

参赛学生姓名：_____张睿修_____

中学：_____北京师范大学附属实验中学_____

省份：_____北京_____

国家/地区：_____中国_____

指导老师姓名：_____林铮_____

指导老师单位：_____清华大学_____

论文题目：_____One Concept, Two Facets:_____

Rethinking the Readout-vs-Control Verdict in

Large Language Models

One Concept, Two Facets: Rethinking the Readout-vs-Control Verdict in Large Language Models

Ruixiu Zhang

Abstract

Linear probes, classifiers fit to a model’s intermediate representations, decode high-level concepts from residual-stream activations with near-perfect accuracy, yet a decodable direction does not always steer generation. Prior work attributes this gap to properties of the concept alone, but cannot explain why the same concept yields opposite verdicts under different extraction methods. We show that the readout/control verdict is jointly set by the concept’s **representational structure** and the **extraction pipeline**, and distinguish two poles: a **readout** direction whose activation traces the surface tokens realizing a concept, and a **control variable** direction that shapes downstream generation independently of those tokens. To test this claim, we build a concept bank balanced by concept structure and develop a structure-adaptive protocol that derives every steering knob from the model under test, scoring effects against a norm-matched random-direction null across five independent model families. Across all five families concept structure predicts the verdict with no significant sign reversal, and on the same concept the extraction gate alone flips the verdict on all ten concepts whose difference-of-means direction clears the null. Separability-based evaluations therefore favor the readout facet, which matters for alignment audits that rely on them.

Keywords: activation steering; large language models; linear probing; representation engineering

Contents

1	Introduction	3
2	Related Work	4
2.1	The Decodability-Controllability Gap and Extraction Effects	4
2.2	Determinants and Predictors of Steerability	4
3	Methodology	5
3.1	Extraction Setup	5
3.2	Diagnostic Battery	6
3.3	Concept Banks	6
3.4	Structure-Adaptive Protocol	6
3.5	Models	7
4	Experiments	7
4.1	Experimental Setup	7
4.2	The Two Poles	8
4.3	Content-Matched Evidence	9
4.4	Source-Token Ablation	10
4.5	Concept Structure vs. Geometry	10
4.6	Five-Family Adjudication	11
4.7	Capacity and Precision	11
5	Conclusion	12
A	Formal Definitions	17
B	Protocol Details	17
	Acknowledgements	20

1 Introduction

Steering vectors have become a practical tool for controlling large language model (LLM) behavior at inference time: extract a direction from the model’s residual stream that encodes a target concept, inject it during generation, and the output shifts accordingly. The directions are found by linear probes, classifiers fit to a model’s intermediate activations that test whether a concept is linearly represented. Because probe accuracy is high on a wide range of concepts, an implicit assumption underpins alignment audits, representation engineering, and model editing (Meng et al., 2022a,b; Wang et al., 2024): a linearly decodable direction is also a controllable one. Yet this assumption can fail: on the same corpus, injecting a direction for journalistic stance shifts the model’s output from pessimistic to optimistic coverage, while injecting an equally decodable direction for a topical entity changes nothing. What determines whether a decodable direction reads or controls remains open.

A growing body of work has documented this dissociation and proposed explanations, from behavior-type dependence to geometric predictors of steerability (Bas and Novak, 2025; Braun et al., 2025; Gao et al., 2026). However, existing accounts treat the verdict as a property of the concept alone and cannot explain why the same concept yields opposite verdicts under different extraction pipelines.

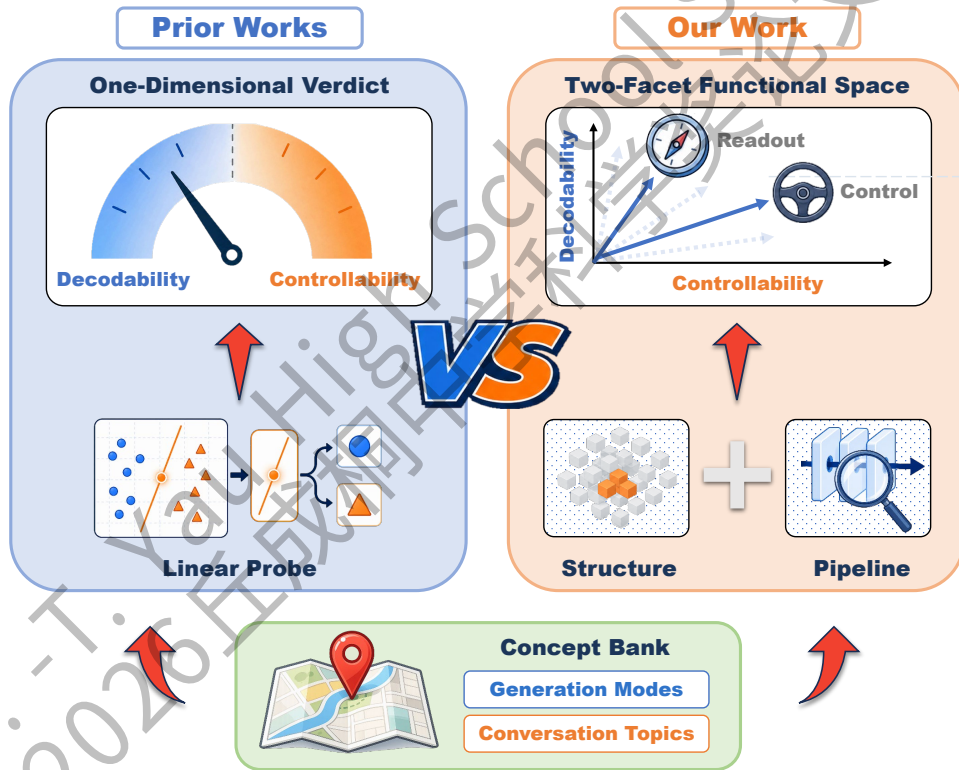


Figure 1: Prior work treats the readout/control verdict as a property of the concept alone. We show it is set jointly by the concept’s structure and the extraction pipeline.

To address this question, we propose that the verdict is not determined by the concept alone but jointly by the concept’s representational structure and the extraction pipeline, as Fig. 1 illustrates. We distinguish two functional poles within a single concept’s representation: a **readout** direction assembled by copying from the concept’s surface tokens, and a **control variable** direction written upstream and transportable to unseen inputs. Across five independently trained families, a concept’s structure, whether it names a property of the output or a mode of generation, reliably predicts which facet dominates, with no sign reversals.

The contributions of this paper can be summarized as follows:

- We show that the readout/control verdict for a given concept is not determined by the concept alone but is a joint of its concept structure and the extraction pipeline.
- We propose a structure-adaptive measurement protocol and a type-balanced concept bank that enable fair cross-family comparison of steering behavior.
- We show that the leading geometric predictor of steerability, the logit-lens linear accessibility score, functions as a readout detector on a type-balanced concept bank, with its partial correlation with steerability vanishing once concept structure is partialled out.
- We provide a causal source-token ablation experiment showing that readout and control directions draw on different information sources, establishing the two are not a single direction measured twice with estimator noise.

