

IdioSteer: Controlling Figurative-vs-Literal Interpretation via Residual-Stream Activation Steering

Bar Cohen^{1,2}, Kai Golan Hashiloni^{1,2},
Kfir Bar^{1,2}

¹Data Science Institute, Reichman University, Herzliya, Israel

²Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel

bar41234@gmail.com, kai.golanhashiloni@post.runi.ac.il,

kfir.bar@runi.ac.il

Abstract

Large language models (LLMs) default to figurative interpretations of idiomatic expressions when context supports them. We investigate whether this bias can be directly controlled via activation steering: asking whether idiomaticity is encoded as a manipulable direction in residual space. We introduce IdioSteer, a small controlled benchmark of 100 sentences across 20 idioms designed to expose interpretation-dependent generation. Using a steering vector applied to five LLMs from two model families, ranging from 2B to 12B parameters, we show that steering systematically shifts model outputs toward literal interpretations while preserving fluency. Three findings stand out: cross-layer steering—vectors from mid-to-late layers injected at early layers—consistently outperforms same-layer interventions; steering vectors transfer to held-out idioms; and effects replicate across architectures and model sizes using the same intervention protocol.

1 Introduction

Natural language frequently exhibits non-compositional semantic constructions, where the meaning of an expression cannot be derived directly from the meanings of its individual components. A central instance of this phenomenon is *idiomatic expressions*, whose interpretation depends on convention and context rather than compositional semantics. For example, the phrase “*bad apple*” can denote either a literal defective fruit or, more commonly, a figurative reference to a corrupt or problematic individual. Resolving such ambiguity requires contextual reasoning, making idioms a core challenge in semantic modeling.

Idioms also vary in the extent to which their constituent words contribute to their figurative meaning, the degree of their compositionality (Nunberg et al., 1994; Gibbs et al., 1989a,b). Their interpretation can therefore depend not only on an expression’s lexical content, but also on conceptual,

Prefix: During the party, Sarah tried to *break the ice*

Unsteered: During the party, Sarah tried to *break the ice* with the other guests by talking about her job.

Steered: During the party, Sarah tried to *break the ice* into pieces so that the ice cubes would melt.

Figure 1: Illustration of the steering effect. Continuations of a prefix without steering (top) follow the conventional figurative reading of *break the ice*, whereas continuations under steering toward literal interpretation (bottom) take its literal reading. Examples for all five models we test are in Table 1 (Appendix A).

cultural, and contextual knowledge (Pollio et al., 1977; Weinreich, 1969; Jackendoff, 1997; Gibbs and Colston, 2012). This tension between an expression’s literal form and its conventional meaning makes idioms a useful setting for investigating how models represent and control non-compositional meaning.

Recent work has shown that large language models (LLMs) capture aspects of non-compositional meaning with increasing robustness. In particular, Hashiloni et al. (2025) and Hefetz et al. (2025) demonstrate that LLMs can reliably identify idiomatic expressions and distinguish them from literal usages across contexts, including in cross-lingual settings. Earlier work shows that while LLMs achieve strong performance on idiom detection, they still exhibit systematic biases and limitations. For example, De Luca Fornaciari et al. (2024) introduce a challenging benchmark for idiom detection and show that performance degrades under more complex contextual conditions, while Phelps et al. (2024) find that even the largest LLMs fall short of fine-tuned models on idiomaticity detection. Similarly, Khoshtab et al. (2025) highlight variability in idiom understanding across languages and figurative constructions. Together, these findings indicate that LLMs encode internal represen-

tations that support idiomaticity, but that these representations also reflect strong default preferences toward figurative interpretation.

A complementary line of work studies how internal representations of high-level model behavior can be *controlled* at inference time through residual-stream activation steering (Turner et al., 2024; Rinsky et al., 2024; Arditi et al., 2024). By identifying a vector in residual space and injecting it during generation, these methods causally bias model behavior in semantically meaningful ways. Most prior work in this area targets relatively low-level properties (sentiment, refusal, factual associations); the application of activation steering to higher-order semantic phenomena such as figurative-vs-literal interpretation has received comparatively little attention.

We investigate whether idiomaticity can be captured as a dense vector that directly controls the figurative-vs-literal balance at inference time. Using a white-box steering approach, we construct an idiomaticity vector from training-time activations and inject it during generation. Without intervention, base models default to figurative continuations; with steering, outputs shift systematically toward literal interpretations. Evaluated via human annotations and LLM-as-a-judge across models of varying architectures and sizes, our results indicate that non-compositional semantic behavior can be captured and manipulated through vector-based interventions in representation space. Figure 1 illustrates this effect qualitatively: the unsteered model follows the conventional figurative reading of *break the ice*, while steering in the literal direction elicits a literal continuation about ice as a physical object. See more examples in Appendix A.

We make the following contributions:

1. We introduce **IdioSteer**, a controlled benchmark of 100 sentence prefixes across 20 English idioms, each ending in an idiomatic expression that admits both figurative and literal continuations.
2. We show that activation steering reliably shifts LLM continuations toward literal interpretation across five base models from two architectural families, ranging from 2B to 12B parameters, while maintaining generation coherence. We provide a comparative analysis demonstrating consistent steering effects across architectures and sizes.
3. We analyze source-layer and injection-layer choices, and identify a cross-layer regime—

constructing the vector at a mid-to-late layer and injecting it at the earliest layers—that substantially outperforms same-layer steering.

4. We show that the steering vector is idiom-general rather than tied to the idioms it was built from: vectors derived from disjoint idiom subsets are geometrically aligned and behaviorally equivalent.

To guide our investigation, we address the following research questions:

RQ1. Can residual-stream activation steering reliably control whether an LLM produces a figurative or literal continuation of an idiomatic expression, across models of varying architecture and scale?

RQ2. Does the resulting steering vector reflect a directional intervention rather than generic activation perturbation?

RQ3. What layer-choice regime makes steering effective?

RQ4. Does the steering vector generalize across idioms, and is idiomaticity what it encodes?

The IdioSteer dataset, construction guidelines, annotation guidelines, and evaluation resources are publicly released to ensure transparency and reproducibility under the CC-BY-4.0 (dataset) and Apache-2.0 (code) licenses: <https://github.com/Intellexus-DSI/IdioSteer>.

2 Related Work

2.1 Idioms in LLMs

Idiomatic expressions are a central case of non-compositional semantics, where meaning cannot be derived directly from the constituent words (Baldwin and Kim, 2010; Constant et al., 2017). Prior work has largely studied idioms through *identification and classification*, that is, determining whether an expression is used figuratively or literally in context. Datasets such as MAGPIE (Haagsma et al., 2020), MIDAS (Kim et al., 2025), and related benchmarks (Mi et al., 2025) have enabled systematic evaluation of this ability. Recent studies show that prompted LLMs can achieve competitive performance on standard idiom detection benchmarks (Hashiloni et al., 2025; Arslan et al., 2025), though performance remains sensitive to contextual complexity and language variation (Phelps et al., 2024; De Luca Fornaciari et al., 2024; Khoshtab et al., 2025; Almheiri et al., 2026; Ye et al., 2026; Hashiloni et al., 2026b). A complementary line of work examines how idiomaticity is represented internally in LLMs. Hefetz et al. (2025) use probing

and Logit Lens (nostalgebraist, 2020) to show that idiom-vs-literal information is partially recoverable from intermediate layers; Oh et al. (2026) use causal tracing to characterize a three-stage mechanism in which early layers retrieve figurative interpretations while parallel pathways carry the literal compositional reading forward.

Most relevantly, Ghosh and Jiang (2026) report that LLMs compress concreteness into a one-dimensional direction in mid-to-late layers and use this direction to bias generation toward more literal or figurative text. Our work shares this motivation but differs in focus: we target interpretation selection rather than generation style, operating at the level of idiomatic expressions, multi-word units whose meaning is not derivable from their components, where the task is to resolve a phrasal ambiguity rather than a lexical one. Additionally, we identify a cross-layer steering regime, separating the layer at which the direction is constructed from the layer at which it is injected, which same-layer approaches do not exploit.

2.2 Activation Steering at Inference Time

Activation steering controls model behavior by intervening on the residual stream of a frozen language model during inference. Turner et al. (2024) introduce activation addition, deriving steering directions from activation differences between contrasting prompts; Rinsky et al. (2024) formalize this as Contrastive Activation Addition (CAA), using differences between class-conditional activation means—the construction we follow in this work for our main steering vector. Some high-level behaviors are mediated by low-dimensional directions in residual space, including a refusal direction (Arditi et al., 2024) and a sentiment direction (Tigges et al., 2024); we compare against a probe-derived direction following this line of work, and evaluate an angular operator inspired by Vu and Nguyen (2025), which steers by rotating activations rather than adding a scaled vector.

Most prior steering work targets surface-level or behaviorally coarse properties (e.g., sentiment, refusal, truthfulness, writing style, factual associations). We extend this line of research to a higher-order linguistic-semantic phenomenon: the figurative-vs-literal interpretation of idiomatic expressions, asking whether activation steering can control non-compositional meaning.

3 IdioSteer Benchmark

To study inference-time control over figurative-vs-literal interpretation, we introduce **IdioSteer**, a controlled benchmark of 100 sentence prefixes spanning 20 English idioms (five variants per idiom) drawn from MAGPIE (Haagsma et al., 2020) and validated by two independent annotators (Appendix B).

Design. Each sentence is constructed so that the target idiom appears as the final tokens of the sentence—with no post-idiom context—and the pre-idiom context is deliberately written to leave the figurative-vs-literal ambiguity unresolved. For example, “During the party, Sarah tried to *break the ice*” admits both a figurative continuation (e.g., ... *by talking to the man next to her*) and a literal one (e.g., ... *into pieces so that the ice cubes would melt*). The two readings need not be equally likely; the criterion is that both continuations are natural. This bidirectional plausibility ensures that any failure to produce a target interpretation reflects a property of the model or the steering method, not a benchmark artifact in which one direction was never available.

Construction and validation. We use Claude 3.5 Sonnet (Anthropic, 2024) only to propose candidate sentences during drafting. All final examples are curated, edited, and validated by human annotators. Two independent human annotators then validate each sentence under a dual-completion protocol: both annotators produce a figurative and a literal continuation; a sentence is *valid* only if both continuations are natural. Sentences flagged by either annotator are revised and re-circulated until all 100 are *valid*. The 20 IdioSteer idioms are entirely distinct from those used to estimate steering vectors. Additional details on annotators, construction constraints, validation, and iterative refinement are provided in Appendix B.

4 Steering Methodology

We use residual stream activation steering (Turner et al., 2024; Rinsky et al., 2024) to bias model continuations toward a literal or figurative reading. Our pipeline has three components: (i) the *steering vector*, a unit-norm direction in residual space; (ii) the *steering operator*, combining the vector with an activation; and (iii) the *application protocol*, specifying where and when the operator fires.

4.1 Steering Vector Construction

We compare three constructions of a steering vector $v \in \mathbb{R}^d$, where d is the residual-stream dimensionality (e.g., $d = 3,072$ for Llama-3.2-3B). All vectors are unit-normalized, so intervention magnitude is controlled entirely by the scaling coefficient α (Section 4.2) and is comparable across vectors and models. Each vector is computed once from a fixed training pool and reused at inference.

Activation collection. All three vectors are derived from *residual-stream activations*: the per-layer, per-token hidden states that pass between transformer blocks. For each training sentence, we run a forward pass through the frozen model and record the residual-stream activation at the final token of the idiomatic expression, at a designated source layer L_s . Note that L_s is distinct from the layer at which the vector is later injected during steering (Section 4.3). Each recorded activation is a vector in $\mathbb{R}^d$.

Training pool. We construct steering vectors from IdioLink (Hashiloni et al., 2026a), a benchmark of sentences spanning both figurative and literal uses of English idioms. We use the idiom-disjoint train split, retaining 1,683 sentences across 22 idioms after filtering for surface-form idiom presence (~ 76 sentences per idiom; 629 figurative, 1,054 literal). Each sentence is annotated as either *figurative* (the idiom carries its conventional non-compositional meaning) or *literal* (the component words are used compositionally outside the idiomatic frame). The idiomatic expression is explicitly marked in each sentence, making it straightforward to identify its constituent tokens. We denote the resulting activation sets at the source layer L_s as H^{fig} and H^{lit} . The class distribution is moderately imbalanced ($\sim 37\%$ figurative / $\sim 63\%$ literal). Additional details are provided in Appendix C.

We experiment with the three vector construction methods, which we describe in the rest of this section: Mean-difference (MD), a contrastive average between class centroids; Probe (PROBE), the weight vector of a learned classifier; and Random (RAND), an untrained perturbation control.

Mean-difference (MD). Following the standard contrastive-activation construction (Rimsky et al., 2024; Arditì et al., 2024), we compute the difference between the class-conditional means of the figurative and literal activation sets, and unit-normalize. Let $c \in \{\text{fig}, \text{lit}\}$ index the two classes, and let $\bar{h}^c$ denote the empirical mean activation of

class c :

$$\bar{h}^c = \frac{1}{|H^c|} \sum_{h \in H^c} h, \quad v_{\text{MD}} = \frac{\bar{h}^{\text{fig}} - \bar{h}^{\text{lit}}}{\|\bar{h}^{\text{fig}} - \bar{h}^{\text{lit}}\|}. \quad (1)$$

By convention, positive scaling along v_{MD} (the steering vector) pushes activations toward the figurative class, negative scaling toward the literal class.

Probe (PROBE). Following Tigges et al. (2024), who define the steering vectors as the unit-normalized weight vector of a logistic regression classifier, we train a binary logistic classifier:

$$p(c = \text{fig} \mid h) = \sigma(w^\top h + b), \quad (2)$$

where $w \in \mathbb{R}^d$ is the classifier weight vector, $b \in \mathbb{R}$ is the bias, $w^\top h$ denotes the dot product, and σ is the sigmoid function. The classifier is trained on the labeled activation set $H^{\text{fig}} \cup H^{\text{lit}}$, with figurative activations labeled 1 and literal activations labeled 0. The steering vector is the unit-normalized weight vector: $v_{\text{PROBE}} = w / \|w\|$.

Like v_{MD} , v_{PROBE} orients toward the figurative class by construction. Probe-derived directions in over-parameterized regimes are known to be non-identifiable (Gao et al., 2026; Venkatesh and Kurupath, 2026); Appendix D reports per-model classification accuracy and provides additional analysis of the probe-derived steering vectors.

Random (RAND). As a control for the possibility that any sufficiently large activation perturbation could shift the figurative-literal balance regardless of direction, we additionally evaluate $K_{\text{rand}} = 3$ random unit vectors, drawn uniformly from the unit hypersphere. Each random vector is applied under the same operator and scaling protocols as the learned vectors. RAND establishes the effect of generic activation perturbation independent of any learned direction.

4.2 Steering Operators

Given a steering vector v and an activation $h \in \mathbb{R}^d$ at the target position, an operator returns a modified activation $h' = \mathcal{O}(h, v; \cdot)$. We evaluate two operators.

Additive ($\mathcal{O}_+$). Following Turner et al. (2024):

$$\mathcal{O}_+(h, v; \alpha) = h + \alpha v, \quad \alpha \in \mathbb{R}. \quad (3)$$

Negative α pushes h along $-v$ (toward literal); positive α along $+v$ (toward figurative in our case). The additive operator changes both the direction and the norm of h .

Angular ($\mathcal{O}_\theta$). We also evaluate a norm-preserving angular operator inspired by Vu and

Nguyen (2025). Instead of adding a scaled vector, angular steering rotates h within the input-dependent plane $\text{span}\{h, v\}$ by an angle $\theta \in [-\pi, \pi]$, where h is the activation at the intervention position and v is the unit-norm steering vector. The rotation controls intervention strength while preserving $\|h\|$. Since v is oriented toward the figurative class, negative θ rotates h toward v (figurative); positive θ rotates away (literal). Full details are given in Appendix E.

4.3 Application Protocol

We apply the steering operator at one input-token position p and one layer L_i , which we call the *injection layer*. The operator is implemented via a forward hook on the residual-stream output of layer L_i .

Token position. We apply the steering operator at a single token position p : the final token of the idiomatic expression in the input sentence. This choice treats the idiom as a completed span, since under causal attention, the final token can incorporate information from the preceding context and earlier idiom tokens. We use this fixed position in all experiments; Appendix F discusses the rationale and alternative intervention sites considered during development.

Cross-layer construction. A central design choice is the separation of the *source layer* L_s , at which the steering vector is constructed (§4.1), from the *injection layer* L_i , at which the operator fires. We call $L_s = L_i$ *same-layer* steering (the standard choice) and $L_s \neq L_i$ *cross-layer* steering. The motivation is that representation quality and intervention leverage need not peak at the same layer: a direction may be easiest to estimate where the distinction is linearly accessible, yet most effective when injected at earlier layers that can still shape the continuation. We ablate L_s and L_i in §6.3.

KV-cache semantics. The intervention fires only once, on the prompt’s forward pass at position p : the modified residual stream at L_i propagates upward, producing modified key-value caches that subsequent generated tokens attend to in the standard way. This single-position, prompt-pass-only protocol differs from CAA (Rimsky et al., 2024), which intervenes at every post-prompt position throughout generation; it is closer in spirit to single-position causal interventions used in mechanistic interpretability. Appendix G gives the implementation details.

Generation settings. Each sentence prefix is

passed verbatim to the model with no instruction or chat template. Completions are sampled at temperature 0.7, top- p 0.9, up to 50 new tokens, stopping at end-of-sequence or the token budget.

5 Experimental Settings

We evaluate five *base* LLMs from two architectural families, 2B–12B parameters: Llama-3.2-3B, Llama-3.1-8B, Gemma-2-2B, Gemma-3-4B, Gemma-3-12B. We use base models (prior to supervised fine-tuning) to isolate next-token generation behavior from instruction-following biases; instruction-tuned models are left to future work.

