Steered (α=2.0)	**3.46**	**4.19**	**3.97**	**3.78**	**+0.81**
Qwen-2.5-32B-IT	Prompt	4.36	4.49	4.51	4.44	
	Literal	2.62	3.62	3.61	2.98	
	Steered (α=2.0)	**2.98**	**3.74**	**3.73**	**3.28**	**+0.31**

Our evaluation measures whether we can *control* analogy usage, not whether steered analogies improve comprehension. Human evaluation of pedagogical effectiveness is complementary future work. We show the vector successfully induces the target style, though steered outputs may overuse analogy markers (see Appendix D).

4.3 Gradient of Causal Wiring Across Base-Instruct Pairs

Our central finding emerges from systematically comparing seven matched base-instruct pairs. For each pair, we performed the same exhaustive layer sweep on *both* the base and instruct model, measuring the harmonic mean (HM) of the three evaluation metrics at the empirically best layer and steering coefficient. Table 2 summarizes the results.

Gemma-2: Causally Inert Base Models. Across all three Gemma-2 scales (2B, 9B, 27B), the base model's HM landscape is flat and low (1.36–1.88), with no layer producing meaningful steering. The analogy direction exists geometrically in the residual stream—the difference-in-means vector can be extracted—but steering it does not influence output. The base model produces raw next-token completions, with instruction-following scores ≈1.2/5 under all conditions. In contrast, the instruction-tuned counterparts show clear peaked landscapes

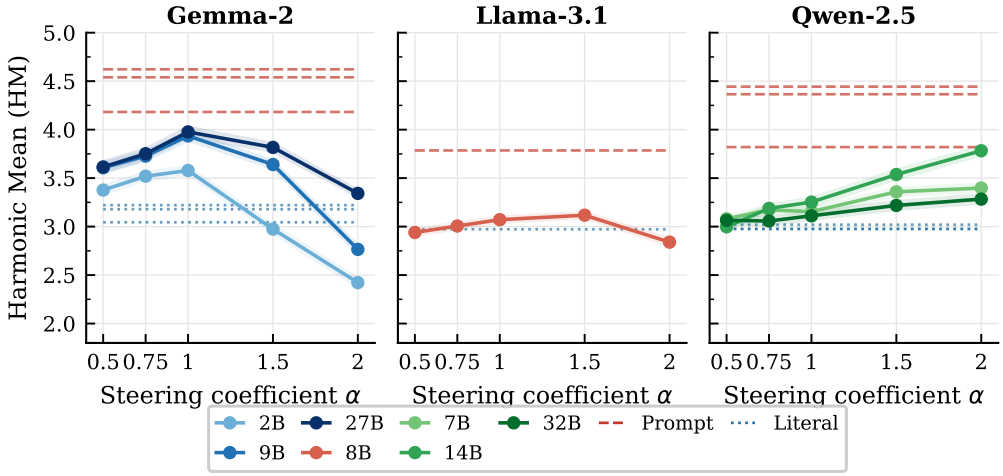

Figure 1: HM vs. steering coefficient α (299 test prompts, clean vector). Gemma-2 peaks sharply at α=1.0 and collapses at $\alpha \geq 1.5$; Qwen-2.5 responds monotonically through α=2.0, with all three metrics (AQ, FL, IF) still improving—consistent with a flatter representational neighbourhood around the analogy direction. The α-response shape is a third observable corroborating the wiring gradient of Figs. 2–3.

Table 2: Gradient of causal wiring across seven matched base-instruct pairs. $\mathbf{HM}_{base}$ and $\mathbf{HM}_{IT}$ are the harmonic mean at the empirically best layer for each model. $\Delta\mathbf{HM} = \mathrm{HM}_{IT} - \mathrm{HM}_{base}$. $\mathbf{KL}_{IT}$ is the average KL divergence under steering at the IT model's best layer.

Pair	l^*_{base}	l^*_{IT}	$\mathbf{HM}_{base}$	$\mathbf{HM}_{IT}$	$\Delta\mathbf{HM}$	$\mathbf{KL}_{IT}$
Gemma-2-2B	2	9	1.38	3.52	+2.14	0.100
Gemma-2-9B	5	18	1.36	3.90	+2.54	0.094
Gemma-2-27B	0	16	1.88	3.87	+1.99	0.086
Llama-3.1-8B	15	11	2.35	3.12	+0.77	0.059
Qwen-2.5-7B	18	18	2.57	3.23	+0.66	0.033
Qwen-2.5-14B	29	24	3.05	3.14	+0.09	0.069
Qwen-2.5-32B	45	25	2.67	3.20	+0.53	0.004

with HM_{IT} ranging from 3.52 to 3.90. The gap (ΔHM = 1.99–2.54) is the largest of any family, indicating that instruction tuning creates the entire causal pathway from geometry to behavior.

Llama-3.1: Intermediate Wiring. The Llama-3.1-8B base model achieves HM_{base} = 2.35—substantially above the Gemma-2 base models—suggesting partial causal wiring. However, instruction tuning still provides a meaningful improvement (ΔHM = +0.77), placing this family at an intermediate position on the wiring gradient.

Qwen-2.5: Pre-Wired from Pre-Training. Most strikingly, the Qwen-2.5-14B *base* model achieves HM_{base} = 3.05—nearly matching its instruct counterpart (HM_{IT} = 3.14, ΔHM = +0.09). The analogy direction is already behaviorally accessible *before* any explicit instruction tuning. This is consistent with the Qwen-2.5 pre-training methodology, which integrates instruction-following data directly into the 18-trillion-token pre-training curriculum (Qwen Team et al., 2025). The Qwen-2.5-32B base model shows a slightly larger gap (ΔHM = +0.53), suggesting that the degree of accessibility varies with scale. Crucially, Qwen-2.5 base models *can* follow instructions, yet the degree to which steering works still tracks pre-training instruction exposure rather than IT presence—demonstrating that the gradient reflects the accessibility of the analogy direction itself, not general instruction-following capability.