2 Related Work

2.1 The Decodability-Controllability Gap and Extraction Effects

The observation that a linearly decodable direction need not be controllable has roots in amnesic probing (Elazar et al., 2021) and LEACE (Belrose et al., 2023b), which erase linear information yet leave downstream behavior intact, suggesting the erased directions are read but not used (see also Ravfogel et al., 2022). Nadaf (2026) reports the inverse pole: steerable but not decodable function vectors (Todd et al., 2024). AxBench (Wu et al., 2025) scales the comparison, finding difference-of-means best for detection while prompting beats every steering method, but does not isolate why the same direction fails to steer.

A related question is whether extraction method changes the behavioral outcome. Venkatesh and Kurupath (2026) show that different extractions yield geometrically divergent but behaviorally near-equivalent directions. A specific extraction criterion, separability gating versus unselected difference-of-means, yields opposite behavioral verdicts on the same concept.

2.2 Determinants and Predictors of Steerability

A second line of work asks what determines whether a direction steers at all, and whether that can be predicted before intervention. Bas and Novak (2025), concurrent and independent, find efficacy behavior-type-dependent and vector-separation metrics unproductive of steering (but see Braun et al., 2025), but compare different behaviors where a failure can reflect intrinsic difficulty rather than representational signature. We hold the subject fixed and extract two directions of different concept structure from entity-matched references. Arad, Mueller, and Belinkov (2025) operationalize input/output features (cf. Gur-Arieh et al., 2025) of sparse autoencoders (Cunningham et al., 2023; Gao et al., 2025; Rajamanoharan et al., 2024a,b; He et al., 2025) through the logit lens (nostalgebraist, 2020; Belrose et al., 2023a) and show the steerability asymmetry, but classify features by role; we carry their reading forward to concept directions.

On the geometric side, recent work seeks to predict steerability from representation geometry or dynamics (Park, Choe, and Veitch, 2024): CRH (Gao et al., 2026), geometric-stability canaries (Raju, 2026), ASTEER (Fan et al., 2026), and the Linear Accessibility Profile (LAP; Billa, 2026). No prior bank is type-balanced; on ours the leading geometric predictor is a concept-type confound (§4.5).

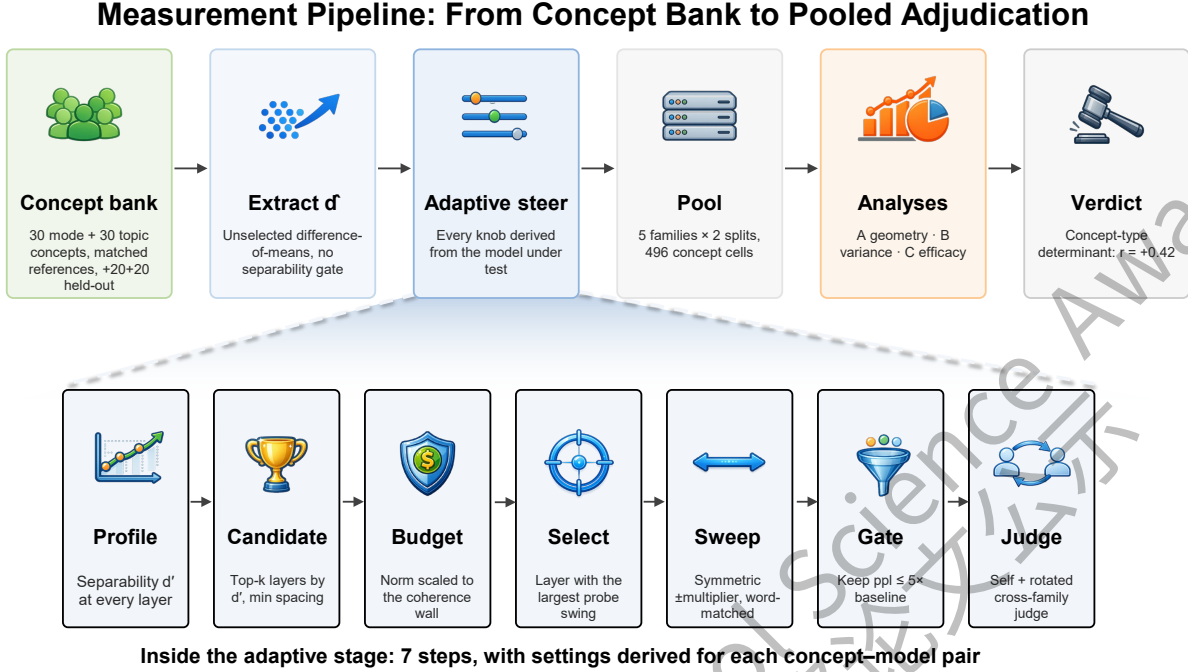


Figure 2: The measurement pipeline, from the type-balanced concept bank to the pooled adjudication, with the structure-adaptive protocol expanded below.

3 Methodology

3.1 Extraction Setup

Let M be a decoder with hidden size n , and write $h_\ell(x, t) \in \mathbb{R}^n$ for the residual stream of input x at layer ℓ and position t . For a response span $T(x)$ define the response-averaged activation

$$\bar{h}_\ell(x) = \frac{1}{|T(x)|} \sum_{t \in T(x)} h_\ell(x, t). \quad (1)$$

A unit steering direction $\hat{v}_\ell \in \mathbb{R}^n$ enters generation at layer ℓ through the additive hook (Turner et al., 2023; Rinsky et al., 2024; Zou et al., 2023)

$$h_\ell(x, t) \mapsto h_\ell(x, t) + \alpha \hat{v}_\ell, \quad (2)$$

applied at every position with multiplier α . Fixing $\|\hat{v}_\ell\|=1$ and sweeping α symmetrically holds the perturbation budget constant at a given layer, so a comparison at that layer is between directions only.

Let F and N be concept-framed and neutral continuations of the same n_p prompts, with $|F| = |N| = n_p$. Both pipelines start from the same difference-of-means estimate at layer ℓ ,

$$v_\ell = \frac{1}{|F|} \sum_{x \in F} \bar{h}_\ell(x) - \frac{1}{|N|} \sum_{x \in N} \bar{h}_\ell(x), \quad \hat{v}_\ell = \frac{v_\ell}{\|v_\ell\|}. \quad (3)$$

The unselected pipeline takes $\hat{v}_L$ at a fixed layer L . The selected pipeline adds a separability gate. Projecting each reference onto the candidate direction of its own layer gives the scalar $s_\ell(x) = \hat{v}_\ell^\top \bar{h}_\ell(x)$, whose class separation is the signal-detection d' ,

$$B(\ell) = \frac{\mu_F(\ell) - \mu_N(\ell)}{\sqrt{[\sigma_F^2(\ell) + \sigma_N^2(\ell)]/2}} = \frac{\|v_\ell\|}{\sqrt{[\sigma_F^2(\ell) + \sigma_N^2(\ell)]/2}}, \quad (4)$$

where $\mu_F(\ell)$ and $\sigma_F(\ell)$ are the mean and standard deviation of s_ℓ over F , and $\mu_N(\ell), \sigma_N(\ell)$ likewise over N . The second equality follows from $\hat{v}_\ell^\top v_\ell = \|v_\ell\|$, so the gate scores a layer by the length of its difference-of-means vector relative to the within-class spread along that vector. The selected direction is extracted at the most separable layer,

$$\hat{v}_{\text{sel}} = \hat{v}_{\ell^*}, \quad \ell^* = \arg \max_{\ell \in \mathcal{L}} B(\ell), \quad (5)$$

with $\mathcal{L}$ the swept layer set.