Steering grid. We evaluate two operators (additive, angular) crossed with two learned vector constructions (mean-difference, probe), plus three random unit-norm vectors as perturbation controls. A *condition* is one (vector, operator, coefficient) combination, or the unsteered baseline. The intervention is global and input-independent. For additive steering, we parameterize intervention strength as $\alpha = \alpha_{\text{factor}} \cdot s_{\text{MD}}$, where α is the raw additive coefficient, α_{factor} is a model-normalized multiplier, and $s_{\text{MD}} = \|\bar{h}^{\text{fig}} - \bar{h}^{\text{lit}}\|$ is the norm of the raw mean-difference vector (with $\bar{h}^{\text{fig}}$ and $\bar{h}^{\text{lit}}$ the empirical class-mean activations). For each model, calibration on IdioLink selects the source layer L_s (where the steering vector is constructed) and the α_{factor} grid. IdioSteer is not used during calibration. The injection layer L_i (where the operator fires) is fixed to $L_i = 2$. We call this per-model setting—the calibrated L_s with $L_i = 2$ and the additive operator—the *canonical* configuration, and use it wherever we do not sweep layers. Angular steering is norm-preserving, so we use a shared grid of rotation angles $\theta \in \{\pm 60^\circ, \pm 70^\circ, \pm 90^\circ, \pm 120^\circ\}$ across all models. The eight coefficient values swept under each operator form its *grid*: the model-specific α_{factor} values for additive steering, and the shared θ values for angular. We sample 5 completions per sentence, giving 500 completions per condition. Full calibration details are in Appendix H.

Evaluation. Each generation is labeled *figurative*, *literal*, or *incoherent*. We report *literal rate* (proportion of generations labeled literal; the directionally harder task since figurative is the conventional default) and *coherence rate* (proportion of generations that are labeled either figurative or literal), with 95% bootstrap confidence intervals (CIs) over evaluation idioms (Appendix I). We

scale evaluation with an LLM-as-a-judge (Zheng et al., 2023), calibrated against 280 Llama-3.2-3B generations labeled by two annotators (Cohen’s $\kappa = 0.867$). The selected judge, **Gemma-4-31B-it** (Google DeepMind, 2026), achieves 90.0% accuracy and $\kappa = 0.821$ against gold labels. Since those 280 items were used to select the judge, we keep them as a separate calibration tier and validate the chosen judge on 795 further human-annotated generations spanning all five models and all steering conditions (91.9% agreement, $\kappa = 0.868$; Appendix J.1). Annotation guidelines, candidate judge models, and prompt iteration are in Appendix J.

6 Results and Discussion

6.1 RQ1: Activation Steering Shifts Figurative-vs-Literal Interpretation Across LLMs

We evaluate whether residual-stream activation steering can reliably shift LLM continuations toward literal interpretation of idioms. For each of five models—Llama-3.2-3B, Llama-3.1-8B, Gemma-2-2B, Gemma-3-4B, Gemma-3-12B—we summarize each method by the highest *literal rate* it attains on IdioSteer within its fixed coefficient grid (Section 5). This best-over-grid statistic measures peak steering capacity without letting us tune the grid post hoc; full curves over the grid are reported in Appendix K. Figure 2 compares these best-over-grid operating points for each vector construction and steering operator.

Steering shifts the figurative-vs-literal balance across all five models. The unsteered baseline literal rate of 0.12–0.14 across models is itself evidence of the strong figurative default in idiom interpretation: even without intervention, the models overwhelmingly produce figurative continuations. All configurations—mean-difference and probe, under both additive and angular operators—produce a literal-rate increase over the unsteered baseline on every model, with one exception: probe-additive on the Gemma family shows no change from baseline. Literal rate increases across the board, though the magnitude varies widely across vector constructions and model families. Mean-difference additive steering is the dominant construction in four of five models, achieving a literal rate of 0.34–0.57 (a $2.4\text{--}4.5\times$ increase over the unsteered baseline); on Gemma-3-4B, the angular operator is marginally preferred (0.37 vs 0.36). Bootstrap confidence

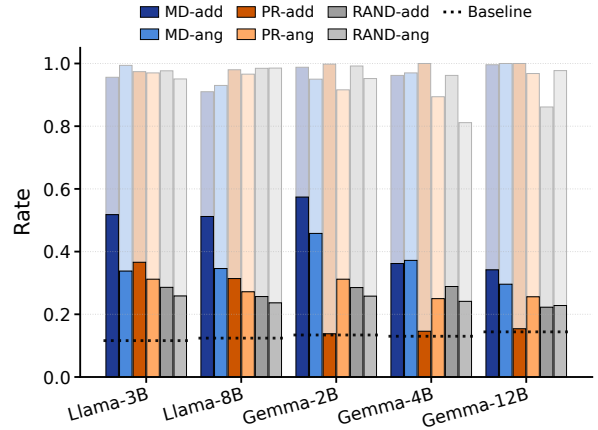


Figure 2: Maximum literal rate per (model, method) at the best operating point. Bars are stacked: solid (darker) = literal rate, lighter top = figurative-coherent rate; total height = coherence rate. Dotted line: unsteered baseline literal rate per model. Random results averaged over three unit-norm vectors. PR=PROBE; add=Additive; ang=Angular.

intervals for the headline operating points are reported in Appendix I. Probe-derived vectors exhibit a family-dependent pattern: PROBE-additive produces moderate effects on the Llama family (0.31–0.37) but collapses to near-baseline on the Gemma family (0.14–0.15 across all three Gemma models), consistent with the probe non-identifiability concern (§4.1, PROBE). PROBE-angular is more robust across families (0.25–0.31 on every model), suggesting that the rotation operator partially compensates for the probe vector’s weaker alignment with the figurative-vs-literal axis (Appendix E discusses how the two operators differ in their sensitivity to activation norm). Random unit-norm vectors reach best-over-grid literal rates of 0.22–0.29 across the five models under both operators, smaller than the strongest learned method on every model. The angular operator does not improve random-vector performance over additive; we characterize the difference between directional and perturbative effects in Section 6.2. A best-over-grid statistic can in principle reward a single fortunate coefficient. We therefore also summarize each method by its mean literal rate over the four literal-direction coefficients of its grid (its *half-grid mean* in Table 8, Appendix K). The ranking is unchanged—mean-difference-additive steering is highest on every model, and exceeds baseline at all four literal-direction coefficients. This shows that the reported effects reflect broad regions of the grid rather than isolated operating points.

The multiplier convention generalizes across

norms. Across the five models, the norm of the raw (pre-normalization) mean-difference vector $\|\bar{h}^{\text{fig}} - \bar{h}^{\text{lit}}\|$ varies by three orders of magnitude (from 4.18 on Llama-3.2-3B to 7,134 on Gemma-3-12B). Despite this, the calibrated multiplier grids $\alpha_{\text{factor}} = \alpha/s_{\text{MD}}$ cluster within a factor of ~ 2.5 across all models, with the optimal operating point falling in $\alpha_{\text{factor}} \in [-2.4, -0.5]$ on every model (see Appendix H for more details). This suggests that the multiplier convention identifies a useful per-norm scale of intervention.

Summary (RQ1). Activation steering reliably shifts continuations toward literal interpretation across two model families with sizes ranging from 2B to 12B. Mean-difference additive steering is the most consistent construction, achieving roughly $2.4\text{--}4.5\times$ over baseline on every model. Optimal intervention strength is consistent across models in multiplier-normalized units; exceeding it trades literal-rate gains for fluency loss (Turner et al., 2024; Rimsy et al., 2024). Coherence at reported operating points remains ≥ 0.91 , indicating that literal-rate gains are achieved without sacrificing generation quality.

6.2 RQ2: The Steering Effect is Directional, not Perturbative

Whether the literal-rate shift reflects a genuine directional intervention or simply the destabilizing effect of a large activation perturbation is addressed with two diagnostic comparisons.

Sign-asymmetric literal rate. Figure 3 overlays the literal rate of mean-difference additive steering as a function of α_{factor} for all five models, with the random-vector average shown as a dashed control line. Mean-difference curves are strongly asymmetric: pushing in the literal direction (negative α_{factor}) produces large literal-rate shifts on every model, while pushing in the figurative direction leaves the literal rate at or below baseline. At the most extreme magnitudes tested, the negative-side literal rate is $3\text{--}4.3\times$ the positive-side rate on four of the five models, reflecting the model’s strong figurative default: there is room to shift toward literal but no room to shift further toward figurative. Random unit-norm vectors at matched magnitudes produce near-symmetric literal rates at the extreme magnitudes tested (0.18–0.29 across the five models on either sign) with no consistent direction, lacking the directional structure of mean-difference steering. This asymmetric mean-difference vs near-symmetric random pattern is visible across all five

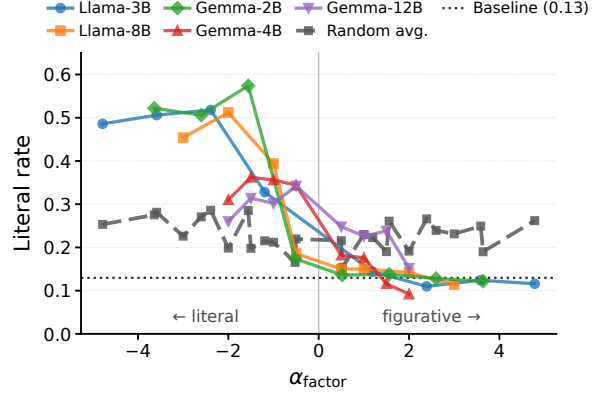


Figure 3: Literal rate as a function of α_{factor} for mean-difference steering, overlaid across all five models. Each curve is a separate model; the dashed grey line is the random-vector average. The unsteered baseline is shown as a dotted horizontal line. Strong asymmetry in α_{factor} sign is consistent across all five models, while random vectors show only mild magnitude-dependent shifts.

models in the per-model literal-rate curves in Appendix K.

Perplexity dissociation. A directional intervention should affect generation quality differently from a perturbative one. We measure per-token perplexity of steered generations under the unsteered model (Appendix L). On Llama-3.2-3B, the perplexity ratio $\text{PPL}^{\text{steered}}/\text{PPL}^{\text{baseline}}$ is roughly symmetric in α_{factor} sign (1.27 at -4.78 , 1.14 at $+4.78$), while the literal-rate effect is strongly asymmetric. Perplexity thus tracks intervention *magnitude*; literal rate tracks its *direction*. Random vectors show the same perplexity profile without producing comparable literal-rate asymmetry (Appendix L).

Summary (RQ2). The mean-difference steering effect is directional, not perturbative. Sign-asymmetric literal rate combined with sign-symmetric perplexity argues against a purely perturbative explanation in which any sufficiently large activation modification produces equivalent results; the asymmetry is consistent across all five models.

6.3 RQ3: Cross-layer Steering Substantially Outperforms Same-layer Steering

The application protocol of Section 4.3 separates the source layer L_s (where the steering vector is constructed) from the injection layer L_i (where it is injected), with $L_s \neq L_i$ in our canonical setup. We ablate this choice with three layer-pair sweeps on three models spanning two families: Llama-3.2-3B, Llama-3.1-8B, and Gemma-2-2B. For each

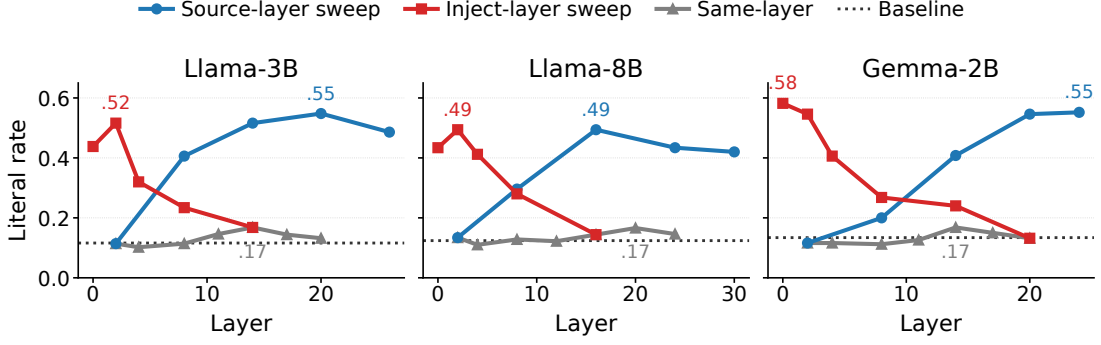


Figure 4: Layer-pair sweeps on three models at $\alpha_{\text{factor}} = -2$. Blue: source-layer sweep ($L_i = 2$ fixed); red: injection-layer sweep (canonical L_s fixed); grey: same-layer diagonal ($L_s = L_i$). Dotted line: unsteered baseline. Cross-layer steering outperforms same-layer steering by $3.0\text{--}3.5\times$.

model we perform a source-layer sweep (varying L_s at fixed $L_i = 2$), an injection-layer sweep (varying L_i at the model’s canonical L_s), and a same-layer diagonal sweep ($L_s = L_i$). All sweeps hold α_{factor} fixed across source layers so that the steering-vector norm is matched within each sweep. Figure 4 shows the results; per-model heatmaps are in Appendix M.

Cross-layer steering substantially outperforms same-layer steering. At the matched coefficient shown in Figure 4 ($\alpha_{\text{factor}} = -2$), the cross-layer regime—mid-to-late source layer paired with $L_i \leq 2$ —achieves the highest literal rates across all three swept models: 0.55 on Llama-3.2-3B, 0.49 on Llama-3.1-8B, and 0.58 on Gemma-2-2B. Same-layer configurations reach only ~ 0.17 on each model, giving a $3.0\text{--}3.5\times$ gap. The optimal injection layer is $L_i \in \{0, 2\}$ and the calibrated source layer falls in the mid-to-late range (50–80% of network depth) across all models.

Source-sweep shape. Displayed in Figure 4 as the blue curves, the literal rate rises steeply as the source layer advances through the early and middle network, peaking in the mid-to-late range. On the two Llama models, it then drops by 0.06–0.07 at the deepest swept layer.

A similar profile appears in how well the distinction can be recovered from the activations themselves. At each layer of Llama-3.2-3B and Gemma-2-2B we train a logistic probe to separate figurative from literal idiom usage and test it on idioms the probe has not seen; we call its accuracy the *decodability* of the distinction at that layer (Appendix O). Decodability starts at 0.63 at the embedding layer—the majority-class baseline set by the 63% literal share of the training pool—climbs sharply through the early layers, and plateaus at 0.91–0.95 across

the mid-to-late network, the region containing each model’s calibrated L_s , before declining at greater depths.

Mechanistic interpretation. We propose a two-fold explanation for this dynamic. First, the distinction is most decodable in mid-to-late layers, making this the ideal region to estimate an informative steering vector; Hefetz et al. (2025) similarly recover this distinction from intermediate representations. Second, injecting the vector is most effective at early layers because the intervention can shape all subsequent computation. This aligns with the causal-tracing results of Oh et al. (2026), who find that figurative interpretations are retrieved early on. Cross-layer steering would then succeed by estimating the vector where the distinction is clearest and injecting it where it exerts the most downstream influence.

Since α is scaled by the source-layer mean-difference norm rather than by the activation norm at the injection site, configurations sharing a multiplier do not deliver identical perturbations relative to the residual stream they modify; a targeted check confirms the advantage persists when the two are matched on that quantity (Appendix N).

Summary (RQ3). Cross-layer steering with L_s at mid-to-late depth and $L_i \in \{0, 2\}$ outperforms the best same-layer configuration by roughly $3\times$ on three models from two families. The regime is consistent across architectures, and the mid-to-late plateau in linear decodability is consistent with the location of the source-sweep optimum.

6.4 RQ4: The Steering Vector Is Idiom-General but Not Purely Idiomatic

The canonical setup already evaluates on idioms disjoint from training: we compute steering vectors

on 22 IdioLink idioms and apply them to the 20 disjoint idioms in IdioSteer. We strengthen this with a cross-fold robustness analysis on Llama-3.2-3B (Appendix P): we partition the 32 idioms from the IdioLink train and validation splits into three disjoint folds, re-derive mean-difference and probe vectors from each fold alone, and evaluate on IdioSteer at the canonical configuration ($L_s = 14$, $L_i = 2$, additive operator). Fold-derived vectors track the canonical 22-idiom vector closely, with a literal-rate spread of only 0.06 across folds. Fold-derived mean-difference vectors converge geometrically (mean pairwise cosine 0.74), while probe-derived vectors do not (pairwise cosine ~ 0.28), each nonetheless retaining substantial alignment (~ 0.59) with its own fold’s mean-difference vector—consistent with probe non-identifiability (Section 4.1).

The effect is also spread across idioms rather than driven by a few: on Llama-3.2-3B, pooling the two mean-difference coefficients spanning the literal-rate plateau (raw $\alpha = -10$ and -15), all 20 idioms exceed their unsteered baseline and 16 of 20 exceed a literal rate of 0.40; at the single headline operating point the corresponding counts are 20 and 15 of 20. Per-idiom rates for all five models are reported in Appendix Q.

A concreteness control. Literal continuations tend to describe more concrete situations than figurative ones, so the mean-difference vector could in principle capture *concreteness* rather than idiomatity. As a complementary ablation, we build a non-idiomatic concreteness vector with our pipeline unchanged, from 250 concrete vs. 250 abstract nouns (Brysbaert et al., 2014), each placed in the same neutral sentence templates so that the two classes differ only in the target word. We then steer with it on Llama-3.2-3B. Full implementation details and results are given in Appendix R.

The concreteness vector v_{conc} alone raises the literal rate from 0.12 to 0.43, compared with 0.52 for the full idiom vector. However, a residual effect remains: removing the component aligned with v_{conc} still yields a literal rate of 0.34. The two components also differ in which idioms they affect. Concreteness steering succeeds when the literal reading is a concrete scene, but fails on *out of the blue*, whose literal reading requires re-parsing the phrase rather than picturing a scene, and on *behind bars*, whose figurative reading is itself a concrete scene; the concreteness-removed component succeeds on both. Thus, the steering vector combines a

general concreteness component with an idiomaticity component that is not reducible to concreteness alone.

Summary (RQ4). The steering vector is idiom-general—recoverable from disjoint idiom subsets with equivalent behavioral effects—but not monolithic: it combines a general concreteness component, dominant for the average effect, with an idiom-specific component that alone controls interpretation where the two dissociate.