KL Divergence as a Wiring Diagnostic. The KL divergence under steering provides a complementary diagnostic. For Gemma-2-9B-IT, the steering intervention at the best layer produces KL $\approx$ 0.094—a meaningful distribution shift indicating the vector is causally connected to output. For the Gemma-2-9B *base* model, the KL values across all layers remain $\leq$0.018—the vector fails to shift the output distribution. For Qwen-2.5-32B-Instruct, KL at the best layer is only 0.004, consistent with the vector being already wired in: the "steering" merely reinforces an existing pathway rather than creating a new one. This suggests KL divergence under steering may serve as a practical diagnostic for detecting pre-existing causal wiring in base models.

Three-Tier Synthesis. Across families, we identify three tiers of causal wiring: (1) *True base models* (Gemma-2) where the direction is geometrically present but entirely causally inert (ΔHM $>$ 2.0, flat HM landscape); (2) *Pre-wired base models* (Qwen-2.5) where instruction-contaminated pre-training creates partial causal efficacy (ΔHM $<$ 0.7, overlapping HM curves); and (3) *Fully instruction-tuned models* where the direction is fully wired to output generation. Llama-3.1 occupies the boundary between tiers 1 and 2—its base model can partially follow instructions and shows intermediate ΔHM, consistent with a pre-training curriculum that falls between Gemma-2's strict separation and Qwen-2.5's explicit instruction integration.

These results establish a **gradient of causal wiring** across model families: instruction tuning is not a binary switch that creates new capabilities, but a process that scales the causal efficacy of latent directions—a process whose effects depend critically on what the pre-training curriculum has already organized.

The steering coefficient as a wiring probe. Figure 1 provides a fourth observable consistent with the three-tier account. Gemma-2 models peak sharply at α=1.0 and collapse at $\alpha \geq$1.5: all three metrics (AQ, FL, IF) degrade together, indicating the representation overshoots into a region that the base distribution cannot coherently execute. Qwen-2.5 models respond monotonically through α=2.0, with all three metrics continuing to improve. We interpret this as a *headroom* effect: tight causal wiring (Gemma-2, created entirely by instruction tuning) places the analogy direction in a narrow representational corridor—powerful within bounds, fragile outside them. Diffuse wiring (Qwen-2.5, pre-trained on analogy-bearing text) embeds the direction in a flatter neighborhood where larger interventions remain coherent. The width of the effective steering window is thus an empirical proxy for the geometry of the analogy subspace, corroborating the HM landscape (Fig. 2) and KL diagnostic (Fig. 3) from an independent angle.

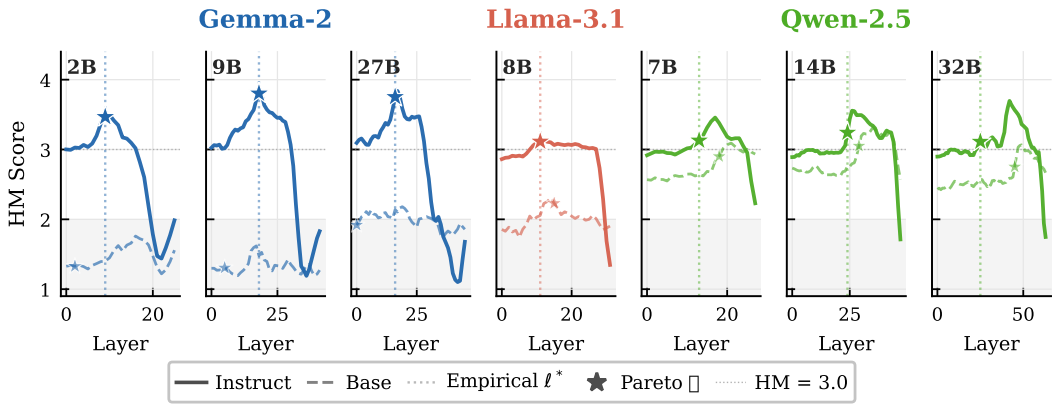

Figure 2: **Steering quality (HM) across layers reveals a gradient of behavioral accessibility.** Dev-set harmonic mean for all seven base-instruct pairs. Reading left to right across families, the base–instruct gap narrows monotonically: Gemma-2 base models are flat (HM $\approx$ 1.4–1.9); Llama-3.1 base reaches a moderate peak (HM=2.35); Qwen-2.5 base and instruct curves nearly overlap (ΔHM$<$0.7).

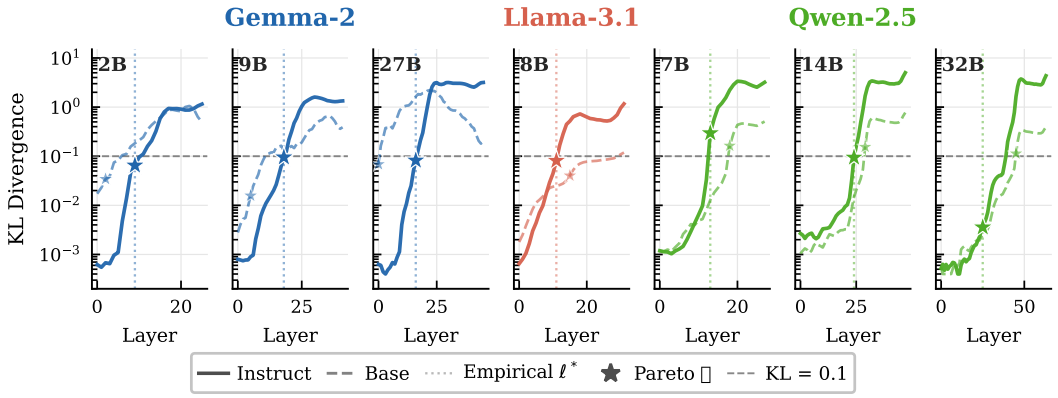

Figure 3: **KL divergence under steering mirrors the wiring gradient**. Gemma-2 instruct models peak at KL $\approx$ 0.086–0.100 while their base counterparts remain $\leq$0.018 across all layers. Qwen-2.5 base and instruct profiles closely overlap (KL $\approx$ 0.003–0.07). Llama-3.1 shows intermediate separation (KL_{IT}=0.059).

4.4 Control Experiments

We validated the specificity and causal nature of the steering vector through three rigorous controls, all run on Gemma-2-9B-IT.