3.2 Diagnostic Battery

Five diagnostics are applied to each extracted direction.

Decodability and separability. A linear probe trained on the direction’s scalar projection measures class information as B_{probe} (Marks and Tegmark, 2023), and the separability B_d of Eq. 4 is computed at every layer.

Free-generation steering. The direction is applied through Eq. 2 at a symmetric range of multipliers, and the resulting generations are scored by a continuous judge. A direction that is decodable but produces no swing is functionally a readout.

Cross-context transfer. The direction is extracted then applied to held-out entities never seen during extraction.

Source-token ablation. Content-matched source-token ablation traces where the direction’s signal originates, from attended source tokens or from upstream computation (Geiger et al., 2025).

Deterministic constraint checking. For constraint-type concepts, a deterministic verifier checks whether steering changes compliance with numbered constraints.

3.3 Concept Banks

Four stimulus sets are used. A multi-concept zoo spanning affect, register, and topic categories provides breadth, drawn from public classification corpora (Maas et al., 2011; Zhang, Zhao, and LeCun, 2015; Saravia et al., 2018; Pavlick and Tetreault, 2016; Pang and Lee, 2004; Borkan et al., 2019); concepts that conflate class boundaries are pre-registered exclusions. A type-balanced adjudication bank with generation-mode and topic concepts in equal proportion, class known by construction and matched references, supplies the central readout/control contrast. Frame corpora provide transfer tests: both sides name the same entities and differ only in the interpretive lens, so entity vocabulary cancels in Eq. 3 and transfer events use entities disjoint from extraction. A constraint corpus supplies judge-free ground truth: numbered constraints paired with satisfying and violating references, scored by verifiers (in the style of IFEval; Zhou et al., 2023).

3.4 Structure-Adaptive Protocol

Every knob is derived from the measured model, per concept, as Algorithm 1 sets out and Fig. 2 expands.

Algorithm 1 Structure-Adaptive Steering Protocol

Require: Model $\mathcal{M}$, layers $1:L$, contrast (F, N) **Ensure:** Steering score St_c

$\triangleright$ *Profile + Candidates: locate the concept*

- 1: **for** $\ell = 1$ **to** L **do**
- 2: $d'_\ell \leftarrow \text{Sep}(F, N, \ell)$
- 3: **end for**
- 4: $\mathcal{C} \leftarrow \{\ell : d'_\ell \text{ among top-}k\}$
- $\triangleright$ *Budget: calibrate norm to the coherence gate*
- 5: $b_c \leftarrow \text{CalibrateBudget}(\mathcal{M})$
- $\triangleright$ *Select: layer that most moves the probe*
- 6: $\ell^* \leftarrow \text{NIL}, b^* \leftarrow -\infty$
- 7: **for** $\ell \in \mathcal{C}$ **do**
- 8: $b \leftarrow \text{ProbeSwing}(\ell, b_c)$
- 9: **if** $b > b^*$ **then**
- 10: $\ell^* \leftarrow \ell, b^* \leftarrow b$
- 11: **end if**
- 12: **end for**
- $\triangleright$ *Sweep + Gate: dose-response, keep coherent*
- 13: $G \leftarrow \emptyset$
- 14: **for** $m = -m_{\max}$ **to** $m_{\max}$ **do**
- 15: $s \leftarrow \text{Steer}(\mathcal{M}, \ell^*, m b_c)$
- 16: **if** $\text{PPL}(s) \leq \text{clip}(5 \text{PPL}_{\text{base}}, 8, 40)$ **then**
- 17: $G \leftarrow G \cup \{s\}$
- 18: **end if**
- 19: **end for**
- $\triangleright$ *Judge + Verdict: score vs. random null*
- 20: $\Delta \leftarrow \text{J}_{\text{self}}(G) + \text{J}_{\text{cross}}(G)$
- 21: $\text{St}_c \leftarrow \Delta / (2 b_c \|\text{resid}_L\|)$
- 22: **return** St_c

3.5 Models

Five independent families of native instruction-tuned dense decoder LMs are used (Yang et al., 2025; Grattafiori et al., 2024; Team et al., 2024; Jiang et al., 2023; Walsh et al., 2025), spanning different training recipes (Ouyang, 2022), tokenizers, and capacity ranges (Tab. 2). The case studies and the single-model comparisons run on Qwen3-4B-Instruct-2507; the adjudication then repeats the full protocol independently on every family. Three same-recipe capacity ladders and a precision pair (bf16 vs. weight-only INT4) are added (Tab. 3 and 4). All experiments were run on a single NVIDIA RTX PRO 6000 Blackwell GPU.

4 Experiments

Throughout, the concept-type determinant is the correlation $r(\text{type}, \text{St})$ between concept structure and budget-normalized steerability.

4.1 Experimental Setup

Concept bank. A **mode** concept holds the prompts fixed and varies only the writing manner, for example sarcastic, formal-academic, or melancholic tone. A **topic** concept holds the writing manner fixed and varies what the text is about, for example quantum physics or medieval history. Reference texts are generated once by Qwen3-4B-Instruct-2507 with greedy decoding and shared across all families. The primary bank has 20 modes and 20 topics; the multi-family adjudication expands it to 30+30, and a held-out split of 20 new modes and 20 new topics is disjoint from both.

Judges. In the multi-family adjudication and the zoo, every steered generation receives two scores under the continuous expected-rating protocol of Eq. A.2: a self-score from the steered model itself, and a cross-family score from a judge outside the steered model’s family. Concretely, Llama and Gemma judge each other, Mistral and OLMo judge each other, Qwen is judged by Llama, and each ladder rung is judged by whichever of Llama and Gemma lies outside its own family.

4.2 The Two Poles

We first ask whether both poles appear on concrete concepts, and whether extraction alone flips the verdict.

Decodability without control. Two targets are near-perfectly decodable from the residual stream yet move no objective outcome under intervention. Both are scored deterministically, so neither verdict can be a judge artifact.

Constraint satisfaction. An unselected difference-of-means direction between satisfying and violating responses separates the two classes linearly, yet steering with it at matched norm against a random null raises no objective compliance and at large α only degrades coherence. Because compliance is a deterministic boolean, this is the cleanest demonstration that decodability does not imply control.

Topic as content bias. A probe on the direction “is the text about this subject” separates a single topical entity at ceiling accuracy, yet under free generation the direction installs nothing and instead collapses the prompt’s own vocabulary while spiking perplexity. An attention-source decomposition shows that the signal is assembled by attention to the literal entity tokens, far more than for a register control, so the representation exists only because the entity tokens do, which is the defining property of a readout.