7 Conclusion

We have shown that the figurative-vs-literal interpretation of idiomatic expressions in LLMs is a property that admits inference-time control via residual-stream activation steering. Mean-difference steering reliably shifts continuations from the model’s default figurative reading toward literal interpretation across five base models from two architectural families, substantially increasing literal-rate over the unsteered baseline, with effects that are sign-asymmetric (ruling out generic activation perturbation), idiom-general (recoverable from disjoint training-idiom subsets), and consistent across architectures when intervention strength is normalized by the raw mean-difference norm. A complementary ablation shows the direction is not purely idiomatic: it mixes a general concreteness component, which dominates the average effect, with an idiom-specific one (Appendix R). The cross-layer protocol introduced here—constructing the vector at a mid-to-late source layer and injecting at the earliest layers—outperforms standard same-layer interventions by roughly $3\times$.

More broadly, our results highlight two diagnostics that may be useful for steering studies beyond idioms. First, separating the layer where a direction is estimated from the layer where it is injected can reveal effective cross-layer intervention regimes that same-layer steering would miss. Second, comparing sign-asymmetric behavioral effects against sign-symmetric perplexity changes helps distinguish semantically meaningful directions from generic activation perturbations.

Limitations

Our study, while providing evidence that residual-stream activation steering can control figurative-vs-literal idiom interpretation in LLMs, is subject to several limitations. First, we focus on English idioms and on base, pre-instruction-tuning

language models. The choice of base models is intentional: because they are used directly as next-token predictors, their continuations provide a relatively direct view of how the model interprets an idiom-final prefix. This is central to the IdioSteer design, where the model receives the sentence prefix verbatim, with no instruction, examples, or chat template, and the continuation reveals whether it follows the conventional figurative reading or the compositional literal reading. However, whether the same figurative-vs-literal direction can be recovered in other languages, particularly typologically distant ones where idioms may behave differently, and whether it persists after instruction tuning or RLHF, remain open questions. A further open question is whether the approach extends to other non-compositional or non-literal phenomena—metaphor, simile, hyperbole, or irony—where a conventional reading similarly competes with a compositional one.

Second, IdioSteer is deliberately small: 20 idioms with 5 sentence prefixes each, selected for bidirectional plausibility. This scale is a direct consequence of the construction constraint rather than of annotation budget: for each idiom, the prefix must end exactly at the idiomatic expression while leaving both the conventional figurative reading and the compositional literal reading naturally available. This rules out most naturally occurring idiom usages, where the preceding context already resolves the interpretation. The benchmark therefore isolates figurative-vs-literal continuation behavior under carefully validated ambiguity, but it does not exhaust the diversity of English idioms, and aggregate literal rates over 20 idioms should be read as estimates over a small, deliberately constrained sample—we report per-idiom breakdowns (Appendix Q) and bootstrap intervals over idioms (Appendix I) for this reason. Future work could extend the benchmark to a larger idiom inventory, additional syntactic forms, and broader contextual settings.

Third, our argument that the steering vector reflects a meaningful representation of figurative-vs-literal interpretation is supported by behavioral evidence—cross-architecture consistency, sign asymmetry, and generalization across held-out idioms—rather than by a complete causal account of how models naturally choose between figurative and literal readings. We do not test whether the same direction is necessary for ordinary, unsteered idiom interpretation outside the interven-

tion regime. The same applies to our account of why cross-layer steering works (Section 6.3): the evidence that mid-to-late layers are the best place to estimate the vector and early layers the best place to inject it is correlational, and causal methods we tried during development were inconclusive on a benchmark of 20 idioms.

Fourth, what the steering vector encodes is only partly characterized. Our concreteness control study (Section 6.4, Appendix R) shows that a substantial share of the average steering effect is attributable to a general concreteness vector that contains no idiom signal, and that the residual, concreteness-ablated component is what controls interpretation on idioms where literalness and concreteness dissociate. We therefore do not claim to have isolated a pure idiomaticity direction, and the decomposition itself is only established behaviorally on one model. Relatedly, the training pool is not matched between classes: IdioLink balances idioms and subject categories by construction, and our filtering leaves that balance essentially intact (Appendix C), but figurative and literal sentences may still differ systematically in syntax and local context within a category. Such covariates could contribute to the estimated direction. A minimal-pair training pool, matched for length, syntax, topic, and register, would be the most stringent control; the concreteness vector of Appendix R, built from identical sentence templates, is a step in that direction but covers only one covariate.

Finally, evaluation at scale relies on an LLM judge. It was selected on 280 Llama-3.2-3B generations and subsequently validated on 795 further human-annotated generations spanning all five models and all steering conditions, where it attains 91.9% agreement (Appendix J.1). Agreement is nonetheless uneven: it is weakest at the coherence boundary under the most aggressive interventions (extreme angular steering, 86.9%), and the direction of judge error is condition-dependent—slightly under-reporting literal continuations under mean-difference steering and over-reporting them under concreteness steering. We report human-corrected rates for the affected conditions, but a fully human-annotated evaluation of every condition remains out of reach at this scale.

Ethics Statement

We use publicly available datasets and model checkpoints in accordance with their intended use and

licenses, as detailed in Table 15 (Appendix S) and Table 16 (Appendix U). IdioSteer consists of newly authored sentence prefixes and annotations created for this work. Although the idiom inventory was selected from MAGPIE (Haagsma et al., 2020), the benchmark examples themselves were manually written, curated, and validated by the authors. IdioSteer is released under a CC-BY-4.0 license. Our use of AI assistants for writing, code, and documentation support is disclosed in Appendix V.

No personally identifiable information is included in IdioSteer, and annotators were instructed to flag and exclude examples containing personal, sensitive, or offensive content. Our work is intended solely for research purposes. Since we evaluate base language models, some generated continuations may contain offensive, biased, or otherwise inappropriate text. We mitigate this risk by using generations only for aggregate analysis and by not releasing raw model completions.

Acknowledgments

This study is supported in part by the European Research Council (Intellexus, Project No. 101118558). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authorities can be held responsible for them.

References

- Saeed Almheiri, Bilal Elbouardi, Salsabila Zahira Pranita, Irina Nikishina, Ashwath Rao B, Parameswari Krishnamurthy, Muhammad Cendekia Airlangga, Rifo Ahmad Genadi, Nguyen Phan Gia Bao, Amir Hossein Yari, Hawau Olamide Toyin, Nurdaulet Mukhituly, Mena Attia, Beshir Hassan, Ahmad Fathan Hidayatullah, Tatsuki Kuribayashi, Haonan Li, Suma Bhat, and Fajri Koto. 2026. [Multilingual idioms in sentences and conversations across high-, medium-, and low-resource languages](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12363–12389, San Diego, California, United States. Association for Computational Linguistics.
- Anthropic. 2024. [Claude 3.5 Sonnet](#). Large language model, accessed via Anthropic API.
- Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. [Refusal in language models is mediated by a single direction](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.
- Doğukan Arslan, Hüseyin Anıl Çakmak, Gülşen Eryiğit, and Joakim Nivre. 2025. [Using LLMs to advance idiom corpus construction](#). In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 21–31, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.
- Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkha and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2nd edition, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.
- Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. [Survey: Multiword Expression Processing: A Survey](#). *Computational Linguistics*, 43(4):837–892.
- Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. [A hard nut to crack: Idiom detection with conversational large language models](#). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
- Ziqi Gao, Yaotian Zhu, Qingcheng Zeng, Xu Zhao, Ziqing Wang, Feng Ruan, and Kaize Ding. 2026. [RAPTOR: Ridge-Adaptive Logistic Probes](#). *Preprint*, arXiv:2602.00158.
- Saptarshi Ghosh and Tianyu Jiang. 2026. [Exploring concreteness through a figurative lens](#). *Preprint*, arXiv:2604.18296.
- Raymond W. Gibbs, Nandini P. Nayak, John L. Bolton, and Melissa E. Keppel. 1989a. [Speakers’ assumptions about the lexical flexibility of idioms](#). *Memory & Cognition*, 17(1):58–68.
- Raymond W. Gibbs, Nandini P. Nayak, and Cooper Cutting. 1989b. [How to kick the bucket and not decompose: Analyzability and idiom processing](#). *Journal of Memory and Language*, 28(5):576–593.
- Raymond W. Gibbs, Jr and Herbert L. Colston. 2012. *Interpreting Figurative Meaning*. Cambridge University Press.
- Google DeepMind. 2026. Gemma 4: Byte for byte, the most capable open models. <https://ai.google.dev/gemma/docs/core/>. Accessed: 2026-05-10.
- Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. [MAGPIE: A large corpus of potentially idiomatic expressions](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages

- 279–287, Marseille, France. European Language Resources Association.
- Kai Golan Hashiloni, Daniel Fadlon, Lior Livyatan, Ofri Hefetz, Jiahuan Pei, and Kfir Bar. 2026a. [IdioLink: Retrieving meaning beyond words across idiomatic and literal expressions](#). *Preprint*, arXiv:2605.22247.
- Kai Golan Hashiloni, Ofri Hefetz, and Kfir Bar. 2025. [Easy as PIE? identifying multi-word expressions with LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23771–23790, Suzhou, China. Association for Computational Linguistics.
- Kai Golan Hashiloni, Lior Livyatan, Ofri Hefetz, Alon Mannor, Bar Cohen, and Kfir Bar. 2026b. [ID10M-JAM: Stress-testing idiom identification under challenging context](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 20846–20864, San Diego, California, United States. Association for Computational Linguistics.
- Ofri Hefetz, Kai Golan Hashiloni, Alon Mannor, and Kfir Bar. 2025. [Not just a piece of cake: Cross-lingual fine-tuning for idiom identification](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2521–2537, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Ray S. Jackendoff. 1997. *The Architecture of the Language Faculty*, volume 28 of *Linguistic Inquiry Monographs*. MIT Press, Cambridge, MA; London, England.
- Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. [Comparative study of multilingual idioms and similes in large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. [Memorization or reasoning? exploring the idiom understanding of LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21678–21699, Suzhou, China. Association for Computational Linguistics.
- Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. [Rolling the DICE on idiomaticity: How LLMs fail to grasp context](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.
- nostalgebraist. 2020. Interpreting GPT: The logit lens. <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>. Accessed: 2026-05-17.
- Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. *Idioms*. *Language*, 70(3):491–538.
- Soyoung Oh, Xinting Huang, Mathis Pink, Michael Hahn, and Vera Demberg. 2026. [Tug-of-war between idioms’ figurative and literal interpretations in LLMs](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2942–2958, Rabat, Morocco. Association for Computational Linguistics.
- Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. [Sign of the times: Evaluating the use of large language models for idiomaticity detection](#). In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.
- Howard R. Pollio, John M. Barlow, Howard J. Fine, and Marilyn R. Pollio. 1977. *Psychology and the Poetics of Growth: Figurative Language in Psychology, Psychotherapy, and Education*. Lawrence Erlbaum, Hillsdale, NJ.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 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)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. [Massive activations in large language models](#). In *First Conference on Language Modeling (COLM)*.
- Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. [Language models linearly represent sentiment](#). In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US. Association for Computational Linguistics.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. [Steering language models with activation engineering](#). *Preprint*, arXiv:2308.10248.
- Sohan Venkatesh and Ashish Mahendran Kurapath. 2026. [On the non-identifiability of steering vectors in large language models](#). *Preprint*, arXiv:2602.06801.
- Hieu M. Vu and Tan M. Nguyen. 2025. [Angular steering: Behavior control via rotation in activation space](#). *Preprint*, arXiv:2510.26243.

Uriel Weinreich. 1969. [Problems in the analysis of idioms](#). In Jaan Puhvel, editor, *Substance and Structure of Language*, pages 23–81. University of California Press, Berkeley and Los Angeles.

Fengying Ye, Yanming Sun, Runzhe Zhan, Lidia S. Chao, Zheqi Zhang, and Derek F. Wong. 2026. [G-IdiomAlign: A gloss-pivoted benchmark for cross-lingual idiom alignment](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 38720–38739, San Diego, California, United States. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Qualitative Steering Examples

Table 1 shows qualitative examples of mean-difference steering in the literal direction: unsteered continuations typically follow the conventional idiomatic reading, while steered continuations use the idiom’s component words compositionally.

Continuation diversity. The examples in Table 1 illustrate a broader pattern: steered literal continuations are qualitatively diverse rather than formulaic. Across the full evaluation, literal completions vary in the compositional strategy they employ—some reinterpret the idiom’s head noun as a physical object (e.g., *ice*, *beans*), others extend a component word into an unexpected compound (*out of the blue room*; *take a bow and arrow*), and others construct novel scenarios around the literal meaning of the full phrase. We do not observe systematic mode collapse toward a single literal template for any idiom. We report this as a qualitative observation and do not quantify continuation diversity here.

B IdioSteer Construction and Validation Details

Idiom selection. We manually selected 20 English idioms from MAGPIE (Haagsma et al., 2020). Selection was performed once, before any steering experiments, and was based only on whether the idiom could support five sentence prefixes satisfying our construction constraints. We did not select idioms based on anticipated steerability or on any

property of the steering results. The IdioSteer idioms are disjoint from all idioms used to estimate steering vectors.

Construction guidelines. Each sentence prefix is written to satisfy five constraints: (i) the target idiom appears at the end of the sentence, with no post-idiom context; (ii) the preceding context does not force either the figurative or literal interpretation; (iii) both figurative and literal continuations are natural; (iv) the sentence is fluent and plausible under both readings; and (v) the five variants for each idiom differ in wording and scenario. The two readings are not required to be equally likely; the goal is bidirectional plausibility rather than balanced prior probability.

Sentence authoring. For each idiom, we manually authored five sentences, each ending in the idiomatic expression. We used Claude 3.5 Sonnet (Anthropic, 2024) only to propose candidate sentences during drafting. All final examples were curated, edited, and approved by the authors; no model-generated candidate was used without human revision and validation.

Validation protocol. Two independent annotators validated each candidate sentence using a dual-completion protocol. For each sentence, annotators were asked to: (i) produce a short natural continuation under a figurative interpretation; (ii) produce a short natural continuation under a literal interpretation; (iii) mark the sentence as *valid* if both continuations were natural, or *invalid* otherwise; and (iv) for invalid cases, propose a revision restoring bidirectional plausibility. Requiring annotators to produce both continuations is a stronger check than asking for an ambiguity judgment alone, since it forces each annotator to commit to a concrete reading before judging validity.

Iterative refinement. The benchmark is constructed and validated through several iterations. Sentences flagged as invalid by either annotator are revised—typically by changing the subject or surrounding clause—and re-circulated. Iteration continued until all 100 sentences received the *valid* label from both annotators.

Idiom inventory. The 20 IdioSteer idioms are: *up in the air*, *in hot water*, *out of the blue*, *break the ice*, *behind bars*, *on ice*, *see red*, *pull someone’s leg*, *at the crossroads*, *lie low*, *shut the door on*, *under the sun*, *take a bow*, *have a ball*, *in stitches*, *in the*

Model	Unsteered continuation	Steered continuation
Gemma-2-2B	During the party, Sarah tried to <i>break the ice</i> by talking to the man next to her.	During the party, Sarah tried to <i>break the ice</i> with her bare hands, but the result was disastrous. The ice was so strong that she couldn't even break it. It was almost as if the ice had been glued together. Sarah was so frustrated that she didn't know what to do.
Llama-3.2-3B	This morning she called me <i>out of the blue</i> and said she had something to tell me.	This morning she called me <i>out of the blue</i> room and asked me to come to her room.
Llama-3.1-8B	He said nothing at first, but deep down he knew he was <i>in hot water</i> .	He said nothing at first, but deep down he knew he was <i>in hot water</i> and had to get out of there.
Gemma-3-4B	Before the competition started, he walked to the center to <i>take a bow</i> and greet the audience.	Before the competition started, he walked to the center to <i>take a bow</i> and arrow.
Gemma-3-12B	During the meeting he accidentally <i>spilled the beans</i> and said he was going to a secret location.	During the meeting he accidentally <i>spilled the beans</i> on the floor.

Table 1: Qualitative examples of mean-difference steering in the literal direction. Unsteered continuations typically follow the conventional idiomatic reading, while steered continuations use the idiom’s component words compositionally.

pink, green light, rough edges, in the fast lane, spill the beans.

Final dataset. The final IdioSteer benchmark contains 100 sentence prefixes, distributed evenly as five prefixes per idiom across 20 English idioms. All prefixes end at the target idiom and contain no post-idiom context. Each prefix was included only if both annotators judged that natural continuations were available under both the figurative and literal readings. The IdioSteer idioms are disjoint from the idioms used to construct the steering vectors. The dataset, construction guidelines, and annotation guidelines are available in the released repository.

C Training Pool: IdioLink Subset

We construct steering vectors from the train split of IdioLink (Hashiloni et al., 2026a), a benchmark containing figurative and literal uses of English idioms. We describe here only the subset used for steering-vector construction; the full dataset construction is described in the IdioLink paper.

Raw and filtered counts. The raw IdioLink train split contains 1,760 sentences across 22 idioms, balanced by construction: 80 sentences per idiom (30 figurative + 50 literal) and 176 sentences per subject category. Our filtering retains only sentences where the idiom string is literally present (no inflected forms), yielding 1,683 sentences (95.6% retention). The dropped 77 sentences mostly involve inflected or altered surface forms, with the largest counts for *see the light* (21), *go to town* (14), *out of pocket* (13), *get stuck in* (13), and *to*

boot (7). Sentences in the retained pool average 107 characters, ranging from 67 to 180.

Class balance after filtering. Filtering leaves the overall class balance essentially unchanged: 37.4% figurative after filtering (629/1,683) versus 37.5% before (660/1,760), and per-category balance is likewise preserved (61–65 figurative and 100–110 literal sentences per subject category). Two idioms lose a substantial share of their figurative examples, because their figurative uses are more often inflected: *see the light* retains 12 of 30 figurative sentences (against 47 literal) and *go to town* retains 18 of 30 (against 48 literal). The remaining 20 idioms retain at least 26 of 30. We note this because the mean-difference construction weights all retained sentences equally; the cross-fold analysis of Section 6.4, in which folds containing and excluding these idioms behave equivalently, indicates the effect does not depend on them.

Label definitions. Each sentence is labeled as either *figurative*, where the idiom carries its conventional non-compositional meaning, or *literal*, where the component words are used compositionally outside the idiomatic frame. For example, a figurative usage may describe someone being *in hot water* with a committee, whereas a literal usage may describe a substance being dissolved in *hot water*.