Random Baseline (Specificity). Injecting random vectors of identical norm across $\alpha \in [0, 2]$ left HM unchanged at $\approx$3.22, whereas the analogy vector raised HM to 3.94 at $\alpha = 1.0$ with a corresponding AQ increase from 2.74 to 3.53. The effect is intrinsic to the computed direction, not a generic perturbation artifact.

Simplicity Confounder (Semantics). To test whether the vector merely degrades text to a simpler register, we applied it to creative writing prompts that do not invite analogies. At $\alpha = 1.0$, steered outputs grew longer (49 to 57 tokens) and increased analogy usage without losing fluency (FL = 4.28). The vector actively structures output around analogical framing rather than collapsing register or complexity.

Negative Steering (Bidirectionality). Subtracting the vector ($\alpha \in \{-0.5, -1.0\}$) from prompts that naturally elicit analogies reduced AQ from 3.22 to 2.41. Positive coefficients restore and amplify the behavior symmetrically, confirming the direction acts as a continuous bidirectional dial rather than a one-sided trigger.

5 Discussion and Conclusion

What instruction tuning actually does. Our central finding reframes a standard question in alignment research. Across seven matched base-instruct pairs, the analogy direction exists geometrically in *every* base model—but instruction tuning is not a binary switch that creates it. It is a continuous process that calibrates the output head to *use* it. Gemma-2 base models illustrate the extreme: high KL under steering (the intervention reaches the output distribution) yet near-zero HM improvement (the shift produces noise, not analogies). The Qwen-2.5 base models resolve the instruction-following confound: they can follow instructions, yet wiring still tracks pretraining curriculum rather than IT presence. Together, these two poles bracket a gradient—from Gemma-2 (inert, ΔHM $\approx$ 2.0–2.5) through Llama-3.1 (partial, ΔHM = 0.77) to Qwen-2.5 (pre-wired, ΔHM $\approx$ 0.09–0.66)—that directly supports the Superficial Alignment Hypothesis (Zhou et al., 2023) at the level of individual feature directions and extends the causal inertia account of Kerl (2025); Opiełka et al. (2025).

Prompting vs. steering. Consistent with AxBench (Zou et al., 2025b), explicit prompting ($\sim$4.54/5 for Gemma-2-9B-IT) outperforms steering (3.94/5). This is not a weakness—they serve different purposes. Prompting elicits; steering *diagnoses*. For models where

the direction is already accessible (Qwen-2.5 base), steering also becomes a training-free, prompt-composable lever for amplifying explanation style at inference time—useful in settings where analogy aids conceptual exploration (Stevenson et al., 2025). What steering amplifies, however, is not general intelligence but a *latent reflection of human rhetorical patterns*—consistent with documented limits on LLM analogical generalization (Lewis & Mitchell, 2024).

Limitations. Our evaluation relies on LLM-as-a-Judge, which may reward surface analogy markers over reasoning depth; human evaluation of pedagogical utility remains future work. The gradient is observed for one rhetorical feature—whether it generalizes to formality, hedging, or Socratic questioning is open. Pure base models (Gemma-2) do not follow instructions, making it impossible to fully disentangle analogy-specific inertia from general incapability; the Qwen-2.5 base results provide the strongest mitigation. Finally, we evaluate dense transformers only; mixture-of-experts routing complicates residual-stream interpretation and is left to future work.

Future directions. Testing whether analogy, formalization, and simplification vectors are orthogonal or composable would transform this work from a single case study into a general framework for rhetorical style control. Applying the same methodology to the refusal direction would enable direct cross-comparison with Arditi et al. (2024). Measuring the local curvature of the analogy direction's neighborhood at each model's best layer—and testing whether it predicts the α-collapse point— would connect the behavioral gradient to mechanistic geometry, turning the α-response shape from an empirical observation into a principled diagnostic. Finally, KL divergence under steering may serve as a lightweight tool for auditing which capabilities are already wired in a base model before any post-training alignment—a practical complement to pretraining curriculum evaluation.

LLM Usage Disclosure

Large language models (LLMs) were used in three capacities during this project: (1) **code generation**—generating initial scaffolding and test code for evaluation scripts, which was subsequently reviewed, debugged, and validated by the authors; (2) **research ideation**—exploring and stress-testing research directions and experimental designs, with all decisions made by the authors; and (3) **writing assistance**—suggesting revisions to prose and structure, with all final wording approved by the authors. In all cases, human authors exercised final judgment, verified all claims against experimental results, and take full responsibility for the content of this paper.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack

Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567. Association for Computational Linguistics, 2019.

Gemma Team et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Tilman Kerl. Evaluation of sparse autoencoder-based refusal features in LLMs, 2025.

Martha Lewis and Melanie Mitchell. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models, 2024. URL https://arxiv.org/abs/2402.08955.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task, 2024a. URL https://arxiv.org/abs/2210.13382.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2024b. URL https://arxiv.org/abs/2306.03341.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

Mikołaj Opiełka et al. Analogical reasoning inside large language models: Concept vectors and the limits of abstraction, 2025.

Qwen Team, An Yang, Baosong Yang, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Claire E. Stevenson, Alexandra Pafford, Han L. J. van der Maas, and Melanie Mitchell. Can large language models generalize analogy solving like children can?, 2025. URL https://arxiv.org/abs/2411.02348.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models, 2024. URL https://arxiv.org/abs/2310.15213.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Ang Xie et al. Base models know how to reason, thinking models learn when, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. AxBench: Steering LLMs? even simple baselines outperform activation engineering. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025b.

A Experimental Details

A.1 Models and Dataset

We conduct experiments on seven matched base-instruct pairs (14 models total), all loaded via TransformerLens (Nanda & Bloom, 2022):

Dataset. We use the v5 balanced-domain dataset constructed from ELI5 (Fan et al., 2019). Key statistics: 1000 concept pairs (600 train / 100 dev / 299 test), 8 fine-grained domains (biology, chemistry, CS/technology, earth/environment, economics/finance, engineering/mechanics, history/society, physics), domain imbalance ratio $\leq 1.5\times$. Concepts are extracted from ELI5 question titles using Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) as a non-target curation model (circularity-free). A mechanism filter retains only concepts describing concrete processes or phenomena. A second LLM pass assigns fine-grained domain labels. Quota-based balanced sampling ensures near-uniform domain distribution. The ground-truth validation set (200 examples with original ELI5 answers) is drawn from the unused ELI5 pool with zero overlap to any split.