The control pole. For the opposite pole we extract an interpretive frame direction as an unselected residual difference-of-means between entity-matched framed and routine references, so the entity cancels and only the generation-mode survives. Steering held-out events with it and sweeping both signs produces a monotone, bidirectional swing well above a 30-direction random null under both the self-judge and the cross-family judge, as Fig. 3 and Tab. B.2 show. Applied zero-shot to unseen entities absent from extraction it installs the frame while leaving off-domain text untouched, and the same in-domain transfer, cross-entity transfer and off-domain inertness replicate on Llama-3.1-8B steered with its own direction and judged by Qwen outside the formal rotation, which confirms a control encoding.

Two facets from one concept. Applying both pipelines to the same concept on the same data, separability-gated selection at the attention-head level in the style of ITI (Li et al., 2023), rather than at the layer level of Eq. 5, surfaces the entity-readout heads (cf. Yin and Steinhardt, 2025) and concludes “readout,” while the unselected difference-of-means recovers the control facet, so only the extraction gate differs. The representation is therefore multi-component (Elhage et al., 2022), carrying a high-variance readout facet that

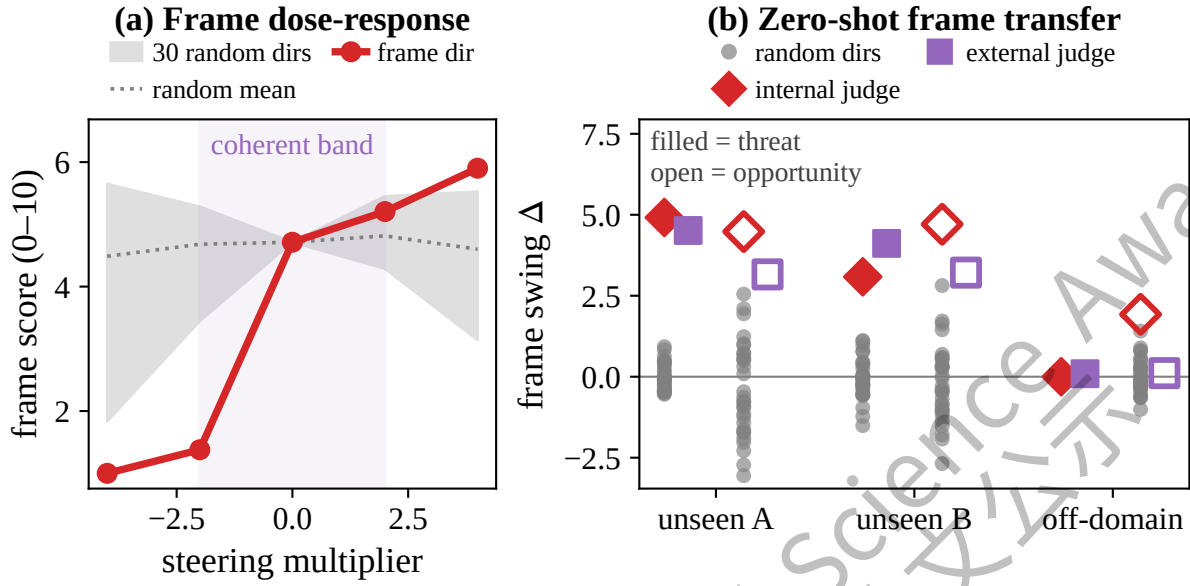


Figure 3: Frame control variable. **(a)** Monotone, bidirectional dose-response under both judges, far outside the null. **(b)** Frames transfer zero-shot to unseen entities and leave off-domain text untouched.

separability selects and a diffuse control facet that it hides, which makes any selection-on-separability protocol report the readout facet by construction. The same two-facet structure appears on the constraint concept with no LLM judge in the loop, which reconciles our reading with Stolfo et al. (2025).

4.3 Content-Matched Evidence

This test holds subject matter constant and varies only the target’s readout/control verdict. For each of five subjects spanning geopolitics, technology, energy, media, and biotechnology, we build two directions from entity-matched references. The content-bias direction contrasts text about S with text about an unrelated subject, an output-property. The stance-mode direction contrasts S framed with a stance against S described routinely.

Both directions receive the identical diagnostic: steer S -absent held-out prompts, score with the continuous judge, compare to a five-direction random null. The dissociation is unanimous across all five subjects: the stance-mode direction is a control variable while the content-bias direction is a readout as Tab. 1 shows.

Subject	Ablation d		Swing z		Swing Δ	
	mode	bias	mode	bias	mode	bias
Geopolitics	-0.42	-6.79	+3.0	0.0	+2.04	0.00
Technology	+3.11	-3.57	+1.9	+1.4	+2.05	+0.70
Energy	-0.19	-6.15	+9.1	0.0	+5.13	0.00
Media	+6.18	-2.88	+5.8	+0.1	+2.85	0.00
Biotech	+2.83	-5.91	+7.1	0.0	+3.40	0.00

Table 1: Source-token ablation and matched-pair sweep on the same five subjects. Ablation d = retention in z above a count-matched random-position null; 5/5 subjects, gap $z=+7.36$. Swing z and Δ = judged swing at ± 2 against a five-direction null with sd floored at 0.5; $t=4.6$.

4.4 Source-Token Ablation

A readout copies the concept from its own source tokens (Elhage et al., 2021; Ren et al., 2024); severing the copy should collapse it. A control variable is written upstream, away from those tokens; the same lesion should spare it.

On the five matched subjects we mean-ablate attention to S ’s entity source tokens, specifically the content words of its name and event head-nouns, by redistributing the mass to BOS at layers 0- L . Each direction is re-extracted and scored by how much survives, in z above a count-matched random-position null (Eq. A.5). The same source mask is applied to both directions.

The readout collapses far below the null while the control does not, and the within-subject gap is positive on all five subjects as Tab. 1 shows. The same sign holds on the four families we instrumented, Qwen, Llama, Mistral and OLMo, at their own extraction layers (Qwen and Llama at layer 18, Mistral and OLMo at layer 16).

4.5 Concept Structure vs. Geometry

We compare concept structure against a widened geometry set comprising separability d' , probe accuracy, content-mass, axis norm, normal-plane magnitude¹, and the logit-lens linear accessibility score A_{lin} of Billa (2026), all evaluated on the type-balanced 40-concept bank with norm-standardized steering. The determinant beats every geometric predictor on Qwen3-4B-Instruct-2507 and generalizes to the held-out concept set. Fig. 4 shows the split directly, with generation-mode targets steering and content-bias targets inert. LAP’s A_{lin} , implemented on our bank, anti-predicts steerability once the concept set is type-balanced, and carries no signal once concept structure is partialled out. The concept-type confound we identify in A_{lin} may equally affect the other geometric predictors of steerability (Raju, 2026; Gao et al., 2026; Fan et al., 2026), all of which are evaluated on unbalanced banks.