Subject distribution. The raw training split is stratified across 10 subject categories: Politics, Sport, Technology, History, Medicine, Culture, Entertainment, Food, Business, and Environment, with 176 sentences per category before filtering.

Annotation provenance. The IdioLink train split combines human-annotated and weakly labeled examples: 271 of the 1,760 raw train sentences (15.4%) are gold-annotated, and the remaining 1,489 (84.6%) are weakly labeled.

Disjointness from evaluation. The 22 idioms used to construct the canonical steering vectors are disjoint from the 20 idioms in IdioSteer. Therefore, all main evaluations are performed on idioms unseen during steering-vector construction.

The 22 training idioms. *cut and dried, get a grip, get stuck in, go to town, in black and white, in the can, in the driving seat, in your face, nuts and bolts, off the shelf, old hat, on paper, on the stump, on the wagon, out of pocket, out of the woods, over the top, point the finger, round the bend, see the light, to boot, under a cloud.*

D Probe construction and PCA analysis

This appendix expands on the brief note in Section 4.1.

Training details. The probe is a binary logistic regression $p(c = \text{fig} \mid h) = \sigma(w^\top h + b)$ fit on 1,683 training activations from IdioLink, with figurative activations labeled 1 and literal activations labeled 0. We use scikit-learn’s LogisticRegression with L2 regularization ($C = 1.0$), the LBFGS solver, maximum 1,000 iterations, default tolerance (10^{-4}), and `fit_intercept=True` (the intercept is fit but not regularized by LBFGS). We do not standardize activations and do not rebalance classes (the 37%/63% imbalance is left as-is). The same hyperparameters are used across all five models; per-model convergence ranges from 30 to 322 iterations on the full fit (Table 2). No model triggered a non-convergence warning.

Training and held-out accuracy. Across all five models, the probe achieves 1.0 training accuracy at the canonical source layer. In the over-parameterized regime (d/n between 1.4 and 2.4), training accuracy is trivially achievable and uninformative on its own. We therefore additionally report 5-fold stratified cross-validation accuracy (Table 2). Held-out accuracy ranges from 0.969 to 0.975 across all five models, indicating that the figurative-vs-literal distinction is linearly decodable from residual-stream activations at L_s on every model. This consistency is notable in light of

the headline finding in Section 6.1 that PROBE-additive collapses on the Gemma family: the probe *does* learn the distinction on Gemma activations, but the recovered weight vector does not function effectively as a steering direction. The disconnect between classification accuracy and steering effectiveness is the non-identifiability concern made concrete (Section 4.1; Venkatesh and Kurapath, 2026).

Model	d	L_s	Train acc.	5-fold CV	Iters
Llama-3.2-3B	3,072	14	1.000	0.975 ± 0.009	33
Llama-3.1-8B	4,096	16	1.000	0.975 ± 0.010	30
Gemma-2-2B	2,304	20	1.000	0.969 ± 0.007	86
Gemma-3-4B	2,560	17	1.000	0.969 ± 0.007	322
Gemma-3-12B	3,840	24	1.000	0.970 ± 0.006	220

Table 2: Per-model probe training accuracy and held-out 5-fold stratified CV accuracy. “Iters” is iterations to convergence on the full fit. All probes converged within the 1,000-iteration cap. Held-out accuracy is tight (0.969–0.975) across models, indicating that classification on held-out activations is consistent even though downstream steering effectiveness varies (Section 6.1).

Non-identifiability concern. The probe’s high held-out accuracy does not imply its weight vector w is unique. In over-parameterized regimes, many distinct w achieve comparable training accuracy, and the specific solution depends on initialization, regularization, and the activation set used for training (Gao et al., 2026; Venkatesh and Kurapath, 2026). This matters because we use w (not the probe’s predictions) as a steering vector.

PCA-based remediation. A natural fix is to reduce dimensionality before fitting, ensuring $n > d_{\text{reduced}}$. We compute the top- k principal components of the training activations $H^{\text{fig}} \cup H^{\text{lit}}$, project all activations onto the resulting k -dimensional basis, fit a logistic classifier with the same hyperparameters as the canonical probe on the projected representations, and reconstruct the steering vector in the original residual space by inverse-projecting the probe weight. We test $k \in \{50, 100, 500\}$ on Llama-3.2-3B.

Results. Table 3 reports the geometric properties of the resulting probe vectors. Three observations stand out. First, training accuracy degrades modestly as k decreases (1.000 at full dimension, 0.977 at $k = 50$), confirming that the full-dimensional probe is over-parameterized and that reducing dimensionality removes some optimization freedom. Second, the cosine between the

PCA-reduced probe and the full-dimensional probe falls sharply with k (0.865 at $k = 500$, 0.611 at $k = 100$, 0.533 at $k = 50$): different dimensionality regimes recover materially different directions, consistent with probe non-identifiability. Third, and most informatively, the cosine between the PCA-reduced probe and the mean-difference vector *rises* as k decreases (0.484 at full dimension, 0.717 at $k = 50$). Reducing the dimensionality available to the probe pushes its solution toward the mean-difference vector.

Interpretation. The PCA results refine the non-identifiability picture. The full-dimensional probe is free to spread its decision boundary across many directions in residual space and finds a solution with cosine 0.484 to the mean-difference vector. Restricting the probe to fewer principal directions removes this freedom, and the recovered weight vector aligns more closely with the principal axis of class-conditional variation—i.e., the mean-difference vector. This is consistent with the cross-fold robustness results in Section 6.4: probe vectors trained on disjoint subsets of idioms have low pairwise cosine (~ 0.28) but each retains substantial alignment with its fold’s mean-difference vector (~ 0.59). The probe-derived direction is non-unique in high dimensions, but its instances share substantial projection onto the mean-difference axis, which is itself uniquely identified by the data.

Decision: keep the full-dimensional probe. We adopted the full-dimensional probe as our primary construction because the steering operator acts on the full residual stream regardless of the dimensionality used to estimate the steering vector; reconstructing a probe from a k -dimensional subspace risks discarding signal that lies in the residual basis. We do not report downstream steering effects for the PCA-reduced probes in this submission; the geometric analysis above is sufficient to characterize how dimensionality affects what the probe recovers.

E Angular Steering Operator

This appendix gives the full definition of the angular steering operator used in Section 4.2. Let d denote the residual-stream dimensionality, let $v \in \mathbb{R}^d$ be a unit-norm steering vector, and let $h \in \mathbb{R}^d$ be the residual-stream activation at the intervention position, i.e., the final idiom token at injection layer L_i . Following Vu and Nguyen (2025), the

k	var. expl.	acc.	$\ w\ $	cos w/ MD	cos w/ full
full	1.00	1.000	6.98	0.484	1.000
500	0.94	0.998	7.08	0.484	0.865
100	0.78	0.983	5.56	0.609	0.611
50	0.70	0.977	4.58	0.717	0.533

Table 3: Probe weight vector properties under PCA-based dimensionality reduction on Llama-3.2-3B at $L_s = 14$. “cos w/ MD” is the cosine with the mean-difference vector; “cos w/ full” is the cosine with the full-dimensional probe. Reduced-dimension probes diverge from the full-dim probe while converging toward the mean-difference vector.

goal is to steer by rotating h rather than by adding a scaled vector. The intervention therefore changes the activation direction while preserving its norm.

We use an input-dependent rotation plane defined by the activation h and steering vector v . First, decompose h into its component parallel to v and its component orthogonal to v :

$$c_1 = v^\top h, \quad r = h - c_1 v, \quad c_2 = \|r\|. \quad (4)$$

Here, c_1 is the scalar coordinate of h along v , r is the residual component orthogonal to v , and c_2 is the norm of that orthogonal component. When $c_2 > 0$, define the second basis vector

$$b_2(h) = \frac{r}{\|r\|}. \quad (5)$$

Then $\{v, b_2(h)\}$ is an orthonormal basis for the two-dimensional plane $\text{span}\{v, h\}$, and h can be written as

$$h = c_1 v + c_2 b_2(h). \quad (6)$$

The angular operator rotates this two-dimensional coordinate representation by an angle θ , measured in radians. Let c'_1 and c'_2 denote the rotated coordinates:

$$c'_1 = c_1 \cos \theta - c_2 \sin \theta, \quad (7)$$

$$c'_2 = c_1 \sin \theta + c_2 \cos \theta. \quad (8)$$

The modified activation is then

$$\mathcal{O}_\theta(h, v; \theta) = c'_1 v + c'_2 b_2(h). \quad (9)$$

In the degenerate case $c_2 = 0$, the activation is already parallel to v , so the plane $\text{span}\{v, h\}$ does not define a unique second basis direction. This case is negligible in practice for high-dimensional residual activations; in implementation, we treat it as a no-op and return h unchanged.

This construction has three useful properties. First, $\theta = 0$ gives the identity intervention. Second, for $c_2 > 0$, the operator is norm-preserving:

$$\|\mathcal{O}_\theta(h, v; \theta)\| = \|h\|. \quad (10)$$

Third, the intervention strength is controlled directly by the angular displacement $|\theta|$ within the rotation plane.

The sign of θ determines the rotation direction. Under the coordinate convention above, negative θ rotates h toward v , while positive θ rotates h away from v . Since our steering vector v is oriented toward the figurative class, negative θ pushes activations toward the figurative direction, and positive θ pushes them away from the figurative direction, which we treat as steering toward the literal side of the figurative-vs-literal axis. This sign convention is opposite to the additive operator, where negative α pushes along $-v$ and therefore toward the literal direction.

The choice between additive and angular steering is consequential. Additive steering is sensitive to the activation norm at the injection layer, which can vary substantially across layers and models (Vu and Nguyen, 2025). Angular steering instead decouples intervention strength from activation norm and bounds the intervention in the geometrically interpretable range $\theta \in [-\pi, \pi]$. We evaluate both operators to assess this trade-off empirically (Section 6).

F Token-Position Selection

In all reported experiments, steering is applied at a single token position p , corresponding to the final token of the idiomatic expression in the input sentence. For example, in “The leaders arrived early to *break the ice*,” the intervention is applied at the final token of *ice*.

This choice is motivated by the causal-attention structure of autoregressive transformers. At the final idiom token, the residual representation can integrate information from the preceding sentence context and from earlier tokens in the idiom. The final token therefore provides a natural intervention site for targeting the idiom as a completed semantic unit.

During development, we also considered other intervention sites, including the first idiom token and interventions over the full idiom span. First-token intervention is less directly aligned with the completed idiom, since later idiom tokens are not

yet available at that position. Full-span intervention is also natural, but it introduces multiple perturbations per prompt and therefore changes the intervention from a single-site causal manipulation into a cumulative multi-site modification. To keep the intervention minimal and comparable across idioms of different token lengths, we use a single final-token intervention throughout.

We do not treat token-position choice as a controlled ablation in this paper. Instead, we fix the intervention site before the main experiments and focus our controlled analyses on vector construction, steering operator, layer choice, and idiom-level generalization.

G KV-cache implementation details

During autoregressive generation, the steering hook at injection layer L_i executes the following conditional logic at every forward pass:

- If the input sequence length exceeds the target position p (the position of the last prompt token), the hook overwrites the residual-stream activation h at position p with the steered activation.
- If the sequence length is 1 (i.e., only the most recently generated token is being processed, with prior tokens served from the KV cache), the condition $p < \text{seq_len}$ fails and the hook is a no-op.

The intervention therefore fires exactly once per generation: on the prompt’s first forward pass, at the canonical target position p . The modified residual stream at layer L_i propagates through subsequent layers $L_i+1, \dots, L_n$ during the same prompt pass, producing modified key-value caches at position p in every attention layer above L_i . All subsequent generated tokens then attend to these modified caches through the standard attention mechanism, without any further hook intervention.

This implementation differs from CAA (Rimsky et al., 2024), which applies the steering vector additively at every post-prompt token position throughout the generation. Our single-position, prompt-pass-only protocol is closer in spirit to the single-position interventions used in mechanistic interpretability, where the goal is to characterize the causal effect of an intervention at one residual-stream location rather than to maintain a sustained behavioral bias across many tokens.

H Experimental and Calibration Details

Motivation for multiplier scaling. The steering intervention in this paper is designed to be global and input-independent. For each model and source layer, we compute a single steering vector from the training pool, unit-normalize it, and reuse it unchanged for all evaluation examples. Likewise, the additive coefficient range is fixed per model: it is not recomputed as a function of the particular input sentence or the activation h being steered. This design is important for interpretability, since it tests whether a single direction and a fixed intervention scale can systematically control figurative-vs-literal interpretation.

However, different models have substantially different residual-stream magnitudes. A raw coefficient such as $\alpha = -20$ therefore does not correspond to the same relative intervention strength across models. To make coefficients comparable while preserving input-independence, we scale additive coefficients by the norm of the raw mean-difference vector computed at the source layer:

$$s_{\text{MD}} = \|\bar{h}^{\text{fig}} - \bar{h}^{\text{lit}}\|. \quad (11)$$

The final steering vector is unit-normalized, but s_{MD} is retained as a model-level scale parameter. Additive steering is then written as

$$\alpha = \alpha_{\text{factor}} \cdot s_{\text{MD}}. \quad (12)$$

Thus, α_{factor} denotes a model-normalized multiplier, while α denotes the raw additive coefficient applied in residual space.

Calibration procedure. For each model, we perform a short calibration procedure before the main evaluation, using only IdioLink training-pool sentences and never the IdioSteer evaluation benchmark. Calibration selects two model-level hyperparameters: the source layer L_s used to compute the steering vector, and the fixed range of α_{factor} values used in the additive grid. The selected values are then frozen and applied to all IdioSteer inputs.

The source-layer search is intentionally coarse rather than exhaustive. We evaluate a small set of candidate layers, with emphasis on middle layers, and choose a layer that produces a clear steering effect without severe fluency degradation on the calibration examples. In most models, the selected layer is approximately halfway through the network; Gemma-2-2B is the only exception, where

the selected source layer is later in the network ($L_s = 20$, approximately 75% depth).

After selecting L_s , we sweep over additive multipliers and identify a range where steering produces a measurable behavioral effect while preserving usable generation quality. We then define a symmetric grid of eight α_{factor} values per model: four negative and four positive. The symmetric grid supports the directional-asymmetry analysis in Section 6.2.

The injection layer L_i is not selected during calibration. We fix $L_i = 2$ for all models in the main experiments, which keeps the application protocol constant across architectures and makes cross-model comparisons easier to interpret. The broader sensitivity to layer choice, including source-layer and injection-layer sweeps as well as same-layer versus cross-layer steering, is analyzed separately in Section 6.3.

Per-model configurations. The calibrated multiplier grids turn out to be comparable across models despite the three-orders-of-magnitude spread in s_{MD} . Table 4 reports the calibrated source layer, fixed injection layer, raw mean-difference norm, and additive coefficient grid for each model.

Across the five models, the union of all grids spans $\alpha_{\text{factor}} \in [-4.78, +4.78]$. Despite s_{MD} varying by three orders of magnitude, the calibrated multiplier ranges are within a factor of ~ 2.5 of one another. We take this cross-model consistency as evidence that the multiplier convention identifies a meaningful per-norm scale at which steering is effective, abstracting over architectural differences in residual-stream magnitude.

Random-vector controls. Random-vector controls use the same parameter grids as the learned methods: for additive steering, each of the three random unit vectors is evaluated across all eight α_{factor} values in the model’s calibrated grid; for angular steering, each is evaluated across all eight θ values in the shared angle grid $\{\pm 60^\circ, \pm 70^\circ, \pm 90^\circ, \pm 120^\circ\}$. This yields $3 \times 2 \times 8 = 48$ random-control conditions per model.

Evaluation grid. The full evaluation grid contains 81 conditions per model: one unsteered baseline, four learned configurations evaluated at eight coefficient values each, and 48 random-control conditions. Each condition is evaluated on the full IdioSteer benchmark with 5 sampled completions per sentence, yielding 500 generations per condition.

Model	L_s	L_i	s_{MD}	α_{factor} grid
Llama-3.2-3B	14	2	4.18	$\pm 1.20, \pm 2.39, \pm 3.59, \pm 4.78$
Llama-3.1-8B	16	2	4.15	$\pm 0.5, \pm 1.0, \pm 2.0, \pm 3.0$
Gemma-2-2B	20	2	96.18	$\pm 0.52, \pm 1.56, \pm 2.60, \pm 3.64$
Gemma-3-4B	17	2	2,299.87	$\pm 0.5, \pm 1.0, \pm 1.5, \pm 2.0$
Gemma-3-12B	24	2	7,134.37	$\pm 0.5, \pm 1.0, \pm 1.5, \pm 2.0$

Table 4: Per-model calibration and evaluation settings. L_s is the source layer used to construct the steering vector; L_i is the injection layer used in the main experiments.

I Bootstrap confidence intervals

We estimate uncertainty over evaluation idioms via a block bootstrap. Each bootstrap sample draws 20 idioms with replacement from the 20 idioms in IdioSteer; each sampled idiom contributes all five sentence variants and all generated completions associated with those sentences. We use $B = 1,000$ resamples with random seed 42 and report percentile 95% confidence intervals (the 2.5 and 97.5 percentiles of the bootstrap distribution) for the headline mean-difference additive operating points and the unsteered baseline. Table 5 reports the resulting intervals for each model.

J Judge Prompt and Validation

Judgment task. Each generated continuation is assigned one of three labels: *figurative*, *literal*, or *incoherent*. A continuation is labeled *literal* only when the idiom’s component words are used compositionally outside the conventional idiomatic frame. A continuation is labeled *figurative* when the idiom is used in its conventional idiomatic sense, even if that sense involves a physical action. For example, a completion in which an actor physically bows after a performance is still treated as figurative for *take a bow*, because the action realizes the conventional idiomatic frame. A continuation is labeled *incoherent* if it is syntactically defective or fails to describe an imaginable situation.

Human annotation sample. To validate the LLM-as-a-judge, we constructed a 280-item human annotation set from Llama-3.2-3B generations. For each of the 20 IdioSteer idioms, we randomly selected one of the five sentence variants. For each selected sentence, we sampled two completions from each of seven steering conditions: the unsteered baseline, mean-difference additive steering at raw $\alpha \in \{-20, -10, +10, +20\}$, and a random-vector control at raw $\alpha \in \{-20, +20\}$. This gives $20 \times 7 \times 2 = 280$ annotated items. Items were

shuffled and assigned blinded IDs, so annotators saw only the sentence prefix and completion, not the steering condition.