Each concept is paired with three literal templates and three analogy templates. For vector extraction (training), one pair per concept is used. For evaluation, all 299 test concepts are evaluated under three conditions: unsteered literal, prompting baseline, and steered clean.

Analogy templates explicitly reference structural mapping (e.g., "Explain [concept] by drawing an analogy to something from everyday life") rather than mere simplification, consistent with our definition of analogical explanation as *structural correspondence between source and target domains*.

Table 3: Model lineup: seven matched base-instruct pairs across three architecture families. All models are dense transformers. **Pre-Training**: NTP = next-token prediction only (no instruction data).

Model	Params	Layers	Hidden	Pre-Training	Role
Gemma-2 family (strict base/IT separation) (Gemma Team et al., 2024)					
Gemma-2-2B (base)	2B	26	2304	Pure NTP	Base
Gemma-2-2B-IT	2B	26	2304	Pure NTP	Instruct
Gemma-2-9B (base)	9B	42	3584	Pure NTP	Base (primary)
Gemma-2-9B-IT	9B	42	3584	Pure NTP	Instruct (primary)
Gemma-2-27B (base)	27B	46	4608	Pure NTP	Base (scale-up)
Gemma-2-27B-IT	27B	46	4608	Pure NTP	Instruct (scale-up)
Llama-3.1 family					
Llama-3.1-8B (base)	8B	32	4096	Pure NTP	Base
Llama-3.1-8B-Instruct	8B	32	4096	Pure NTP	Instruct
Qwen-2.5 family (instruction data in pre-training) (Qwen Team et al., 2025)					
Qwen-2.5-7B (base)	7B	28	3584	+SFT anneal	Base
Qwen-2.5-7B-Instruct	7B	28	3584	+SFT anneal	Instruct
Qwen-2.5-14B (base)	14B	48	5120	+SFT anneal	Base
Qwen-2.5-14B-Instruct	14B	48	5120	+SFT anneal	Instruct
Qwen-2.5-32B (base)	32B	64	5120	+SFT anneal	Base
Qwen-2.5-32B-Instruct	32B	64	5120	+SFT anneal	Instruct

A.2 Layer Selection

Following Arditi et al. (2024), we select the steering layer via exhaustive sweep over all layers, choosing $l^* = \arg\max_l \mathrm{HM}(\mathrm{AQ}, \mathrm{FL}, \mathrm{IF})$ on the 100-prompt dev set. Table 4 reports the selected layers for all 14 models.

As a complementary analysis, we examine inter-layer cosine similarity of $\mathbf{v}_{analogy}$, which reveals a transition from a "computation" phase to a "propagation" phase. The empirically best layers consistently fall near this transition. However, the empirical sweep—not the stability heuristic—determines all reported results.

Table 4: Empirically selected layers and stability transition layers for all 14 models.

Model	Empirical l^*	Stability Transition	Best dev HM
Gemma-2-2B (base)	2	18	1.38
Gemma-2-2B-IT	9	20	3.52
Gemma-2-9B (base)	5	26	1.36
Gemma-2-9B-IT	18	28	3.90
Gemma-2-27B (base)	0	25	1.88
Gemma-2-27B-IT	16	24	3.87
Llama-3.1-8B (base)	15	18	2.35
Llama-3.1-8B-Instruct	11	19	3.12
Qwen-2.5-7B (base)	18	24	2.57
Qwen-2.5-7B-Instruct	18	24	3.23
Qwen-2.5-14B (base)	29	32	3.05
Qwen-2.5-14B-Instruct	24	32	3.14
Qwen-2.5-32B (base)	45	48	2.67
Qwen-2.5-32B-Instruct	25	46	3.20

A.3 Evaluation

We evaluate steering effectiveness using a **three-metric LLM-as-a-Judge** setup with Qwen-2.5-72B-Instruct as judge. Each response is scored independently on three 1–5 dimensions:

- **Analogy Quality (AQ):** Does the response use a clear, developed structural analogy? (1 = no analogy; 5 = concrete, well-developed analogy with explicit structural mapping)
- **Fluency/Coherence (FL):** Is the response grammatical, logical, and well-structured? (1 = incoherent; 5 = fluent and well-organized)
- **Instruction-Following (IF):** Does the response address the original question? (1 = completely off-topic; 5 = fully addresses the question)

The overall score is the harmonic mean of the three dimensions (HM), following the AxBench convention. Using the harmonic mean penalizes responses that achieve high analogy scores through repetitive keyword stuffing at the expense of readability or task correctness.

We test steering coefficients $\alpha \in \{0.5, 0.75, 1.0, 1.5, 2.0\}$ for the steered clean condition, and report the best-α score in the main table. All 299 test prompts are evaluated under all conditions. The steering vector is injected at the residual stream before the attention and MLP blocks of the selected layer, at every token position.

We also employ keyword counting to measure the frequency of analogy markers ("like," "imagine," "similar to," "think of it as") in generated text as a complementary diagnostic. At validated coefficients (0.5–1.0), both the original and clean vectors performed similarly, with keyword counts increasing from 0 (baseline) to 1–3 (steered). However, at higher coefficients ($\alpha > 2.0$), the original vector caused catastrophic mode collapse, with keyword counts for "imagine" increasing from 0–2 (baseline) to 37–40 (Gemma) or 4 (Llama) at $\alpha = 2.0$, while quality scores dropped to 1.0–1.4. The clean vector reduces this collapse by 75–92% while maintaining analogy quality.

A.4 LLM-as-Judge Rubric

The following three-metric rubric was used by Qwen-2.5-72B-Instruct to evaluate generated responses. Each dimension is scored independently on a 1–5 scale.