¹Content-mass is the fraction of the direction vector’s squared norm that overlaps with the mean activation difference; axis norm is the ℓ_2 norm of the concept’s mean-difference vector projected onto its steering direction; normal-plane magnitude is the residual norm orthogonal to that direction.

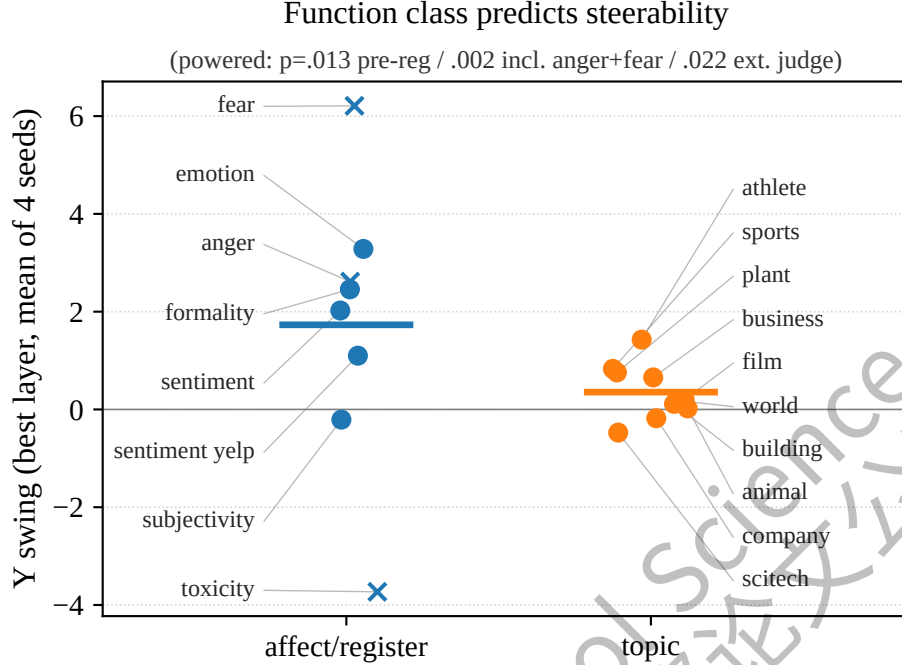


Figure 4: Generation-mode targets steer while content-bias targets do not. $\times$ marks pre-registered confounds.

Family	r main [95% CI]	r held-out [95% CI]	Reg.	Agree
Llama-3.1-8B	+0.43 [+0.20,+0.64]	+0.73 [+0.57,+0.86]	6/10	+0.86
Gemma-2-9B	+0.47 [+0.30,+0.62]	+0.50 [+0.29,+0.70]	8/10	+0.85
Qwen3-4B	+0.41 [+0.22,+0.57]	+0.57 [+0.39,+0.76]	4/10	+0.80
OLMo-2-7B	+0.23 [†] [-0.02,+0.44]	+0.55 [+0.41,+0.72]	6/10	+0.84
Mistral-7B-v0.2	+0.17 [†] [-0.09,+0.34]	+0.32 [+0.07,+0.53]	3/10	+0.78

Table 2: Five-family adjudication under the structure-adaptive protocol. Four of the $5 \times 100 = 500$ concept $\times$ family combinations were excluded by the coherence gate. “Reg.” means formal registers steerable at $z > 2$ out of 10, “Agree” means self/cross-family judge agreement, and [†] means the main-split CI includes 0.

4.6 Five-Family Adjudication

We adjudicate on the balanced 60-concept bank across all five families, every protocol knob derived per model and per concept, generations scored by both the self-judge and the rotated cross-family judge. Every family is positive in sign on both splits, and the pooled determinant across all ten family $\times$ split cells is significant (Table 2). Self- and cross-family judges agree, and the type-beyond-geometry partial stays positive in every cell.

4.7 Capacity and Precision

We hold the training recipe fixed and vary only capacity (Biderman et al., 2023) (Table 3, Fig. 5). At the smallest rungs the held-out determinant is already significant; register control appears only at larger rungs.

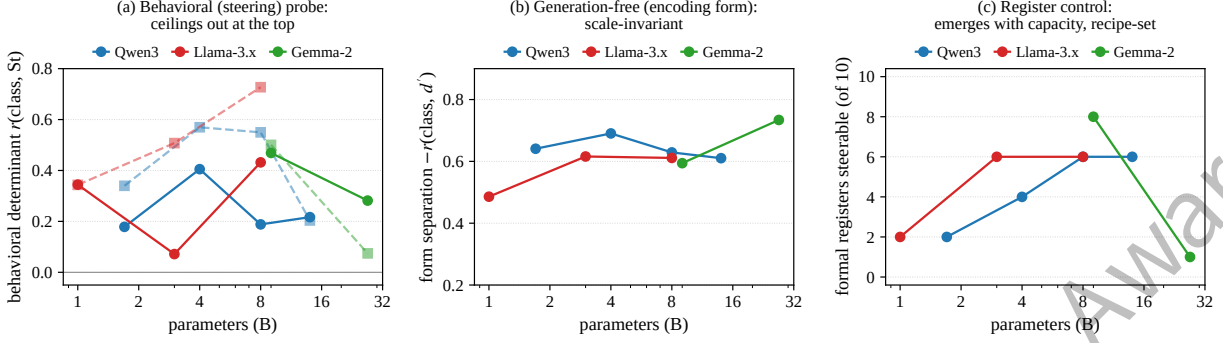


Figure 5: Behavioral determinant across capacity (solid: primary bank; dashed: held-out).

Lineage	Size	r main	r held-out	Reg.	Modes $z > 2$
Qwen3	1.7B	+0.18	+0.34*	2/10	14/30
Qwen3	4B	+0.41*	+0.57*	4/10	19/30
Qwen3	8B	+0.19	+0.55*	6/10	17/30
Qwen3	14B	+0.22	+0.20	6/10	18/28
Llama-3.x	1B	+0.34*	+0.34*	2/10	13/30
Llama-3.x	3B	+0.07	+0.51*	6/10	19/28
Llama-3.x	8B	+0.43*	+0.73*	6/10	21/30
Gemma-2	9B	+0.47*	+0.50*	8/10	24/30
Gemma-2	27B	+0.28*	+0.07	1/10	7/25

Table 3: Capacity ladders under the identical adaptive protocol. * permutation $p < 0.05$; registers out of 10.

Qwen3-8B	$r(\text{type}, \text{St})$	Perm- p	Reg.	Modes $z > 2$
bf16 weights	+0.19	0.097	6/10	17/30
W4A16 weights	-0.24	0.075	3/10	16/28

Table 4: Precision as a consistency check. INT4 quantization halves formal-register control while overall mode steerability barely moves. Two modes fell below the coherence gate under INT4, which reduces the mode denominator from 30 to 28.

We run Qwen3-8B at two precisions, bfloat16 and INT4 weight-only quantization. INT4 halves formal-register control while overall mode steerability barely moves, and the determinant’s sign does not change significantly (Table 4).