Human annotation protocol. Annotators labeled each item in two steps. First, they applied a reading-independent coherence check: the continuation was marked incoherent only if it was syntactically defective or failed to describe an imaginable situation, treating the idiom’s words as ordinary words if necessary. Second, for coherent continuations, annotators compared the figurative and literal readings of the idiom in context and selected the better-supported interpretation. If neither interpretation was clearly preferred, annotators defaulted to *figurative*, reflecting the conventional idiomatic reading.

Annotator information. Two authors annotated and revised the IdioSteer examples and also participated in the human evaluation used to validate the LLM-as-a-judge. One annotator is a PhD student and the other is a Master’s student. Both performed the annotation as part of their work on the paper and received no additional monetary compensation. During judge-validation annotation, annotators were blinded to the steering condition. Difficult cases were resolved through discussion among the authors. Because both annotators are authors of the paper, they were aware that their annotations would be used for dataset validation, judge validation, aggregate analysis, and release of the curated benchmark.

Human agreement. Two annotators independently labeled the 280 items. Inter-annotator agreement reached Cohen’s $\kappa = 0.867$. Disagreements were resolved jointly to produce gold labels. The final gold set contains 158 figurative, 95 literal, and 27 incoherent continuations. Per-label one-vs-rest agreement was $\kappa = 0.899$ for *figurative*, $\kappa = 0.858$ for *literal*, and $\kappa = 0.802$ for *incoherent*. Disagreements were concentrated near the expected decision boundaries: literal-vs-incoherent

Model	Baseline	MD-add literal	MD-add coh.
Llama-3.2-3B	0.116 [0.070, 0.170]	0.518 [0.442, 0.594]	0.956 [0.934, 0.976]
Llama-3.1-8B	0.124 [0.072, 0.182]	0.512 [0.440, 0.586]	0.910 [0.848, 0.958]
Gemma-2-2B	0.134 [0.088, 0.184]	0.574 [0.490, 0.662]	0.988 [0.974, 0.998]
Gemma-3-4B	0.130 [0.074, 0.188]	0.362 [0.280, 0.444]	0.962 [0.940, 0.980]
Gemma-3-12B	0.144 [0.086, 0.210]	0.342 [0.268, 0.418]	0.996 [0.990, 1.000]

Table 5: Bootstrap 95% confidence intervals (percentile) for the unsteered baseline literal rate and the headline mean-difference additive operating point per model. “Baseline” is the unsteered literal rate; “MD-add literal” and “MD-add coh.” are the mean-difference additive literal rate and coherence rate at the per-model best operating point.

cases, where the continuation extended the idiom’s words in an awkward but potentially interpretable way, and literal-vs-figurative cases, where both readings remained plausible.

Candidate judges and prompt iteration. We evaluated several candidate judge models from both open and proprietary providers, including Gemini 2.5 Flash, GPT-4o, GPT-4o-mini, Claude Sonnet 3.5, Claude Haiku, DeepSeek, Llama-3.3-70B-Instruct, Qwen3, and Gemma-4-31B-it. We paired these models with approximately 14 prompt variants, selecting the model-prompt pair with the highest agreement against the human gold labels. Prompt development progressed from a direct single-label classification prompt to a structured-output prompt that requires the judge to reason separately about coherence, the figurative reading, and the literal reading before assigning the final label.

Prompt-design failure modes. Two recurring failure modes guided prompt iteration. First, earlier prompts sometimes conflated literal physical actions with literal interpretation. This is incorrect when the physical action is part of the conventional idiomatic meaning, as in *take a bow*. The final prompt therefore explicitly distinguishes using an idiom *as an idiom* from using its component words compositionally. Second, earlier prompts sometimes treated unusual literal uses as incoherent. The final prompt addresses this by applying coherence independently of idiom interpretation: a surprising literal continuation can be coherent if it describes an imaginable situation.

Final judge structure. The selected judge prompt elicits structured intermediate fields rather than only a single label. The judge first states the conventional figurative meaning of the idiom and the literal compositional meaning of its component words. It then applies each reading to the generated text, identifying cues that support or contradict

each interpretation. The final three-way label is derived from these fields: *incoherent* if the continuation fails the coherence check; *literal* if the literal reading is clearly and substantially better supported than the figurative reading; and *figurative* otherwise.

Judge performance. The selected judge, **Gemma-4-31B-it** (Google DeepMind, 2026), achieved 90.0% accuracy (252/280) and Cohen’s $\kappa = 0.821$ against the human gold labels, close to the human-human agreement of $\kappa = 0.867$. Per-label F1 scores were 0.93 for *figurative*, 0.86 for *literal*, and 0.88 for *incoherent*. Coherent-vs-incoherent agreement was 97.5%, with only 7 of 280 samples crossing the coherent/incoherent boundary. Conditional on both the human gold label and the judge labeling a sample coherent, figurative-vs-literal agreement was 91.5% (227/248, $\kappa = 0.817$).

Calibration scope. The judge was *selected* using the 280-item Llama-3.2-3B annotation set described above, which spans the unsteered baseline, mean-difference additive steering, and random controls. Because that set was used for model and prompt selection, its agreement figures are optimistic; we therefore report it separately as a calibration tier and validate the selected judge on additional, independently annotated generations.

J.1 Extended human validation

We annotated a further 795 generations under the protocol above, covering all five models and all method families. The sample comprises 600 items from the four models not used for judge selection (150 each: Llama-3.1-8B, Gemma-2-2B, Gemma-3-4B, Gemma-3-12B), stratified across steering methods, operators, and coefficient strengths and spread across idioms (at most two items per idiom per stratum per model), plus 195 Llama-3.2-3B items covering the strata absent from the calibra-

tion set: probe steering, extreme angular steering, and the concreteness-steering conditions of Appendix R. Items were shuffled and blinded to steering condition; gold labels were produced by the two annotators with joint adjudication, so no independent inter-annotator agreement is reported for this tier.

Agreement. Table 6 reports judge–gold agreement. Overall accuracy on the 795 validation items is 91.9% (Cohen’s $\kappa = 0.868$), comparable to the 90.0% obtained on the calibration set, indicating that the judge generalizes beyond the model and conditions used to select it. Per-model accuracy ranges from 88.2% to 94.7%; the lowest value is Llama-3.2-3B, whose validation tier deliberately contains only the hardest strata. Errors are concentrated at the coherence boundary rather than in interpretation: only 26 of 795 items cross the figurative–literal boundary, whereas collapsing the three labels to coherent vs. incoherent gives 95.2% agreement, with most residual disagreement arising where the judge accepts degraded text that annotators marked incoherent.

Where the judge is weakest. Per-stratum accuracy (Table 6) is high for the conditions underlying our headline results (98.0% for moderate and 95.8% for extreme mean-difference additive steering, 97.0% for probe). Two strata are weaker: extreme angular steering (86.9%) and the concreteness-ablated condition $v_{\perp}$ (84.3%). Both involve heavily perturbed generations, and in both the dominant error is the judge accepting near-degenerate text as coherent; angular results at the largest rotations should therefore be read with this caveat. Within the concreteness conditions, agreement on the four idioms highlighted in Figure 19 is 100% for *out of the blue*, *behind bars*, and *spill the beans*, and 85.7% for *in the pink*.

Effect on reported rates. We use the validation labels to re-estimate literal rates: for a condition judged with label distribution $P(\text{judge})$, the corrected rate is $\sum_j P(\text{judge}=j) \cdot P(\text{gold}=\text{literal} \mid \text{judge}=j)$, with $P(\text{gold} \mid \text{judge})$ taken from the matching audited stratum. Table 7 reports raw and corrected literal rates at each model’s headline mean-difference additive operating point. All corrections are non-negative and small (between +0.000 and +0.035): in this regime the judge slightly *under*-reports literal continuations, so the rates in Section 6.1 are conservative rather than

	Acc.	κ	N
<i>Calibration tier (used for judge selection)</i>			
Llama-3.2-3B	0.900	0.821	280
<i>Validation tier, by model</i>			
Llama-3.2-3B [†]	0.882	0.806	195
Llama-3.1-8B	0.947	0.917	150
Gemma-2-2B	0.933	0.892	150
Gemma-3-4B	0.927	0.867	150
Gemma-3-12B	0.920	0.868	150
All	0.919	0.868	795
<i>Validation tier, by stratum</i>			
baseline	0.933	0.753	60
MD-additive, moderate	0.980	0.964	100
MD-additive, extreme	0.958	0.934	120
MD-angular, extreme	0.869	0.779	175
probe	0.970	0.921	100
random	0.910	0.849	100
v_{conc}	0.900	0.830	70
$v_{\perp}$	0.843	0.748	70

Table 6: Judge–gold agreement. The calibration tier was used to select the judge model and prompt and is reported separately; validation items are independently annotated and were not used for selection. [†]: the Llama-3.2-3B validation tier covers only the strata absent from calibration (probe, extreme angular, and the concreteness conditions), which are the hardest strata; its baseline, mean-difference, and random coverage is in the calibration tier.

inflated. The concreteness-steering conditions behave oppositely—the judge over-reports literal continuations on that text style—and their corrected rates are lower (v_{conc} : 0.432 $\rightarrow$ 0.370; $v_{\perp}$: 0.344 $\rightarrow$ 0.320 at the best operating point). This condition-dependence is the practical argument for validating an LLM judge separately on each new intervention type rather than once at calibration. Corrections of this kind are also sensitive to which audited stratum is used: applying the extreme-coefficient matrices instead would shift two models’ estimates by 0.04–0.05. We therefore match each condition to its own coefficient band and report raw rates as primary throughout.

Prompt release. The full human annotation guidelines, annotation form, judge prompt, structured-output schema, and decoding settings are included in the released repository.

For reproducibility, Appendix T reports implementation details, random seeds, hardware, software versions, and computational cost.

K Full Coefficient Sweeps

The headline results in Section 6.1 summarize each method by its best operating point within the cali-

Model	Raw	Corrected	Δ
Llama-3.2-3B [‡]	0.518	0.535	+0.017
Llama-3.1-8B	0.512	0.512	+0.000
Gemma-2-2B	0.574	0.574	+0.000
Gemma-3-4B	0.362	0.397	+0.035
Gemma-3-12B	0.342	0.342	+0.000

Table 7: Raw and human-corrected literal rates at each model’s headline mean-difference additive operating point. Corrections are non-negative throughout, indicating the reported rates are not inflated by judge error. Each model’s correction uses the audited stratum matching its headline coefficient band; all five headline coefficients fall in the moderate rather than the extreme band. [‡]: Llama-3.2-3B uses the mean-difference confusion matrix pooled over the four other models, since its own mean-difference gold labels belong to the calibration tier.

brated coefficient grid (Table 4, Appendix H). To show that these results are not driven by a single selected coefficient, we report the full coefficient sweeps for each model below.

For additive steering, Figures 5, 7, 9, 11, and 13 show literal rate and coherence rate as a function of α_{factor} for each of the five models. Mean-difference steering exhibits a clear directional effect across all models: literal rate rises sharply for negative α_{factor} (the literal direction in our sign convention), reaching its maximum within the calibrated grid, while positive α_{factor} remains near the unsteered baseline. Probe additive steering varies in strength across model families (consistent with the family-dependent pattern reported in Section 6.1), while random vectors produce smaller and less directionally structured shifts on every model. Coherence remains high across most moderate coefficients, with degradation only at the largest-magnitude interventions.

For angular steering, Figures 6, 8, 10, 12, and 14 show the corresponding sweep over rotation angles θ . Under our angular sign convention, positive θ rotates activations away from the figurative direction and corresponds to literal steering. Mean-difference angular steering increases literal rate over baseline across a range of positive angles on every model, while very large rotations reduce coherence. Probe and random angular steering show weaker gains and stronger sensitivity to rotation magnitude.

Together, these sweeps support the use of best-over-grid headline summaries in Section 6.1: the reported effects reflect broad regions of effective

steering within the calibrated grid, rather than isolated outlier coefficients.

Table 8 summarizes these sweeps numerically: for each method we report the mean literal rate over the literal-direction half of the calibrated grid. Mean-difference additive steering has the highest half-grid mean on every model—including Gemma-3-4B, where the best-over-grid statistic marginally favors the angular operator—and exceeds baseline at every coefficient in its half-grid on all five models, whereas probe-additive’s half-grid mean is indistinguishable from the unsteered baseline on every Gemma model.

L Perplexity analysis

To complement the judge-derived coherence metric used in our main results, we compute per-token perplexity on the canonical Llama-3.2-3B run. For each generation, we score the steered output under the *unsteered* Llama-3.2-3B (i.e., we measure how surprising the steered generation is to the unmodified model). We report token micro-perplexity (per-token, robust to varying completion lengths) and the perplexity ratio $\text{PPL}_{\text{steered}}/\text{PPL}_{\text{baseline}}$, following the convention of Turner et al. (2024). We aggregate over 500 completions per condition (100 sentences $\times$ 5 samples). No condition produces a degenerate (infinite-loss) generation.

Table 9 reports the literal rate, coherence rate, and perplexity ratio for the principal methods at four representative steering strengths on Llama-3.2-3B.

Two observations from the table. First, perplexity rises modestly with steering: at the best literal point ($\alpha_{\text{factor}} = -2.39$ for Llama-3.2-3B), mean-difference achieves a perplexity ratio of 1.10, a 10% relative increase over baseline. The strongest steering tested ($\alpha_{\text{factor}} = -4.78$) reaches 1.27, still well within the range of fluent generation as confirmed by the judge-coherence rate (0.84 at this point).

Second, perplexity tracks intervention magnitude, not direction. At $\alpha_{\text{factor}} = -4.78$ vs $+4.78$, mean-difference perplexity ratios are 1.27 and 1.14—comparable in magnitude—while the literal rate is 0.49 vs 0.13, a $3.8\times$ asymmetry. Random vectors at the same magnitudes produce a comparable perplexity pattern (1.17 vs 1.10) and near-symmetric literal rates (0.25 vs 0.26). This dissociation supports the Section 6.2 interpretation: mean-difference identifies a directional axis in residual

Llama-3B

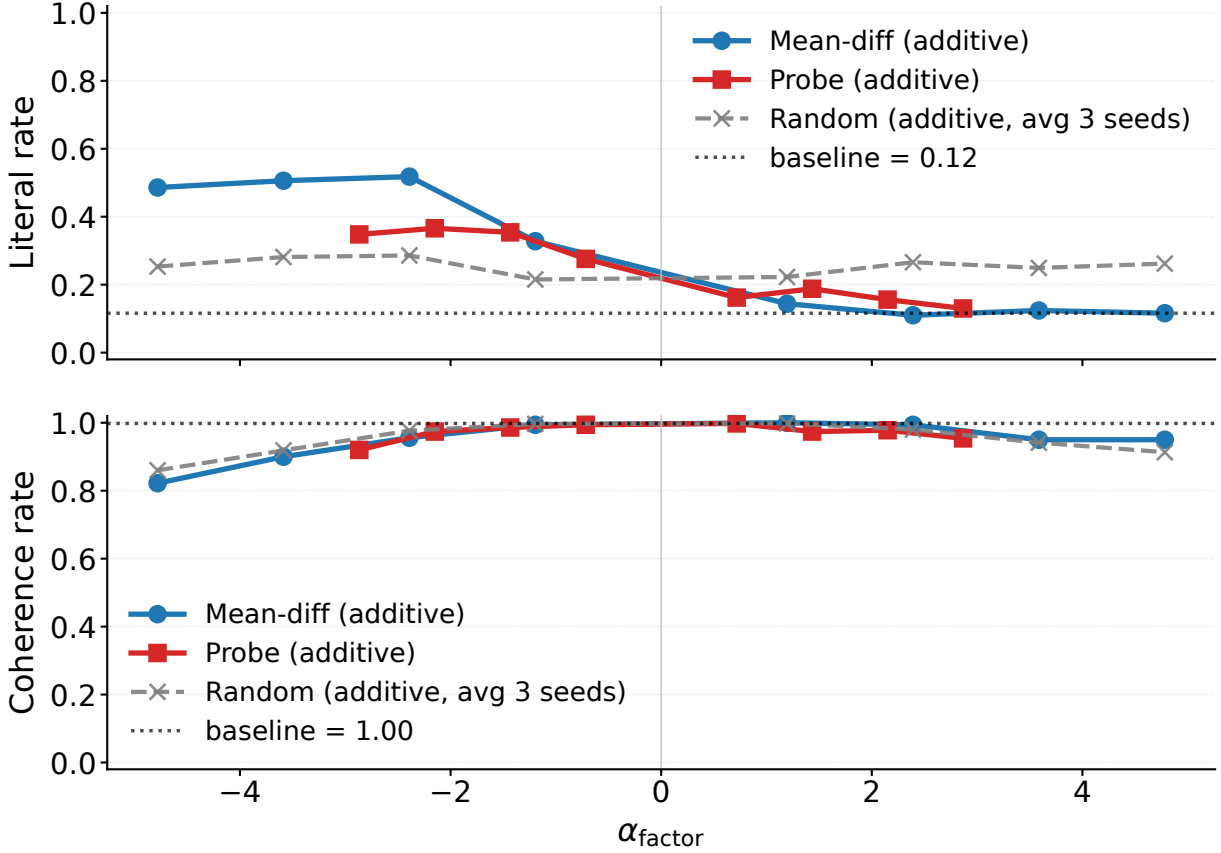


Figure 5: Full additive coefficient sweep on Llama-3.2-3B. Top: literal rate vs α_{factor} for mean-difference, probe, and random-vector steering. Bottom: coherence rate.

space, while perplexity reflects only the perturbative magnitude of any sufficiently large activation modification.

M Layer-sweep heatmaps

Figure 15 shows the evaluated (L_s, L_i) grid for Llama-3.1-8B at four multiplier values. Dotted cells denote layer pairs that were not evaluated. The cross-layer advantage is visible as a hot region in the bottom rows ($L_i \leq 2$) at mid-to-late source layers; the diagonal is uniformly weak. The contrast between negative and positive multiplier panels mirrors the sign asymmetry reported in Section 6.2.