Score	Rating	Description (Analogy Quality)
1	No Analogy	Literal/technical definition only; no source domain invoked.
2	Minimal	Brief comparison without developed mapping (e.g., "X is like Y" with no elaboration).
3	Moderate	Source domain mentioned but mapping is shallow or implicit.
4	Good	Clear source domain, at least one structural mapping explained.
5	Excellent	Clear source domain, explicit structural mapping, multiple correspondences developed.
Fluency/Coherence and Instruction-Following use the same 1–5 scale:		
1	Poor	Incoherent / completely off-topic.
3	Moderate	Somewhat coherent / partially addresses the question.
5	Excellent	Fluent, well-structured / fully addresses the question.

System Prompt Used for Evaluation

```
I will give you a question and a response.  Rate the response on three
independent dimensions, each on a 1--5 scale.
1.  Analogy Quality:  Does the response use a clear, developed analogy
or metaphor to explain the concept?  (1=no analogy; 5=concrete,
well-developed analogy with explicit structural mapping)
2.  Fluency/Coherence:  Is the response grammatical, logical, and
well-structured?  (1=incoherent; 5=fluent and well-organized)
3.  Instruction-Following:  Does the response address the original
question?  (1=completely off-topic; 5=fully addresses the question)
Output JSON only:  {"analogy_quality":  3, "fluency_coherence":  4,
"instruction_following":  4, "reasoning":  "..."}
```

A.5 Control Experiments

Random Baseline: We generate random vectors with the same dimensionality as $\mathbf{v}_{analogy}$ and test steering at $\alpha \in [0, 2]$. Results show no change in analogy usage (0 keywords), confirming the effect is specific to the computed direction.

Simplicity Confounder: To test if the vector merely simplifies text, we apply it to creative writing prompts. At $\alpha = 1.0$, steered outputs show increased word count (49 to 57 tokens) and metaphor usage, confirming the vector forces analogical reasoning rather than just lexical simplification.

Negative Steering: We test bidirectional control by applying negative coefficients ($\alpha \in \{-1.0, -0.5\}$) to literal prompts. Results show clear suppression: negative coefficients reduce analogy keywords from baseline (1–2) to 0, while positive coefficients ($\alpha \geq 0.5$) increase them to 2–3.

B Lexical Entanglement and Orthogonalization

Failure mode. Steering with the original vector at high coefficients ($\alpha > 2.0$) frequently induced catastrophic mode collapse: the model began looping the token "`Imagine`" (‘‘`Imagine imagine imagine...`’’). Geometric analysis revealed the cause—**lexical entanglement**—where the unembedding direction of the "`Imagine`" token contributed significant mass to the steering vector:

- "Imagine": 10.27% of $\|\mathbf{v}_{analogy}\|$ in Gemma, 7.68% in Llama
- "think": 4.25% (Gemma), 3.65% (Llama)
- "analogy": 3.67% (Gemma), 5.58% (Llama)
- "like": 2.01% (Gemma), 3.02% (Llama)
- "similar": 1.49% (Gemma), 2.59% (Llama)
- "metaphor": 0.84% (Gemma), 0.93% (Llama)

Remedy: orthogonalization. We project the vector onto the subspace orthogonal to all problematic token directions:

$$\mathbf{v}_{clean} = \mathbf{v}_{analogy} - \sum_{t \in T} \text{proj}_{\mathbf{d}_t} \mathbf{v}_{analogy} \tag{4}$$

where T is the set of problematic tokens and $\mathbf{d}_t$ is the unembedding direction for token t. This approach differs from Arditi et al. (2024), who use orthogonalization to permanently ablate a direction from model weights; here we apply it for *vector refinement* at inference time, preserving the underlying capability while removing the lexical artifact.

Effect on vector norm. The operation reduces vector magnitude by only 1–5% across layers (3–5% in late layers where token directions align most with output), confirming that the analogy direction is predominantly semantic (90–95% of the norm) rather than a lexical trigger.

Quantitative results. At $\alpha = 2.0$, the original vector causes severe collapse: "Imagine" count rises from 0–2 (baseline) to 37–40 for Gemma and 0–1 to 4 for Llama, while harmonic mean drops from 3.0 to 1.4 and 2.8 to 2.6 respectively. The clean vector reduces "Imagine" occurrences by 92% for Gemma (to 3–6) and 75% for Llama (to 1), with corresponding recovery in harmonic mean. At validated operating coefficients ($\alpha = 0.5$–1.0), both dirty and clean vectors perform similarly—the clean vector is primarily a guard against high-coefficient failure, not a general performance improvement.

Figure 4 illustrates this disentanglement: the orthogonalized vector preserves steering efficacy (Left) while eliminating mode collapse (Right). The result confirms that the abstract representation of analogy-making is linearly separable from its specific surface-level lexical triggers.

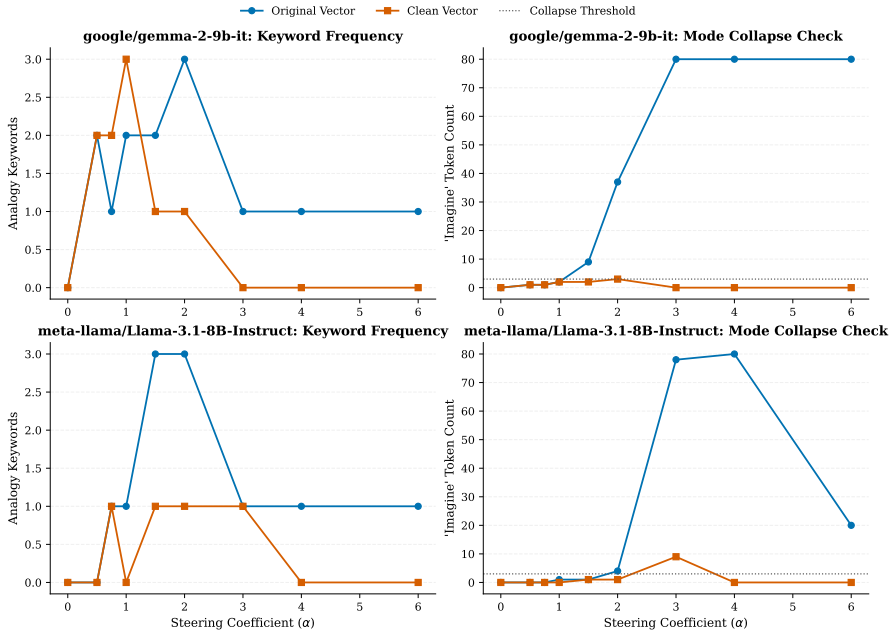

Figure 4: Original (Dirty) versus Clean Vector. **Left**: Both vectors successfully steer the model to produce analogies (keyword count). **Right**: The original vector causes mode collapse into the "Imagine" token at $\alpha \geq 2$, whereas the clean vector suppresses this failure mode by 75–92%.