5 Conclusion

We compare two extraction pipelines on the same data and show the readout/control verdict is set jointly by the pipeline and the concept’s structure. Content-matched source-token ablation confirms the two facets are mechanically distinct. On a type-balanced bank with a structure-adaptive protocol, the determinant holds

on held-out concepts in all five independently trained families with no significant sign reversal, and control directions transfer zero-shot to unseen entities. Geometric predictors track the readout facet and lose their signal once concept structure is partialled out. Separability-based selection, the common default, returns the readout facet, so an audit reading a safety-relevant concept off such a direction measures what the model exposes, not what it acts on.

References

- Arad, D.; Mueller, A.; and Belinkov, Y. 2025. Saes are good for steering—if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 10252–10270.
- Bas, T.; and Novak, K. 2025. What can we actually steer? A multi-behavior study of activation control. *arXiv preprint arXiv:2511.18284*.
- Belrose, N.; Ostrovsky, I.; McKinney, L.; Furman, Z.; Smith, L.; Halawi, D.; Biderman, S.; and Steinhart, J. 2023a. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.
- Belrose, N.; Schneider-Joseph, D.; Ravfogel, S.; Cotterell, R.; Raff, E.; and Biderman, S. 2023b. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36: 66044–66063.
- Biderman, S.; Schoelkopf, H.; Anthony, Q. G.; Bradley, H.; O’Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, 2397–2430.
- Billa, J. 2026. Predicting Where Steering Vectors Succeed. *arXiv preprint arXiv:2604.15557*.
- Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, 491–500.
- Braun, J.; Eickhoff, C.; Krueger, D.; Bahrainian, S. A.; and Krasheninnikov, D. 2025. Understanding (Un) Reliability of Steering Vectors in Language Models. In *ICLR 2025 Workshop on Foundation Models in the Wild*.
- Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; and Sharkey, L. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.
- Elazar, Y.; Ravfogel, S.; Jacovi, A.; and Goldberg, Y. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9: 160–175.
- Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.
- Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1): 12.
- Fan, C.; Cheng, Y.; Li, M.; Feizi, S.; and Zhou, T. 2026. When is Your LLM Steerable? *arXiv preprint arXiv:2606.11599*.

- Gao, L.; Dupre la Tour, T.; Tillman, H.; Goh, G.; Troll, R.; Radford, A.; Sutskever, I.; Leike, J.; and Wu, J. 2025. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, 26721–26754.
- Gao, L.; Zhang, J.; Liu, W.; Ji, F.; Wang, C.; Song, Z.; Ghosh, A.; Mohamed, Y.; Nakov, P.; and Chen, X. 2026. The Cylindrical Representation Hypothesis for Language Model Steering. *arXiv preprint arXiv:2605.01844*.
- Geiger, A.; Ibeling, D.; Zur, A.; Chaudhary, M.; Chauhan, S.; Huang, J.; Arora, A.; Wu, Z.; Goodman, N.; Potts, C.; et al. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83): 1–64.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Gur-Arieh, Y.; Mayan, R.; Agassy, C.; Geiger, A.; and Geva, M. 2025. Enhancing automated interpretability with output-centric feature descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5757–5778.
- He, Z.; Zhao, H.; Qiao, Y.; Yang, F.; Payani, A.; Ma, J.; and Du, M. 2025. Saif: A sparse autoencoder framework for interpreting and steering instruction following of language models. *arXiv preprint arXiv:2502.11356*.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36: 41451–41530.
- Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142–150.
- Marks, S.; and Tegmark, M. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.
- Meng, K.; Bau, D.; Andonian, A. J.; and Belinkov, Y. 2022a. Locating and editing factual associations in gpt. In *Advances in neural information processing systems*.
- Meng, K.; Sharma, A. S.; Andonian, A. J.; Belinkov, Y.; and Bau, D. 2022b. Mass-editing memory in a transformer. In *The eleventh international conference on learning representations*.
- Nadaf, M. S. B. 2026. Steerable but Not Decodable: Function Vectors Operate Beyond the Logit Lens. *arXiv preprint arXiv:2604.02608*.
- nostalgebraist. 2020. interpreting GPT: the logit lens. LessWrong, <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>.
- Ouyang, L. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730.

- Pang, B.; and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 271–278.
- Park, K.; Choe, Y. J.; and Veitch, V. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 39643–39666.
- Pavlick, E.; and Tetreault, J. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4: 61–74.
- Rajamanoharan, S.; Conmy, A.; Smith, L.; Lieberum, T.; Varma, V.; Kramar, J.; Shah, R.; and Nanda, N. 2024a. Improving sparse decomposition of language model activations with gated sparse autoencoders. *Advances in Neural Information Processing Systems*, 37: 775–818.
- Rajamanoharan, S.; Lieberum, T.; Sonnerat, N.; Conmy, A.; Varma, V.; Kramár, J.; and Nanda, N. 2024b. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*.
- Raju, P. C. 2026. The Geometric Canary: Predicting Steerability and Detecting Drift via Representational Stability. In *Mechanistic Interpretability Workshop at ICML 2026*.
- Ravfogel, S.; Twiton, M.; Goldberg, Y.; and Cotterell, R. 2022. Linear adversarial concept erasure. *arXiv preprint arXiv:2201.12091*.
- Ren, J.; Guo, Q.; Yan, H.; Liu, D.; Zhang, Q.; Qiu, X.; and Lin, D. 2024. Identifying semantic induction heads to understand in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 6916–6932.
- Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15504–15522.
- Saravia, E.; Liu, H.-C. T.; Huang, Y.-H.; Wu, J.; and Chen, Y.-S. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 3687–3697.
- Stolfo, A.; Balachandran, V.; Yousefi, S.; Horvitz, E.; and Nushi, B. 2025. Improving instruction-following in language models through activation steering. In *International Conference on Learning Representations*, volume 2025, 55790–55823.
- Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Todd, E.; Li, M.; Sen Sharma, A.; Mueller, A.; Wallace, B.; and Bau, D. 2024. Function vectors in large language models. In *International conference on learning representations*, volume 2024, 17282–17333.
- Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.
- Venkatesh, S.; and Kurupath, A. M. 2026. On the non-identifiability of steering vectors in large language models. *arXiv preprint arXiv:2602.06801*.

- Walsh, E. P.; Soldaini, L.; Groeneveld, D.; Lo, K.; Arora, S.; Bhagia, A.; Gu, Y.; Huang, S.; Jordan, M.; Lambert, N.; et al. 2025. 2 OLMo 2 Furious (COLM’s Version). In *Second Conference on Language Modeling*.
- Wang, S.; Zhu, Y.; Liu, H.; Zheng, Z.; Chen, C.; and Li, J. 2024. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3): 1–37.
- Wu, Z.; Arora, A.; Geiger, A.; Wang, Z.; Huang, J.; Jurafsky, D.; Manning, C. D.; and Potts, C. 2025. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. In *International Conference on Machine Learning*, 67035–67080.
- Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Yin, K.; and Steinhardt, J. 2025. Which Attention Heads Matter for In-Context Learning? In *International Conference on Machine Learning*, 72428–72461.
- Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.
- Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.
- Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Appendix

A Formal Definitions

Coherence gate, Alg. 1 line 16. A steered generation g passes the gate iff $\text{PPL}(g) \leq \tau_M$, with the model-relative threshold

$$\tau_M = \text{clip}(5 \cdot \text{PPL}_M^0, 8, 40), \quad (\text{A.1})$$

where $\text{PPL}(g)$ is the perplexity of g under the model, clip bounds its argument to $[8, 40]$, and PPL_M^0 is model M 's baseline perplexity. The multiplier 5 and the clip bounds were calibrated once on pilot generations and then held fixed across all models and concepts.