N Matched relative intervention strength

Intervention strength is parameterized as $\alpha = \alpha_{\text{factor}} \cdot s_{\text{MD}}(L_s)$, normalized by the mean-difference norm at the *source* layer. How much the intervention perturbs the forward pass, however, depends on its size relative to the activation it

modifies at the *injection* site,

$$\text{rel} = \frac{|\alpha|}{\|h(L_i)\|}, \quad (13)$$

with $\|h(L_i)\|$ the mean activation norm at L_i over the training pool. The two coincide only when $s_{\text{MD}}(L_s)/\|h(L_i)\|$ is constant, which it is not across layer pairs: on Llama-3.2-3B, $\|h(2)\| = 3.06$ against $\|h(14)\| = 11.46$, so at $\alpha_{\text{factor}} = -2$ the cross-layer configuration $(14, 2)$ reaches $\text{rel} = 2.73$ while the diagonal $(14, 14)$ reaches 0.73. Matching α_{factor} therefore equalizes intervention strength in units of the steering vector’s norm but not in units of the residual stream.

We ran a targeted check on Llama-3.2-3B, re-running the diagonal ($L_s = L_i = 14$) at four additional multipliers chosen so that rel spans $0.5\times$ to $4\times$ the cross-layer value, holding everything else fixed (same unit-normalized direction, same 100 prefixes, $K = 5$, same generation settings, seed, hook, and judge). Table 10 and Figure 16 report the result. At matched relative strength (2.738 vs.

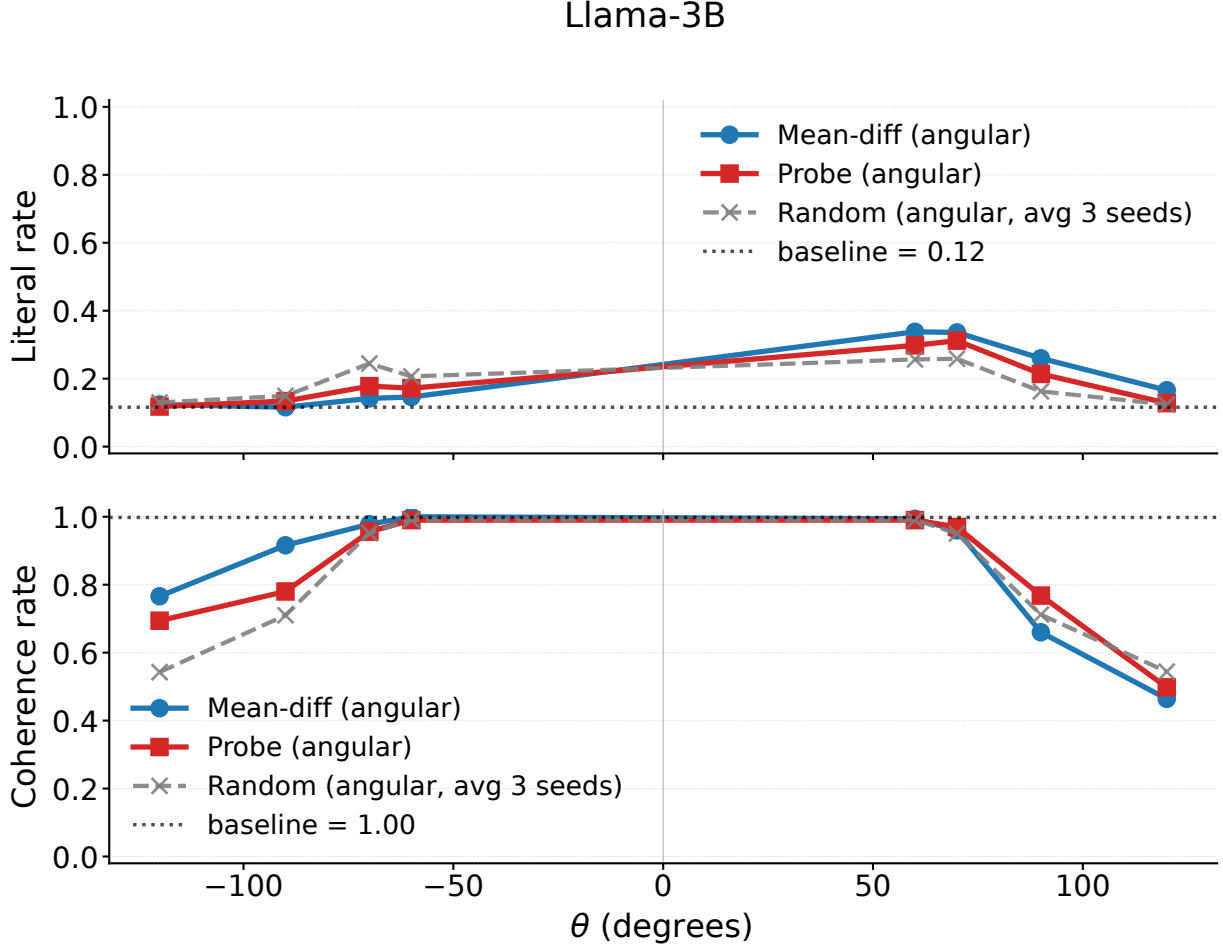


Figure 6: Full angular coefficient sweep on Llama-3.2-3B. Top: literal rate vs rotation angle θ . Bottom: coherence rate.

2.733) and matched coherence (0.978 vs. 0.980), the diagonal attains a literal rate of 0.182 against the cross-layer’s 0.527: the gap does not close. Pushing the diagonal to $2\times$ and $4\times$ raises literal rate only to 0.360 and 0.406, and does so by collapsing coherence to 0.672 and 0.438, so the cross-layer configuration Pareto-dominates every diagonal configuration tested. Between $0.5\times$ and $1\times$ the diagonal is flat at ~ 0.18 , consistent with the dose-response saturation visible at fixed layers in Table 9.

This is a single layer pair on a single model, and we do not treat it as a full characterization: a systematic study of how relative intervention strength interacts with layer choice—including under the norm-preserving angular operator, which fixes rel by construction—is left to future work. It does indicate that the cross-layer advantage reported in Section 6.3 is not an artifact of unequal relative perturbation.

O Per-layer decodability

To characterize the source-layer sweep of Section 6.3 at the representation level, we ask at which depth the figurative-vs-literal distinction becomes *linearly decodable*. We train the probe of Appendix D (same hyperparameters) on the IdioLink training-pool activations at every layer of Llama-3.2-3B and Gemma-2-2B, extracting activations at the final idiom token exactly as for steering-vector construction. Because the pool contains only 22 idioms and the question is whether the distinction generalizes rather than whether idiom identity is memorized, we use 5-fold cross-validation grouped by idiom, so each test fold contains only idioms unseen by the probe. This is stricter than the stratified cross-validation of Table 2, which shares idioms across folds, and therefore yields a slightly lower ceiling (0.943 at the canonical layer of Llama-3.2-3B versus 0.975 stratified).

Figure 17 shows the curves. Both models follow the same profile. At the embedding layer, accuracy

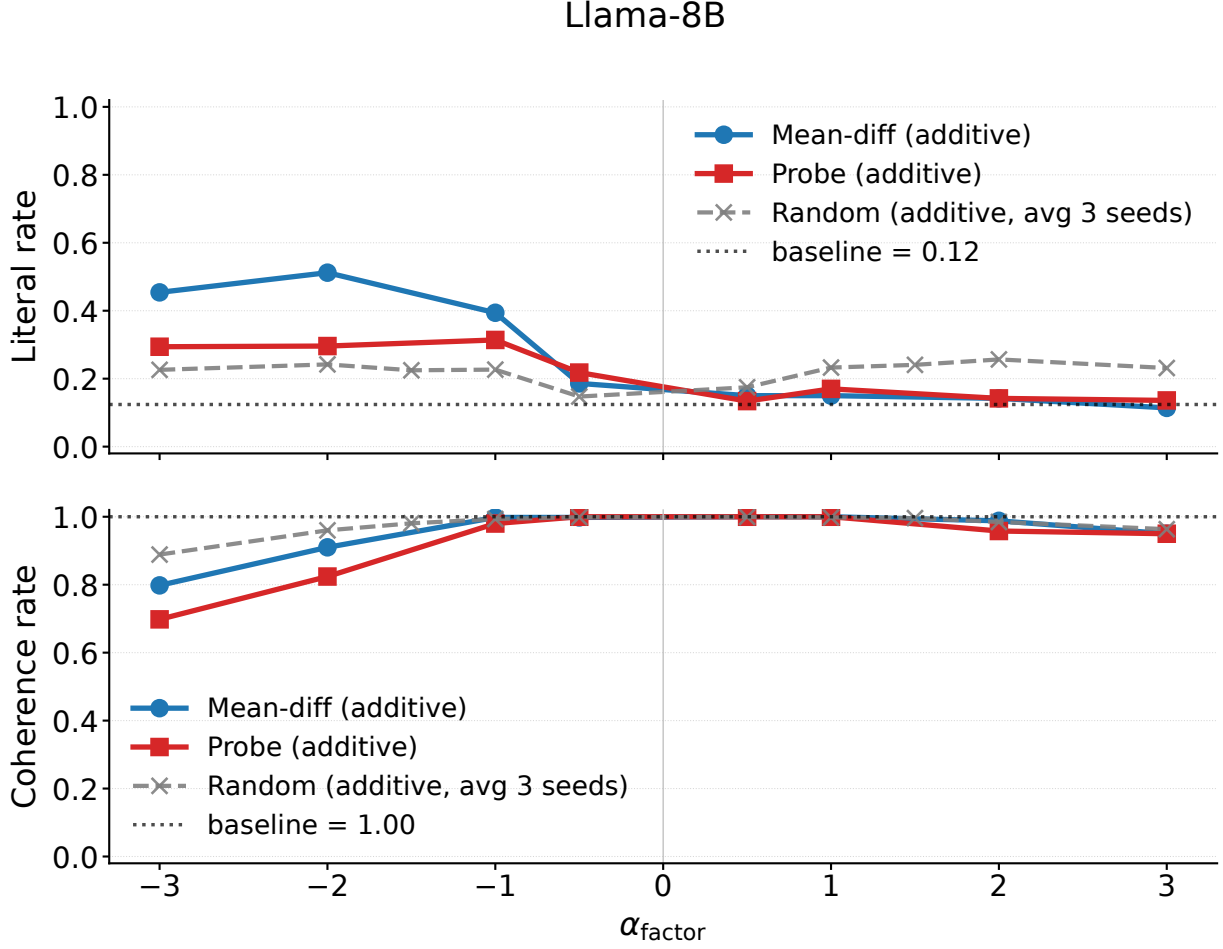


Figure 7: Full additive coefficient sweep on Llama-3.1-8B. Top: literal rate vs α_{factor} . Bottom: coherence rate.

is at the majority-class rate of the pool (62.6% of the pool is literal): 0.626 on Llama-3.2-3B and 0.621 on Gemma-2-2B—the distinction is not linearly available before any transformer block has run. It then rises steeply through the early layers (0.90 by layer 5 on Llama-3.2-3B, 0.85 on Gemma-2-2B) into a broad high plateau, and declines mildly at the deepest layers (0.910 and 0.880 at the final layer). Maxima fall at layer 12 (0.949) on Llama-3.2-3B and layer 14 (0.942) on Gemma-2-2B.

The relationship to the source-layer sweeps of Section 6.3 is informative but not tight. On both models the calibrated source layer lies inside the high plateau: $L_s = 14$ on Llama-3.2-3B (0.943, versus a maximum of 0.949 two layers earlier) and $L_s = 20$ on Gemma-2-2B (0.91, versus a maximum of 0.942 at layer 14). Linear decodability and steering effectiveness therefore agree at the level of *regions*—both identify mid-to-late depth, and both fall off at the embedding layer and the deepest layers—but decodability alone does not pinpoint the best source layer, which on Gemma-2-2B lies

well past the decodability maximum. Accessibility of the distinction is therefore plausibly necessary for a good source layer, but not sufficient to predict one.

Finally, decodability is measured on idioms the probe never saw, reaching 0.94 on both models. What is recoverable at these layers is therefore the figurative-vs-literal contrast itself rather than idiom identity, consistent with the idiom-generalizability reported in Section 6.4.

P Cross-fold Robustness Curves

This appendix shows the literal-rate-vs- α_{factor} curves for the cross-fold robustness analysis (Section 6.4). Figure 18 overlays the curves for the canonical 22-idiom run (four grid points) and the three folds (each trained on 10–11 disjoint idioms and evaluated at $\alpha_{\text{factor}} \in \{\pm 3.59, \pm 4.78\}$). The fold curves track the canonical curve closely at the tested points, indicating that the steering effect is robust to the specific choice of training idioms.

The three fold-derived vectors (10–11 idioms

Llama-8B

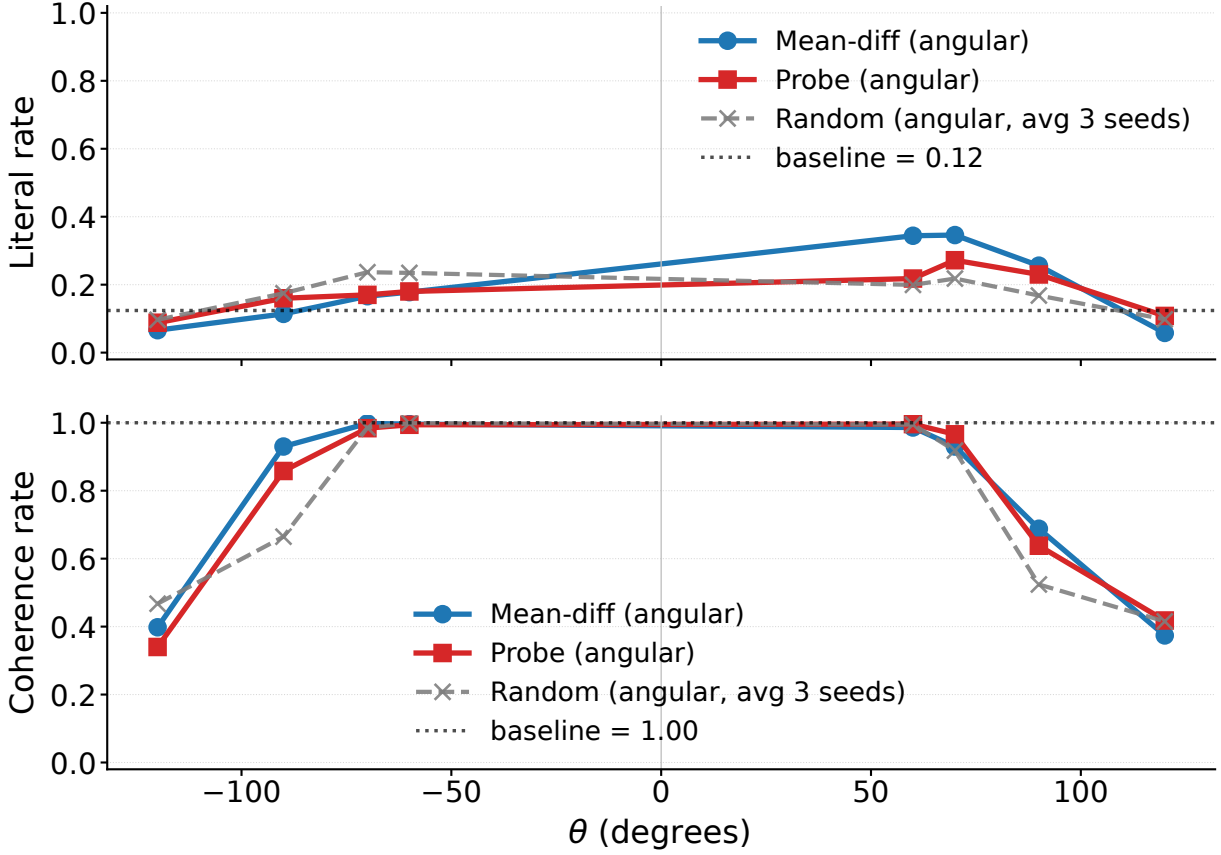


Figure 8: Full angular coefficient sweep on Llama-3.1-8B. Top: literal rate vs θ . Bottom: coherence rate.

each) achieve literal rates of 0.48, 0.46, and 0.42 at $\alpha_{\text{factor}} = -4.78$, versus 0.49 for the canonical vector. Pairwise cosines between fold-derived mean-difference vectors are $\{0.80, 0.71, 0.72\}$.

Q Per-idiom results

The aggregate literal rates of Section 6.1 could in principle be driven by a few highly steerable idioms. Table 11 therefore reports, for every model, the literal rate per idiom at that model’s headline mean-difference additive operating point, alongside the unsteered per-idiom baseline. Rates follow the main-paper convention (all 25 generations per idiom in the denominator).

The effect is broadly distributed. Of the 100 (model, idiom) cells, 97 improve on the unsteered baseline, and each of the 20 idioms improves on at least four of the five models; 17 improve on all five. Per-model mean gains over baseline are +0.40 (Llama-3.2-3B), +0.39 (Llama-3.1-8B), +0.44 (Gemma-2-2B), +0.23 (Gemma-3-4B), and +0.20 (Gemma-3-12B). The three cells that do not

improve are confined to the Gemma-3 models. Two are small negative deviations on idioms whose unsteered baselines are already comparatively high, leaving little headroom (*rough edges*, $0.36 \rightarrow 0.32$ on Gemma-3-4B; *in the fast lane*, $0.56 \rightarrow 0.44$ on Gemma-3-12B); the third is *lie low* on Gemma-3-4B, where steering leaves the literal rate at its floor of 0.00.

Which idioms are easiest to literalize is only partly shared across models: pairwise correlations between per-idiom rate profiles range from 0.21 to 0.81 (mean 0.54). Some idioms are uniformly easy—*shut the door on* and *have a ball* gain more than 0.5 on average—while others are uniformly hard: *under the sun* and *up in the air* stay at or below 0.32 on every model, and *out of the blue*, which requires reanalyzing the phrase rather than depicting a concrete scene (Appendix R), reaches 0.36–0.44 on the Llama models and Gemma-2-2B but only 0.04–0.08 on the Gemma-3 models.