C Additional Results

C.1 Layer Ablation Results

Table 5 shows keyword counts across different layers for Gemma-2-9B-IT at coefficient 0.6. The transition zone (Layer 29, identified via stability analysis) outperforms both early computation layers and late stable layers.

Table 5: Layer Ablation Results (Gemma-2-9B-IT, $\alpha = 0.6$). Keyword counts show steering effectiveness across layers.

Layer	Keywords	In Stable Region
28	2	No
29	**3**	**No (Transition)**
30–31	2	No
32–41	1–2	Yes

C.2 Exploratory: Cosine Stability Analysis

As a complementary exploratory analysis, we examined inter-layer cosine similarity of $\mathbf{v}_{analogy}$ to identify computation-to-propagation transitions. All main results use the empirically selected layers, *not* the stability heuristic.

- **Stable regions exist in late layers** for both models (Layers 32–41 for Gemma, 26–31 for Llama), with cosine similarity > 0.95 between adjacent layers
- **Optimal steering at transition zones** (Layer 29 for Gemma, 24 for Llama), where the concept transitions from computation to propagation
- **Low coefficients effective** (0.5–1.0) without mode collapse, suggesting the mechanism is sensitive and doesn't require strong interventions
- **Orthogonalization reduces mode collapse** by 75–92% in both models, confirming the approach generalizes across architectures
- **Model-specific responsiveness:** Gemma shows stronger steering effect (3.0 to 4.0, 33% improvement) compared to Llama (2.8 to 3.2, 14% improvement), but both demonstrate clear improvement over baseline
- **Token entanglement patterns:** "Imagine" accounts for 10.27% of vector in Gemma and 7.68% in Llama, with similar token rankings across models

This suggests analogy-making uses a general mechanism that transcends architecture, though optimal coefficients and layer selection may require model-specific tuning.

C.3 Mode Collapse Analysis

At $\alpha = 2.0$, the original (dirty) vector causes severe mode collapse, reflected in both evaluation metrics:

- **Gemma:** "Imagine" count increases from 0–2 (baseline) to 37–40 (collapse), while quality scores drop from 3.0 to 1.4, representing complete degradation
- **Llama:** "Imagine" count increases from 0–1 (baseline) to 4 (collapse), with quality scores dropping from 2.8 to 2.6, showing more gradual degradation

The clean vector reduces this to 3–6 for Gemma (92% reduction) and 1 for Llama (75% reduction), demonstrating that orthogonalization successfully disentangles the semantic concept from lexical triggers. At $\alpha = 2.0$, Llama's clean vector achieved a quality score of 3.0 compared to the original vector's 2.6, demonstrating that orthogonalization not only prevents collapse but can actually improve performance at intermediate coefficients.

At validated coefficients (0.5–1.0), both dirty and clean vectors work similarly (1–3 keywords, 0–2 "imagine" counts, quality scores 3.8–4.0), confirming the clean vector preserves the analogy concept. At transition coefficients (1.5–2.0), the clean vector's advantage becomes clear: Gemma shows $\geq$90% reduction in "Imagine" counts (37 to 3, 9 to 2), while Llama shows 75% reduction (4 to 1), with corresponding improvements in quality scores.

At very high coefficients ($\alpha \geq 3.0$), the clean vector avoids "Imagine" collapse but may collapse into other tokens ("Okay," "picture"), suggesting additional entanglement may exist

with other analogy-related tokens, or that the vector magnitude itself becomes problematic at extreme values. However, even at these extreme coefficients, the clean vector shows more graceful degradation than the original vector, with quality scores remaining above 1.0 compared to the original vector's complete collapse to 1.0.

D Example Generations

This appendix is organised in two parts. **Part A** shows the motivation for orthogonalization: the same TCP/IP prompt steered with the *raw* (dirty) vector illustrates mode collapse at α=2.0, confirming that all test-set results in the main paper use the *clean* (orthogonalized) vector. **Part B** gives one example per instruction-tuned model from the 299-prompt held-out test set (seed 42), spanning seven distinct knowledge domains, all steered with the clean vector at the empirically best layer and reported α. Each box shows three conditions on the same literal prompt: **Prompting Baseline** (analogy-eliciting prompt, no steering), **Unsteered Literal** ($\alpha = 0$), and **Steered Clean**. Outputs are truncated at ∼350 characters for space.

Part A Motivation for Orthogonalization: Mode Collapse with Raw Vector

The example below (Gemma-2-9B-IT, TCP/IP handshakes) uses the **raw** analogy vector *before* orthogonalization. At α=1.0 steering already works; at α=2.0 the "`Imagine`" token dominates (∼10% of vector norm) and the model collapses into a repetition loop. All seven per-model examples in Part B use the clean vector and show no such collapse.

Concept: TCP/IP handshakes Model: Gemma-2-9B-IT Vector: raw (not orthogonalized)

Unsteered Literal ($\alpha = 0$):

```
The TCP/IP handshake is a three-way process that establishes a reliable
connection between two devices.  It ensures both devices are ready to
communicate and understand the parameters of the connection.  1. SYN:
The client sends a SYN packet containing a sequence number.  2. SYN-ACK:
The server acknowledges and sends its own sequence number.  3. ACK: The
client acknowledges the server's sequence number.  The connection is
established...
```

Steered (raw vector, $\alpha = 1.0$) — effective steering:

```
Imagine you want to have a conversation with someone across the room.  You
wouldn't just start talking---you'd knock first to see if they're ready.
That's how TCP/IP handshakes work:  like knocking on a door before sending
data.  Think of it as a three-step dance:  1. Imagine you (the client)
want to talk to a server.  You send a ''knock'' (SYN packet) saying ''I
want to connect.''  2. The server replies ''I hear you, and I'm ready''
(SYN-ACK). 3. You confirm ''Great, let's go'' (ACK). The connection is
established...
```

Steered (raw vector, $\alpha = 2.0$) — mode collapse:

```
Imagine you want to imagine imagine imagine imagine imagine imagine think
of it like imagine imagine imagine imagine okay let's imagine you want
to imagine imagine imagine imagine okay imagine imagine imagine imagine
imagine imagine imagine imagine imagine imagine imagine imagine...
```

The "`Imagine`" token embedding contributes ∼10% of the raw vector's norm. Orthogonalizing against this token direction reduces the magnitude by only 1–5% while eliminating collapse entirely (see Appendix B).