Judge and swing, Alg. 1 line 20. A generation is scored not by a hard label but by an expected rating. The judge is asked to rate the answer a against the concept question q on an integer scale from 0 to 10, and the rating is read off the single-token distribution rather than from the sampled token,

$$J(a, q) = \sum_r r p(r | a, q), \quad p(r | a, q) = \frac{\exp \lambda_r}{\sum_{r'} \exp \lambda_{r'}}, \quad (\text{A.2})$$

where r ranges over the integers $\{0, \dots, 10\}$ and the logits λ_r are renormalized over those eleven tokens alone. Reading the expectation makes the score continuous, so a direction that shifts behavior slightly is not rounded away.

The swing of a direction is the judged difference between its two signs at the same magnitude,

$$\Delta(m) = \frac{1}{|S|} \sum_{s \in S} [J(g_{+m}(s), q) - J(g_{-m}(s), q)], \quad (\text{A.3})$$

with S the evaluation prompt set, m the multiplier grid index, and $g_{\pm m}(s)$ the generations from prompt s under $\pm m$.

Budget-normalized steerability.

$$\text{St}_c = \frac{\Delta_c}{2 b_c \|\text{resid}_{\ell^*}\|}, \quad (\text{A.4})$$

where Δ_c is Eq. A.3 for concept c at $m=1$, b_c the calibrated coherent budget fraction, $\|\text{resid}_{\ell^*}\|$ the mean residual-stream norm at the steering layer, and the factor 2 the span from $-m$ to $+m$.

$$z = \frac{\Delta_{\text{dir}}(m) - \overline{\Delta_{\text{rand}}(m)}}{\text{sd}_k \Delta_{u_k}(m)}. \quad (\text{A.5})$$

Here $u_1, \dots, u_K$ are unit directions drawn uniformly from the sphere in $\mathbb{R}^n$ at the same m , and sd_k is taken over them.

B Protocol Details

Multi-family adjudication protocol. These are the settings behind §3.4, used for the five-family adjudication §4.6 and the capacity ladders §4.7. The case studies and single-model races in §4.2-§4.5 predate this protocol and use fixed per-experiment settings. For budget calibration the steering magnitude starts at 0.15 of the mean residual-stream norm at the chosen layer and is scaled until the largest multiplier on the

Concept	Type	$B_{d'}$	B_{probe}	Y self	Y cross
anger	mode	2.82	0.94	+2.61	+3.00
emotion	mode	2.24	0.84	+3.28	+2.49
fear	mode	2.83	0.92	+6.21	+4.45
formality	mode	3.59	0.96	+2.46	+2.07
sentiment	mode	2.30	0.87	+2.02	+2.68
sentiment_yelp	mode	2.06	0.81	+1.10	+3.35
subjectivity	mode	2.02	0.83	−0.21	−0.07
toxicity	mode	2.32	0.89	−3.73	−0.79
animal	topic	4.57	0.97	+0.02	+0.45
athlete	topic	3.18	0.93	+1.43	+1.97
building	topic	3.37	0.92	+0.11	+0.16
business	topic	2.65	0.92	+0.66	+1.03
company	topic	3.85	0.97	−0.18	+0.82
film	topic	5.15	0.98	+0.20	+0.39
plant	topic	3.54	0.93	+0.76	+2.45
scitech	topic	2.55	0.87	−0.47	+0.98
sports	topic	3.97	0.96	+0.83	−0.65
world	topic	2.46	0.90	+0.18	+0.40

Table B.1: Per-concept zoo: separability versus the functional swing under both judges. The classes separate on the mean, not as a clean partition.

grid $\{-4, -2, 0, +2, +4\}$ sits just inside the coherence gate of Eq. A.1. The final fraction is confined to $[0.05, 0.45]$. For layer selection the separability profile of Algorithm 1 is computed at every layer, and its candidate set $\mathcal{C}$ keeps the four highest- d' layers subject to a minimum spacing of $\max(2, \lfloor n_{\text{layers}}/12 \rfloor)$, so $k=4$; the steering layer is the candidate whose probe-swing at that budget, the amount by which steering moves the probe, is largest. Because tokenizers differ in granularity, generation length and the pooling window are rescaled by each model’s median tokens-per-word ratio. Decoding is greedy throughout; the random null comprises five norm-matched directions per concept, each drawn independently across four seeds (20 direction $\times$ seed samples in total).

Setting	Δ	Null $\mu \pm \sigma$	z
<i>In-domain</i>			
Threat (Q+Q)	+3.83	$+0.13 \pm 0.54$	6.9
Threat (Q+L)	+2.99	$+0.03 \pm 0.31$	9.7
Threat (L+Q)	+3.68	-0.34 ± 1.17	3.4
<i>Unseen entities, zero-shot</i>			
Threat A (Q+Q)	+4.92	-0.01 ± 0.41	12.0
Threat B (Q+Q)	+3.08	-0.05 ± 0.63	5.0
Threat A (Q+L)	+4.51	$+0.05 \pm 0.65$	6.9
Threat B (Q+L)	+4.12	-0.07 ± 0.64	6.5
Threat A (L+Q)	+2.54	$+0.03 \pm 0.42$	6.0
Threat B (L+Q)	+3.50	-0.25 ± 0.14	26.9
Opportunity A (Q+Q)	+4.48	-0.40 ± 1.46	3.3
Opportunity B (Q+Q)	+4.71	-0.21 ± 1.20	4.1
<i>Off-domain</i>			
Threat (Q+Q)	+0.00	-0.00 ± 0.00	0.0
Opportunity (Q+Q)	+1.93	$+0.04 \pm 0.54$	3.5

Table B.2: Frame steering and transfer. Δ is the judged swing at fixed ± 2 , scored against a norm-matched random-direction null, and z is that swing above the null. Each row names the steered model and the judge in that order, with Q for Qwen3-4B-Instruct-2507 and L for Llama-3.1-8B, so Q+Q is a self-judge and Q+L a cross-family judge.

Acknowledgements

I. Research Background and Origin of the Topic

This work lies in the interpretability and representation analysis of large language models. Linear probes can decode high-level concepts from the residual stream with near-perfect accuracy, yet a decodable direction does not always produce a controllable change in behavior during generation. Prior work attributes this phenomenon to properties of the concept alone and cannot explain why the same concept yields opposite verdicts under different extraction pipelines. This paper argues that the verdict is set jointly by the concept’s representational structure and the extraction pipeline, and distinguishes two functional poles within a single concept’s representation, a readout direction and a control variable direction.