Gemma-2B

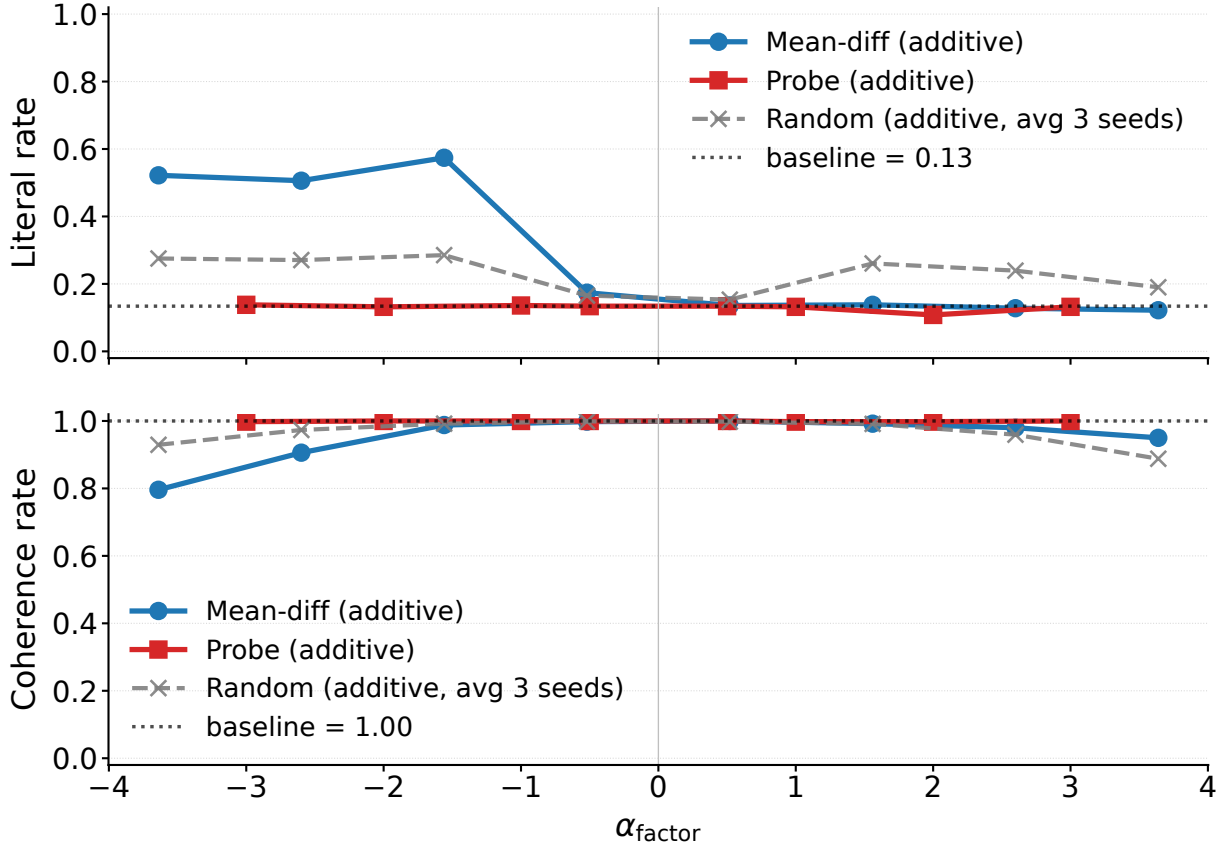


Figure 9: Full additive coefficient sweep on Gemma-2B. Top: literal rate vs α_{factor} . Bottom: coherence rate.

R Concreteness Control Analysis

This appendix details the concreteness control reported in Section 6.4. Literal continuations of idioms tend to describe more concrete situations than figurative ones, so a mean-difference vector estimated from figurative-vs-literal idiom usages could partly reflect concreteness rather than idiomaticity. We therefore construct a concreteness vector that contains no idiom signal by construction, compare it geometrically to the idiom direction, and test both behaviorally.

R.1 Constructing a non-idiomatic concreteness vector

Word selection. We use the concreteness norms of Brysbaert et al. (2014), which provide human ratings for $\sim 40\text{k}$ English words on a 1–5 scale (1 = abstract, 5 = concrete). We select 250 highly concrete nouns (rating ≥ 4.5 ; e.g., *hammer*, *puddle*) and 250 highly abstract nouns (rating ≤ 2.0 ; e.g., *justice*, *notion*), restricted to lowercase alphabetic single words tagged as nouns and tokenizing into

at most four subwords in every model’s tokenizer. Word lists are released with the code.

Carrier sentences. Because our idiom pipeline extracts activations from expressions in sentential context, each word is placed in three semantically neutral template sentences: “*She thought about the ____ for a while.*”, “*They discussed the ____ during the meeting.*”, “*He mentioned the ____ in his letter.*” The templates use epistemic verbs that are equally natural for concrete and abstract nouns, and both classes pass through identical templates, so topic, genre, syntax, and length are held constant and the two activation sets differ only in the target word. This is a stricter control than the idiom training pool itself affords (Appendix C).

Extraction and construction. Activations are collected exactly as for the idiom vector: a forward pass through the frozen model, recording the residual stream at the *final subword of the target word*, at the model’s canonical source layer L_s , cast to fp32. The concreteness vector v_{conc} is the unit-normalized difference of class means (concrete –

Gemma-2B

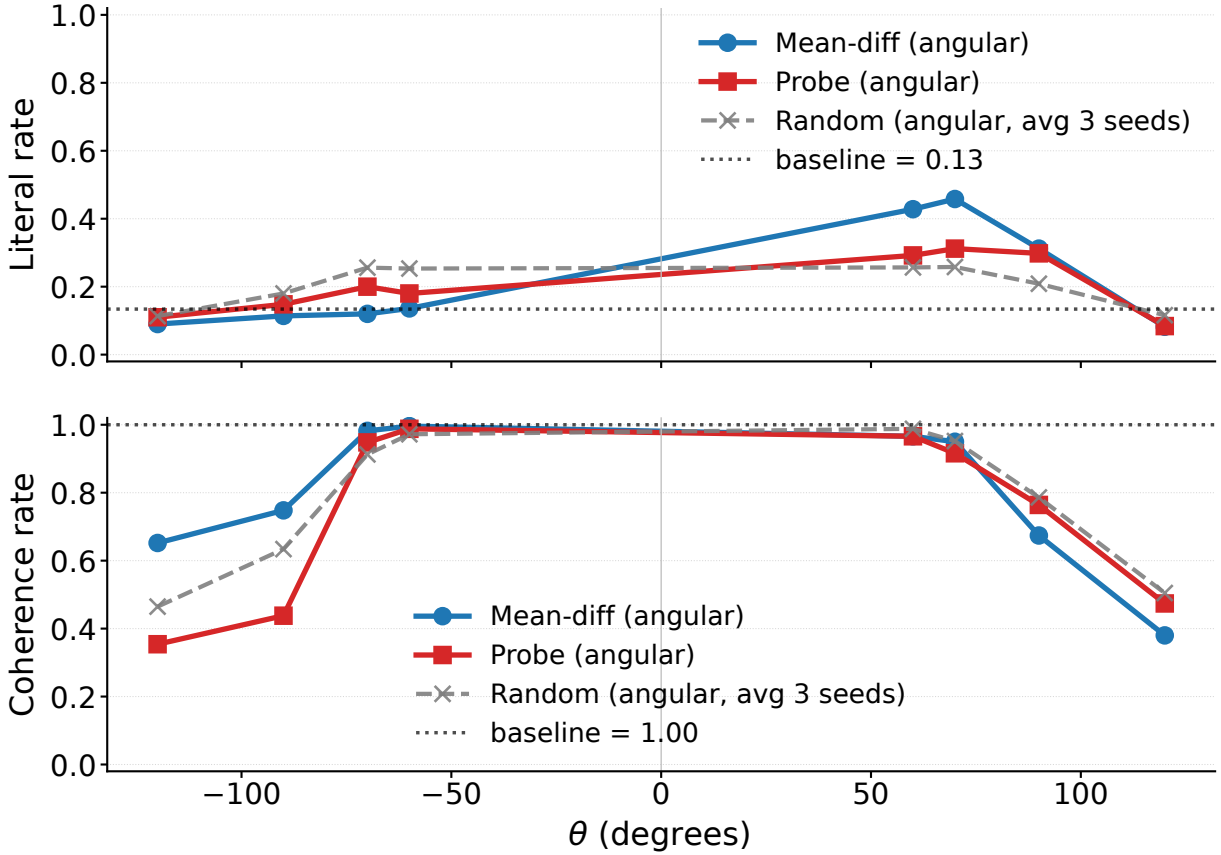


Figure 10: Full angular coefficient sweep on Gemma-2B. Top: literal rate vs θ . Bottom: coherence rate.

abstract), i.e. Equation 1 with the concrete/abstract classes substituted for figurative/literal. Its positive pole is *concrete*; the idiom vector’s positive pole is *figurative*, so a negative cosine between them indicates that figurative aligns with abstract.

Robustness to construction choices. As an independent construction we also build a concreteness vector following the recipe of Ghosh and Jiang (2026) (thresholds $\geq 4.0 / \leq 2.0$, no part-of-speech filter, bare-word extraction). The two constructions recover closely aligned axes ($\cos = 0.65$ on Llama-3.2-3B; $0.59\text{--}0.66$ on the three models whose raw geometry is unaffected by the anisotropy discussed below), indicating the operationalization is not an artifact of our thresholds, filtering, or carrier templates.

Validation. Split-half cosines are computed by partitioning the word list into two halves (word-level splits, so all sentences of a given word fall on the same side), estimating a mean-difference vector independently from each half, and taking the cosine between the two. Both directions are stably

estimable on every model: split-half cosines are ≥ 0.96 throughout. Held-out concrete-vs-abstract separability by projection is ≥ 0.996 for v_{conc} on Llama-3.2-3B, Llama-3.1-8B, and Gemma-2B, but only $0.63\text{--}0.64$ on the Gemma-3 models in raw space; this is a consequence of the massive-activation anisotropy discussed below, and separability recovers to ≥ 0.99 on all five models once activations are whitened per dimension. Per-model values are in Table 13.

R.2 Geometry

Table 12 reports cosines between the idiom and concreteness vectors for all five models. On Llama-3.2-3B, $\cos(v_{\text{MD}}, v_{\text{conc}}) = -0.39$: the expected sign, but well below the $0.71\text{--}0.80$ agreement between idiom vectors derived from disjoint idiom subsets (Section 6.4) and below the 0.48 alignment between our probe and mean-difference vectors (Appendix D). The concreteness subspace accounts for $\sim 17\%$ of the idiom vector’s squared norm.

Gemma-4B

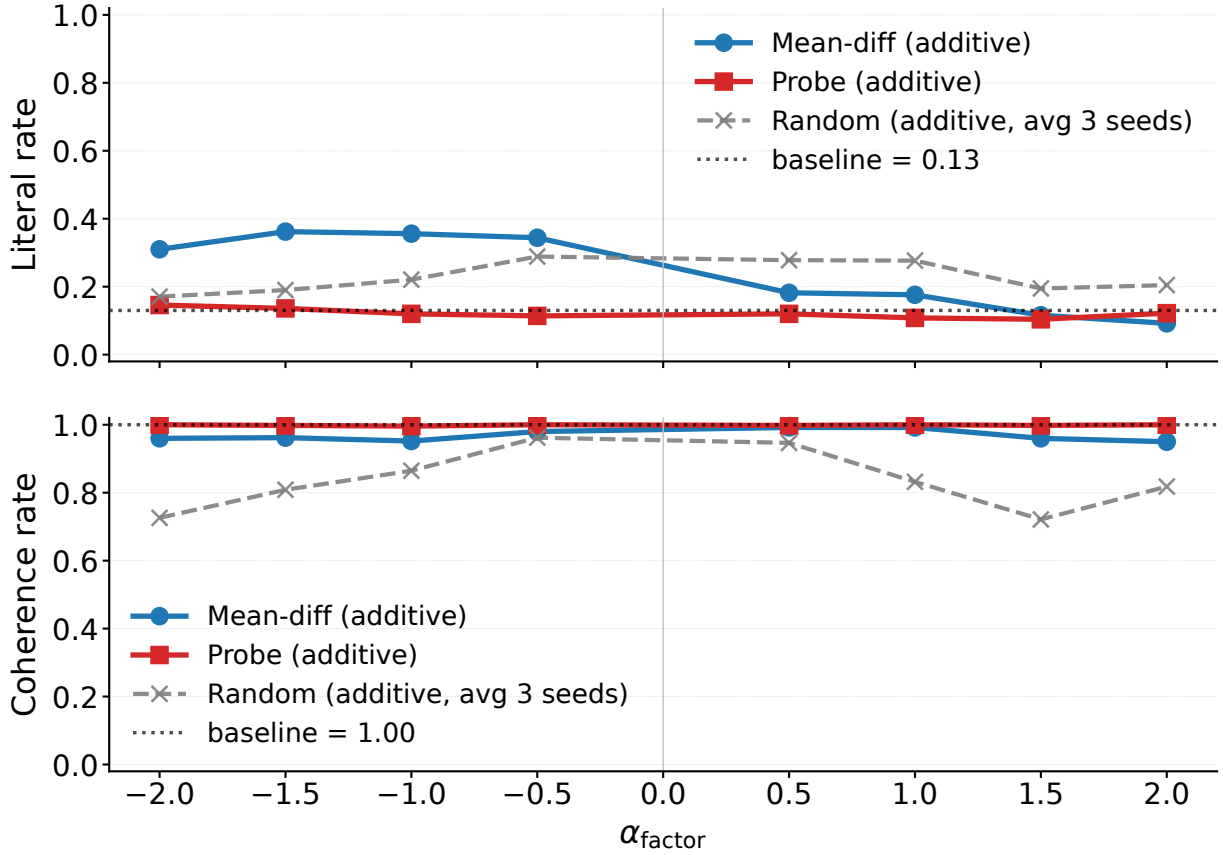


Figure 11: Full additive coefficient sweep on Gemma-3-4B. Top: literal rate vs α_{factor} . Bottom: coherence rate.

Massive-activation dimensions inflate raw cosines. Raw cosines vary widely across models (-0.32 to -0.89), with extreme values on the Gemma-3 family. This is an artifact of massive-activation dimensions (Sun et al., 2024): on Gemma-3-4B/12B a single dimension carries 84%/95% of the idiom vector’s squared norm and 39%/77% of the concreteness vector’s, and zeroing that single shared dimension collapses the cosines to $-0.39/-0.37$. As a control, a semantically unrelated contrast—a mean-difference vector between IdioLink subject categories, computed on the same activations—attains $\cos = +0.79/+0.73$ with the idiom vector on Gemma-3, versus $+0.27/+0.31$ on the other three models: under this anisotropy, *any* two mean-difference vectors appear aligned. Under per-dimension whitening, the idiom–concreteness alignment is modest and uniform across all five models (-0.17 to -0.27), and the concreteness vector’s held-out separability, which fails on Gemma-3 in raw space ($0.63-0.64$), recovers to ≥ 0.99 . Whitening statis-

tics (per-dimension mean and standard deviation) are estimated from the pooled union of the IdioLink training-pool activations and the concreteness carrier-sentence activations at the model’s canonical source layer L_s ; each mean-difference vector is z-scored with these statistics and renormalized to unit ℓ_2 norm before the cosine is taken. These same dimensions account for the three-orders-of-magnitude spread in raw mean-difference norms noted in Section 6.1. We therefore report whitened cosines for cross-model comparison, and caution that raw cosine similarity between mean-difference steering vectors is unreliable on models exhibiting massive activations.

R.3 Behavioral decomposition

We steer IdioSteer on Llama-3.2-3B with three vectors at matched intervention strength (raw additive coefficients $\alpha \in \pm\{5, 10, 15, 20\}$; all vectors unit-norm, so magnitude is comparable), evaluated with the same generation and judging protocol as the main experiments:

Gemma-4B

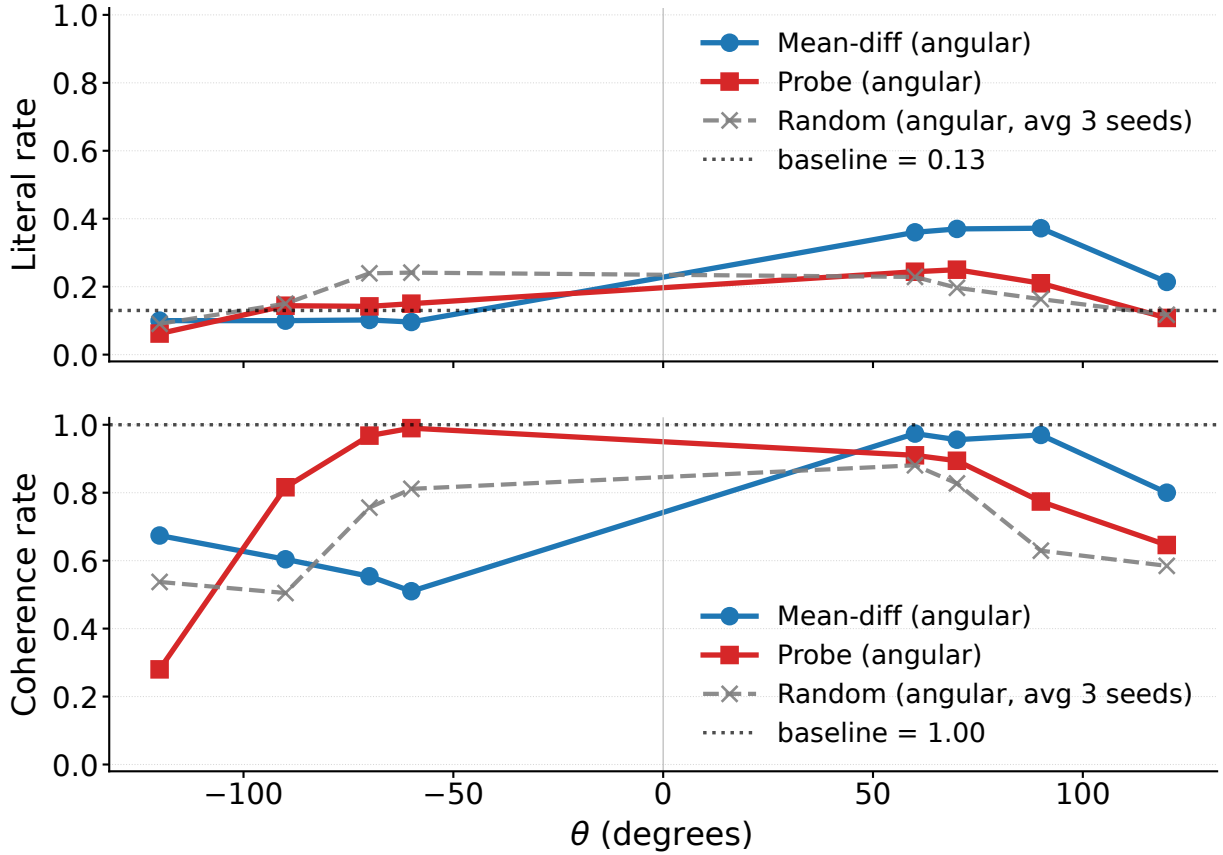


Figure 12: Full angular coefficient sweep on Gemma-3-4B. Top: literal rate vs θ . Bottom: coherence rate.