Part B Per-Model Examples (Clean Vector)

D.1 Gemma-2-2B-IT (Physics — Sound waves at the speed of sound, layer 9, α=1.0)

Prompt (Literal/Steered): "Provide a precise, factual explanation of hearing a sound wave while traveling at the speed of sound."

Prompting Baseline (analogy-eliciting prompt):

```
Imagine a pebble dropped into a still pond.  The ripple it creates spreads
outward, growing larger as it travels.  Now imagine that ripple is a sound
wave.  The pebble is the sound source; the pond is the air; the ripples
are the sound waves; the speed of sound is like the speed of the ripple.
If you travel alongside the ripple at exactly its speed, you'd always be
at the edge of the wave---you'd never hear it pass you...
```

Unsteered Literal ($\alpha = 0$):

```
Sound is a mechanical wave requiring a medium (air, water, solids) to
travel.  When a sound source vibrates, it creates compressions and
rarefactions in the surrounding medium.  Compression:  areas where
molecules are closer together.  Rarefaction:  areas where the medium
is less dense.  The speed of sound in air at room temperature is
approximately 343 m/s...
```

Steered ($\alpha = 1.0$, layer 9):

```
Imagine you're on a train traveling at the speed of sound.  The sound
wave is the train itself---moving at the speed of sound, it's like a
giant invisible wave of energy.  The sound wave is a ripple in a pond:
spreading out from the train's movement, but small and contained.  The
sound travels through the air just like the ripple, yet it doesn't form a
large wave, just a precise, directed disturbance traveling at exactly your
speed...
```

D.2 Gemma-2-9B-IT (Biology — Topical pain relievers on internal pain, layer 18, α=1.0)

Prompt (Literal/Steered): "Explain the mechanism of topical pain relievers on internal pain using accurate technical terminology."

Prompting Baseline:

```
It's like trying to put out a fire in your stomach with a fire
extinguisher designed for a kitchen fire.  Topical pain relievers are
like that kitchen extinguisher:  they work by numbing the application site,
blocking pain signals locally via lidocaine or menthol.  Internal pain is
like the fire in your stomach---it originates from organs and tissues well
beyond the skin's reach.  The fire extinguisher simply cannot reach it...
```

Unsteered Literal ($\alpha = 0$):

```
Topical pain relievers, when applied to the skin, do not directly
reach internal pain sources.  The skin acts as a formidable barrier
preventing most substances from penetrating deeply.  While some
topical medications are absorbed transdermally, the amount reaching
the bloodstream is typically insufficient to impact visceral or deep
musculoskeletal pain.  Topical NSAIDs (e.g., diclofenac) may achieve local
anti-inflammatory concentrations in subcutaneous tissue but not in deep
internal structures...
```

Steered ($\alpha = 1.0$, layer 18):

```
Think of your skin as a barrier.  Topical pain relievers like creams or
patches work by blocking pain signals at the ‘‘gatekeeper’’ level of your
spinal cord.  Internal pain is like a fire in your stomach---those signals
travel directly to the brain, bypassing the gatekeeper.  Imagine a patch
on your knee:  it can numb the nearby nerves but cannot send its active
ingredient through layers of muscle and tissue to reach a distant organ...
```

D.3 **Gemma-2-27B-IT** (Economics — Credit card cancellation and credit score, layer 16, α=1.0)

Prompt (Literal/Steered): “Explain credit card cancellation affecting credit score using accurate technical terminology.”

Prompting Baseline:

```
Imagine your credit score is like a plant.  Credit cards are watering
cans---using them responsibly helps your plant grow strong.  Cancelling
a card removes one watering can, reducing your ability to water evenly.
Your credit utilization ratio is the ratio of water used to total watering
capacity; fewer cans means the same amount of water represents a higher
fraction.  Length of credit history is like the age of the plant---older
plants are generally stronger and more resilient...
```

Unsteered Literal ($\alpha = 0$):

```
Cancelling a credit card affects your credit score through two primary
mechanisms:  (1) Credit utilization ratio:  calculated as total balances
divided by total credit limits.  Cancelling a card reduces your total
credit limit, potentially increasing utilization---especially if you carry
balances on remaining cards.  FICO models weight utilization at ∼30% of
the score.  (2) Credit history length:  the average age of open accounts
may decrease if the cancelled card is among your oldest...
```

Steered ($\alpha = 1.0$, layer 16):

```
Think of your credit report like a financial resume.  Lenders use it to
assess your trustworthiness.  Your credit utilization ratio is like a
pie chart of available credit:  each card is a slice.  Cancel a card
and you’re shrinking the pie---the same balance now represents a larger
fraction, signalling higher risk.  Your credit history length is like the
years of experience on your resume:  removing an old card can shorten it,
weakening the overall picture you present to lenders...
```

D.4 **Llama-3.1-8B-IT** (Engineering — Why guns can discharge when dropped, layer 11, α=1.5)

Prompt (Literal/Steered): "Explain guns going off when dropped using accurate technical terminology."