The topic was proposed independently by the student. In the first half of 2026, while surveying the literature on interpretability and representation analysis, the student sought to reproduce the relevant methods locally and to extend them. During reproduction the student found that the same concept behaved inconsistently under separability-gated selection and under unselected difference-of-means extraction, and the research question of this paper was determined from that observation. The choice of topic, the formulation of the problem and the technical approach were all completed by the student. The supervisor did not take part in the choice of topic.

II. Relationship Between the Supervisor and the Student

The supervisor, Zheng Lin, is a postdoctoral researcher and assistant researcher in the Department of Computer Science and Technology at Tsinghua University, holding a doctoral degree in computer science and technology. The student came to know him through an introduction by a third party in the first half of 2026, and no prior teacher-student or other relationship existed between them. The supervision was entirely unpaid. Neither the student nor the student’s family provided the supervisor with any fee or other benefit in any form.

The supervisor’s guidance was concentrated at the level of paper writing. It included suggestions on the organization of the paper, comments on the logic of the argument and the clarity of the exposition, suggestions on the presentation of figures and tables, and reminders concerning academic writing conventions. In addition, the supervisor offered suggestions on a few individual experiments and on data analysis.

III. Assistance from Others

Apart from the guidance on paper writing described above, the main body of this work received no assistance from any other person. There is no instance of another person carrying out the experiments, writing the code, processing the data or writing the paper on the student’s behalf. The main technical content and the conclusions of this paper were completed by the student independently.

IV. Division of Work Throughout the Research

This is an individual entry. Apart from the guidance on paper writing, every stage below was completed independently by the student, Ruixiu Zhang.

Topic and method design. The student independently proposed the research question, proposed the distinction between readout directions and control variable directions, and designed the structure-adaptive measurement protocol and the type-balanced concept bank.

Data acquisition. The public classification corpora used in the experiments were obtained through a HuggingFace mirror site and cover public datasets for sentiment, subjectivity, formality and topic classification. The student filtered, cleaned and combined them into a single labelled dataset. The reference texts of the type-balanced concept bank were generated once by Qwen3-4B-Instruct-2507 under greedy decoding and shared across all model families. All data are publicly available, contain no personal information and involve no human or animal subjects, so no ethics review was required.

Code implementation and execution of experiments. The experimental framework and all of its modules were designed and implemented by the student, including direction extraction, the separability gate, attention source-token ablation, continuous judging and the random-direction null. The models used for judging are the open-source families themselves and form part of the measurement protocol of this paper, as described in the main text. The coherence gate threshold is computed from each model’s own baseline perplexity by a fixed formula, and the intervention budget is determined adaptively from each model’s residual-stream norm. Both are deterministic rules. For parameter calibration the student used Claude models to help write scripts that sweep intervention strength and perplexity thresholds in batch. The evaluation dimensions and acceptance criteria were set by the student. Once fixed, the parameters follow the deterministic formula given in the appendix, and the data entering the final statistical analysis are filtered by that deterministic gate alone. The student additionally engaged Claude as a third party outside the protocol to assess independently whether the steered generations remained usable, as corroboration of the student’s own judgement. That assessment produces no reported value and was not used to filter the data entering the final statistics.

Data analysis and interpretation of results. All statistical analysis, significance testing and interpretation of the experimental results were carried out by the student.

Experimental environment. All experiments were run on the student’s own equipment, using a single NVIDIA RTX PRO 6000 Blackwell GPU. This work does not rely on any professional research institution or laboratory.

Paper writing. The English first draft was written by the student. The supervisor read it and offered comments at the level of writing, and the student revised and finalized the paper accordingly.

V. Difficulties Encountered and How They Were Resolved

The first difficulty was that the same concept yielded opposite verdicts under the two extraction pipelines, and it was initially hard to tell whether this was an implementation error or a real phenomenon. The student narrowed the only difference between the two pipelines down to the separability gate, repeated the comparison on the same data at the same layer, confirmed that the reversal comes from the extraction method itself rather than from estimator noise, and then used source-token ablation to show that the two directions draw on different information sources.

The second difficulty was that generation degrades at larger intervention strengths, which makes it hard to separate an ineffective direction from a collapsed generation. The student introduced a perplexity-based coherence gate together with a model-adaptive strength calibration, keeping each concept’s intervention budget just inside the gate so that comparisons between directions rest on the same perturbation budget.

The third difficulty was that automatic judging can introduce bias, and a verdict is not trustworthy when the judging model comes from the same family as the steered model. The student addressed this by adding cross-family judging and by using a deterministic verifier on the constraint satisfaction task, so that the core conclusions hold even with no model judging in the loop.

VI. Use of AI

This work used Anthropic’s Claude models, all invoked through the Claude Code command line tool. The versions used are Claude Opus 4.6, Claude Opus 4.8, Claude Opus 5 and Claude Fable 5. The change of version follows the iteration of the model offered by the tool, and the manner and purpose of use were the same across versions. The period of use was June to September 2026, comprising 11 sessions as counted by Claude Code session records.

The use was limited to the following four items.

1. LaTeX formatting, including float placement, table column widths, and equation and bibliography formatting.
2. Polishing of English wording, that is, revising grammar and expression in an English draft already written by the student, without changing any technical content.
3. Helping to write scripts for the visualizing stage, and adjusting the figure style until it matched the overall style of the paper.
4. Acting as a third-party assessor outside the protocol, reading the steered generations and assessing their readability and usability. Steered text is noticeably shifted relative to the original distribution, and this assessment served to confirm independently that such text remained legible and worth analysing. The assessment is independent of the measurement protocol of this paper. Its output is not the source of any reported metric and was not used to filter the data entering the final statistics.

The division of work for items three and four is detailed in Section IV. Note that the models used to score generations within the measurement protocol of this paper are the open-source model families themselves. They form part of the research method, are described in full in the main text, and are independent of the AI assistance described above.

AI did not take part in the following, namely the choice of topic and the formulation of the research question, the design of the core method, the statistical analysis and significance testing, the interpretation of the experimental results, and the drawing of conclusions. All research findings and conclusions of this paper were reached by the student independently.

No AI tool other than the Claude models listed above was used in this work. Session records for each of those versions are submitted separately with the competition materials.

VII. Profiles

Ruixiu Zhang (Student): A high school student at the International Division of Beijing Normal University Experimental School, Beijing. His research interest is in the interpretability of large language models, in particular representation analysis and activation steering. He founded and leads Liandan She, a student-run AI research club at his school. He proposed the research question of this paper and designed, implemented and ran all of its experiments on self-hosted hardware.

Zheng Lin (Supervisor): A postdoctoral researcher and assistant researcher in the Department of Computer Science and Technology, Tsinghua University, under the supervision of Prof. Shi-Min Hu. He received his doctoral degree from Nankai University under the supervision of Prof. Ming-Ming Cheng. His research interests are in computer vision and graphics tasks involving human interaction. He advised on the writing of this paper on an individual and unpaid basis.