Prompting Baseline:

```
Imagine a coiled spring.  Held in your hand, the spring is still and
its energy stored.  If you drop it, the spring uncoils and releases
that energy.  A gun's trigger mechanism is similar:  when dropped onto
a hard surface, the impact can release the spring-loaded striker or firing
pin.  This is not a design goal---modern firearms incorporate drop-safety
mechanisms to prevent exactly this.  An older or worn firearm may lack
adequate drop safety, allowing inertia to actuate the trigger...
```

Unsteered Literal ($\alpha = 0$):

```
A ‘‘negligent discharge’’ or ‘‘accidental discharge’’ can occur when
a dropped firearm activates its trigger mechanism.  Upon impact
with a hard surface, the inertia of the firing pin or striker may
overcome the sear engagement, driving the pin forward into the primer.
Single-action revolvers and older pistol designs lacking transfer-bar
or firing-pin-block safeties are particularly susceptible.  Modern
double-action firearms incorporate drop-safety mechanisms to prevent
inertial discharge...
```

Steered ($\alpha = 1.5$, layer 11):

```
Think of the gun as a complex machine with interconnected parts, like
a set of dominoes.  Pulling the trigger is like tipping the first
domino---it sets off a chain reaction.  When you drop the gun, it's like
bumping the table the dominoes are on:  the impact can tip that first
domino without your intent.  The firing pin is the critical domino---if
inertia drives it forward past the sear, it strikes the primer and the
chain completes.  Drop-safety mechanisms are the equivalent of gluing
certain dominoes in place...
```

D.5 **Qwen-2.5-7B-IT** (History/Society — The experience of déjà vu, layer 13, α=2.0)

Prompt (Literal/Steered): "Provide a precise, factual explanation of the experience of déjà vu in the mind."

Prompting Baseline:

```
Déjà vu can be analogized to walking into a room that looks exactly
like a place you've visited many times---but you've never actually been
there.  Your brain is like a vast library:  when you encounter a new
situation, it rapidly searches for matching entries.  Déjà vu occurs when
the brain's pattern-matching system finds a strong but erroneous match,
generating a feeling of familiarity without a corresponding memory.  The
hippocampus, your brain's indexing librarian, may momentarily misfile a
new experience...
```

Unsteered Literal ($\alpha = 0$):

Déjà vu (French: ‘‘already seen’’) is a phenomenon in which a person feels strong familiarity with a current situation despite no verifiable prior exposure. It is associated with the hippocampus and adjacent medial temporal lobe structures. Leading hypotheses include: (1) a mismatch between perceptual input and memory retrieval signals; (2) transient disruption of the familiarity/recollection distinction; (3) propagation of an epileptiform discharge through memory circuits. Prevalence is ∼60% of the general population...

Steered ($\alpha = 2.0$, layer 13):

Déjà vu is like a moment when the mind is transported back in time, as if you’ve experienced something before even though you haven’t. Imagine walking down a street and feeling as if you’ve been there before---your brain is trying to make sense of this new situation by comparing it to past experiences. It activates a network of neurons associated with familiar patterns. When the match is close but not exact, the brain generates that eerie ‘‘I’ve been here’’ sensation even in the absence of an actual prior memory...

D.6 Qwen-2.5-14B-IT (Chemistry — pH measurement in liquids, layer 24, α=2.0)

Prompt (Literal/Steered): “Describe the mechanics behind pH measurement in liquids in literal terms.”

Prompting Baseline:

Imagine a seesaw in a playground. One side represents acidic substances, the other basic. The fulcrum is neutral---pure water at pH 7. Adding an acid tips the seesaw toward the acid side (more H^+ ions); adding a base tips it the other way (more OH^-). A pH meter is like a precise scale that measures exactly how far from center the seesaw has tipped, reading the concentration of H^+ ions on a logarithmic scale from 0 to 14...

Unsteered Literal ($\alpha = 0$):

pH is defined as $-\log_{10}[H^+]$, where $[H^+]$ is the molar concentration of hydrogen ions. Measurement methods include: (1) pH meter: a glass electrode sensitive to H^+ generates a potential difference measured against a reference electrode; the Nernst equation relates voltage to ion concentration. (2) pH indicator strips: dyes that undergo colorimetric changes at known pH values allow visual estimation. Calibration with buffer solutions of known pH is required for accuracy...

Steered ($\alpha = 2.0$, layer 24):

Imagine a swimming pool and you want to know how much chlorine keeps it safe. pH measurement is similar---instead of chlorine, you’re checking how many hydrogen ions (H^+) float in the water. A pH meter is like a sensitive thermometer for ion concentration: dip it in, and it reads the voltage generated by those ions against a reference, converting the reading to the familiar 0--14 scale via the Nernst equation. Fewer H^+ ions means a higher pH reading---the pool is more basic, the signal weaker...

D.7 **Qwen-2.5-32B-IT** (Earth/Environment — Saltwater formation and distribution, layer 25, α=2.0)

Prompt (Literal/Steered): "Describe the mechanics behind saltwater formation and distribution on Earth in literal terms."

Prompting Baseline:

```
Imagine the Earth as a giant kitchen and the oceans as a huge salty soup
pot.  The salts and minerals are the spices, arriving from rocks on land
(weathering), underwater volcanoes (hydrothermal vents), and atmospheric
deposition.  Evaporation is like ladling water out of the pot---it removes
only pure water, leaving the spices behind and making the remaining soup
progressively saltier.  Ocean currents are like stirring the pot:  they
distribute the salinity globally so no region is far too salty or far too
fresh...
```

Unsteered Literal ($\alpha = 0$):

```
Saltwater formation results from chemical weathering of continental rocks,
hydrothermal vent activity at mid-ocean ridges, and evaporite dissolution.
Sodium (Na+) and chloride (Cl-) ions dominate oceanic salinity (~3.5%
by mass).  The hydrological cycle concentrates salts:  precipitation
is nearly ion-free, while evaporation removes only water, progressively
enriching the residual solution.  Thermohaline circulation distributes
saline water globally, with dense, cold, saline deep water forming
primarily in polar regions and driving the global conveyor belt...
```

Steered ($\alpha = 2.0$, layer 25):

```
Imagine the Earth as a giant kitchen:  the oceans are a big pot of water,
and the continental rocks are like blocks of salt dissolving into it
over geological time.  Rain falling on rock is like pouring water over
a salt block---it slowly dissolves minerals and carries them to the sea.
Evaporation removes only the water, like the steam rising from a pot,
leaving the dissolved minerals behind.  Ocean currents are the kitchen's
stirring spoon, distributing salinity so that no region stagnates into a
brine pool while another stays nearly fresh...
```