- v_{MD} : the canonical idiom vector.
- v_{conc} : the pure concreteness vector. Its positive pole is concrete, so the literal direction here is *positive* α , opposite to v_{MD} .
- $v_{\perp}$: the idiom vector with its concreteness component removed, $v_{\perp} \propto v_{\text{MD}} - \Pi_C(v_{\text{MD}})$, where Π_C projects onto the two-dimensional subspace spanned by our two concreteness constructions, renormalized to unit norm ($\|\Pi_C(v_{\text{MD}})\| = 0.41$, $\|v_{\text{MD}} - \Pi_C(v_{\text{MD}})\| = 0.91$).

The complementary component $\Pi_C(v_{\text{MD}})$ is nearly collinear with v_{conc} ($|\cos| = 0.95$), so the v_{conc} condition also serves as the test of whether the concreteness component alone suffices.

Aggregate effects. Table 14 reports literal and coherence rates. At the best operating point, v_{conc} reaches a literal rate of 0.43 (human-corrected 0.37; Appendix J.1) against 0.52 for v_{MD} , 0.12 unsteered, and 0.20–0.26 for random unit vectors on this model at matched magnitudes: a generic concreteness vector accounts for much of the aver-

age effect. It is not necessary, however: $v_{\perp}$ reaches 0.34 (corrected 0.32), retaining the majority of the effect after the concreteness subspace is removed.

Per-idiom dissociation. Figure 19 breaks the three conditions down by idiom (rates over coherent generations, pooled over the plateau coefficients, with Wilson 95% intervals). The per-idiom profiles of v_{conc} and $v_{\perp}$ correlate only weakly ($r = 0.25$), and five idioms are significantly separated between them. The pattern is interpretable: v_{conc} succeeds where the literal reading is a concrete physical scene (*spill the beans* 0.47, *break the ice* 0.61, *lie low* 0.70) and fails where literalness and concreteness come apart—*out of the blue* (0.07), whose literal reading requires reanalyzing the phrase rather than depicting a concrete scene, and *behind bars* (0.23, no better than its unsteered baseline of 0.28 on the same coherent-only basis), whose *figurative* reading is itself a concrete physical scene, so pushing toward concreteness cannot discriminate the readings. On both of these, $v_{\perp}$ succeeds (0.68 and 0.54). Each component therefore

Gemma-12B

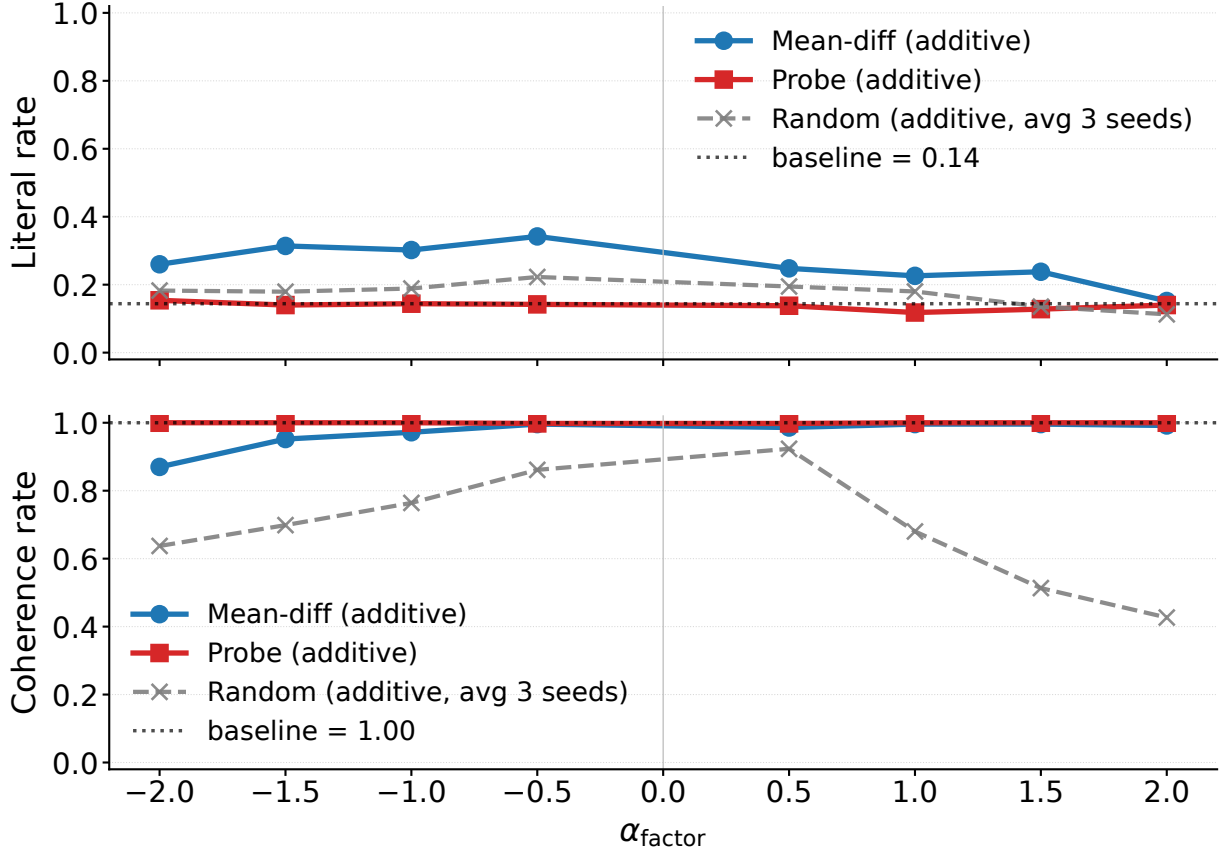


Figure 13: Full additive coefficient sweep on Gemma-3-12B. Top: literal rate vs α_{factor} . Bottom: coherence rate.

controls interpretation on idioms the other cannot.

Interpretation and scope. The mean-difference vector is best understood as a mixture: a general concreteness component that exploits the typical correlation between literal readings and concrete scenes, and an idiom-specific component that operates where that correlation breaks down. This sharpens the distinction from Ghosh and Jiang (2026) drawn in Section 2.1: interpretation selection is not reducible to generation style along the concreteness axis. Two scope notes: the behavioral decomposition is run on Llama-3.2-3B only, with geometry reported for all five models; and per-idiom rates use coherent-generation denominators, whereas aggregate rates follow the main-paper convention over all generations.

S Model Checkpoints

Table 15 lists the base-model checkpoints used for the steering experiments. All five models are evaluated in their pre-instruction-tuning form.

Auxiliary models. Claude 3.5 Sonnet was used only to propose candidate IdioSteer sentence prefixes during dataset drafting; all final examples were manually curated, edited, and validated by the authors. Gemma-4-31B-it was used as the LLM-as-a-judge after validation against human annotations (Appendix J).

T Reproducibility and computational cost

This appendix documents implementation details, hyperparameter choices, random seeds, and computational cost for reproducibility.

Hardware and inference setup. All experiments run on a single host with four NVIDIA RTX 3090 GPUs (24 GB each). Llama-3.2-3B, Llama-3.1-8B, Gemma-2-2B, and Gemma-3-4B each fit on a single GPU; Gemma-3-12B is loaded across two GPUs via HuggingFace `device_map="auto"`. We use HuggingFace transformers 4.55.0 with PyTorch 2.5.1 (CUDA 12.1 build). Llama-3.2-3B, Llama-3.1-8B, and Gemma-2-2B are loaded in fp16; Gemma-3-4B and Gemma-3-12B in bf16.

Gemma-12B

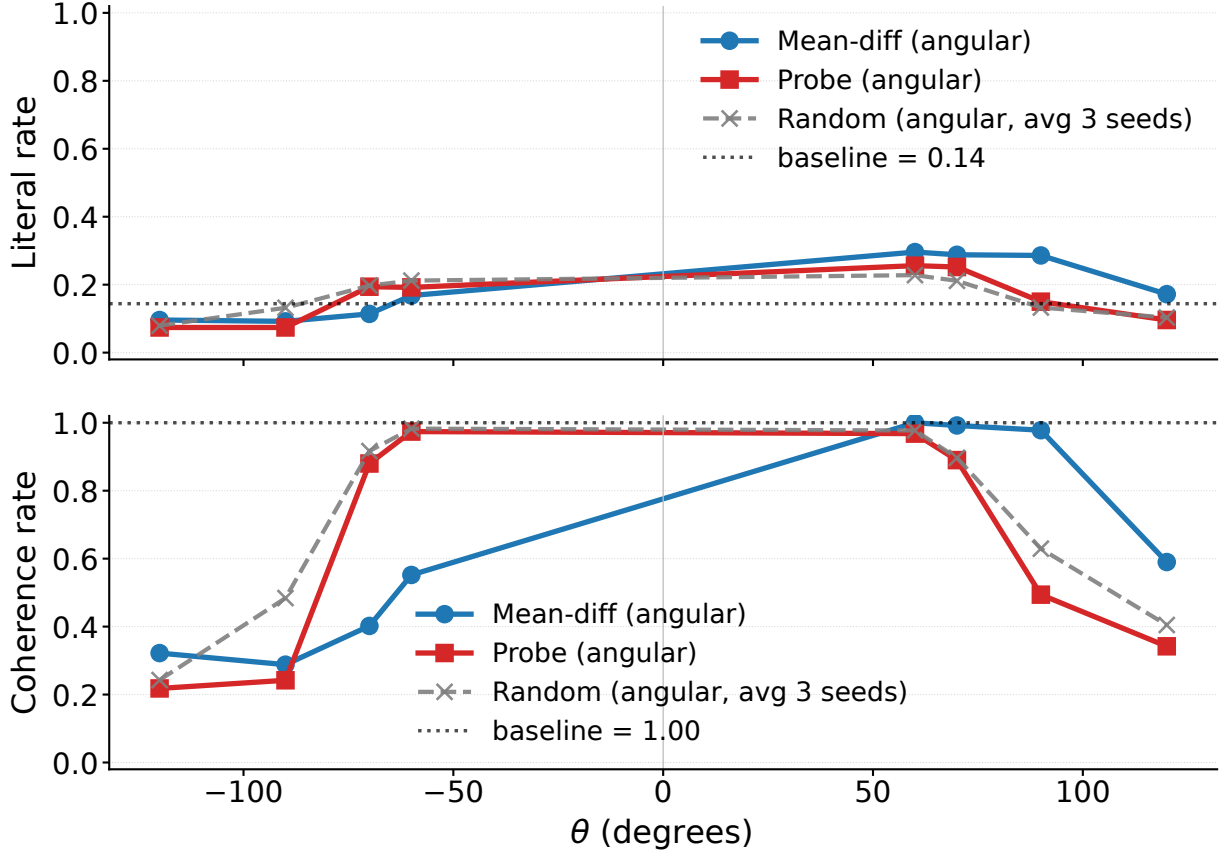


Figure 14: Full angular coefficient sweep on Gemma-3-12B. Top: literal rate vs θ . Bottom: coherence rate.

Attention implementation defaults are kept for the Llama and Gemma-2 families; the Gemma-3 family uses `attn_implementation="sdpa"`. KV-cache is enabled at inference (HuggingFace default). Inference batch size is 1 (sentence-by-sentence forward-and-generate loop).

Layer indexing. We refer to layers via the HuggingFace `output_hidden_states` convention: `hidden_states[0]` is the embedding output, and `hidden_states[N]` is the output of physical decoder block $N-1$. The source layer L_s uses this index for activation collection. The injection hook attaches to `model.model.layers[L_i]` via a forward hook on the full block output (post-attention + post-MLP residual stream). $L_i = 2$ therefore modifies the output of decoder block 2.

Activation and steering vector precision. Activations are collected in the model’s native dtype (fp16 or bf16) and cast to fp32 before being stacked into the design matrix. The mean-difference vector v_{MD} and the probe weight vector are computed in

fp32. At inference, the additive operator adds $\alpha \cdot v$ in the residual stream’s native dtype; the angular operator casts intermediate quantities through fp32 and back to fp16/bf16 before writing to the residual stream. Steering vectors are not persisted to disk; they are recomputed on every evaluation run from the same IdioLink training set, ensuring bit-for-bit reproducibility given the fixed seed.

Generation and random seeds. We set `torch.manual_seed(42)`, `numpy.random.seed(42)`, and `random.seed(42)` once at the start of each evaluation run; the seed is not re-set per condition. The 5 completions per sentence are produced by a sequential loop of 5 `generate()` calls drawing from the advancing RNG state (not `num_return_sequences=5`). This is deterministic across reruns given fixed hardware and library state; we do not enable `torch.use_deterministic_algorithms(True)`, so cuBLAS/SDPA nondeterminism can introduce small variation across distinct hardware. The three random unit-norm control

Model	Baseline	Mean-difference		Probe		Random	
		additive	angular	additive	angular	additive	angular
Llama-3.2-3B	0.116	0.460	0.275	0.336	0.238	0.259	0.201
Llama-3.1-8B	0.124	0.387	0.251	0.281	0.207	0.213	0.171
Gemma-2-2B	0.134	0.444	0.320	0.135 [†]	0.247	0.249	0.210
Gemma-3-4B	0.130	0.343	0.329	0.129 [†]	0.203	0.218	0.177
Gemma-3-12B	0.144	0.305	0.260	0.145 [†]	0.189	0.193	0.169

Table 8: Robustness across coefficients: mean literal rate over the literal-direction half of each model’s calibrated grid (four coefficients per learned method; twelve conditions for random, i.e. three vectors $\times$ four coefficients). Bold marks the best method per model; mean-difference additive is highest on every model, and exceeds baseline at every coefficient in its half-grid on every model. [†]: probe-additive’s half-grid mean is within 0.001 of the unsteered baseline on every Gemma model, consistent with the family-dependent collapse in Section 6.1. Best-over-grid values for the same conditions are in Figure 2.

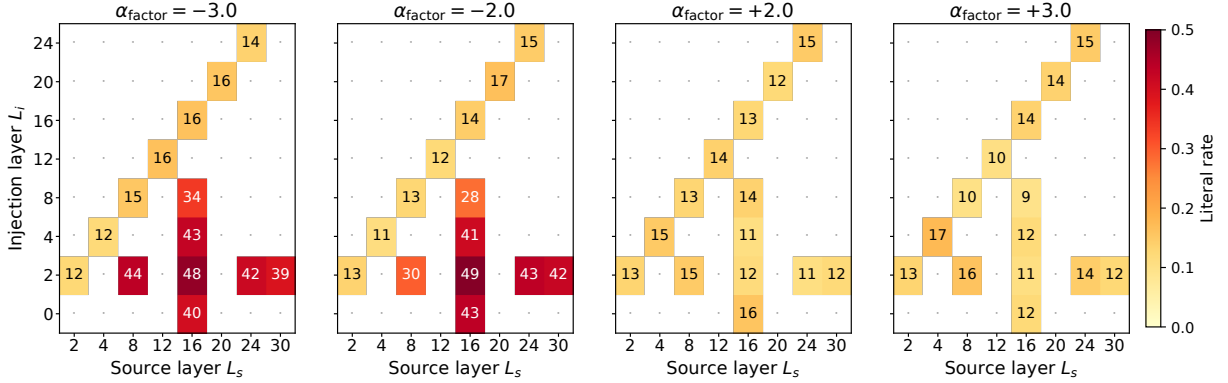


Figure 15: Literal rate over the evaluated (L_s, L_i) grid on Llama-3.1-8B at four multiplier values. Cell labels are literal rates $\times 100$; dotted cells denote layer pairs that were not evaluated. The optimal evaluated region clusters at low L_i and mid-to-late L_s under negative multipliers.

vectors (Section 4.1) are generated with `np.random.RandomState(s).randn(d)` for $s \in \{42, 123, 7\}$, yielding three different vectors per model (because d varies) under the same three seeds. The bootstrap procedure (Section 5, Appendix I) uses $B = 1000$ resamples with seed 42. The cross-fold shuffle (Section 6.4) uses seed 42; only a single shuffle was performed.

Probe hyperparameters. The probe uses scikit-learn 1.6.1 LogisticRegression with L2 regularization at $C = 1.0$ (the sklearn default), LBFGS solver, maximum 1,000 iterations, tolerance 10^{-4} , and `fit_intercept=True`. We did not perform a hyperparameter search over C , regularization type, or solver. Per-model classification accuracy is reported in Appendix D, Table 2.

Calibration. For each model, we manually select the source layer L_s by inspecting steered generations on a small subset of IdioLink training-pool idioms (three sentences per training idiom). Selection is qualitative: L_s is chosen to produce visibly literal completions without obvious fluency degra-

dation. The α_{factor} grid is then expanded around the selected layer to span a literal-rate plateau and a fluency-loss regime, yielding the eight α_{factor} values per model reported in Appendix H. The IdioSteer benchmark is never used during calibration.

Compute cost. The main evaluation grid (81 conditions $\times$ 500 completions per condition per model $\times$ 5 models = 202,500 generations) runs in approximately 85.7 GPU-hours: 5.96 h for Llama-3.2-3B, 9.24 h for Llama-3.1-8B, 9.27 h for Gemma-2-2B, 13.68 h for Gemma-3-4B, and 23.79 h $\times$ 2 GPUs for Gemma-3-12B. The cross-fold robustness evaluation (Section 6.4) adds 7.06 GPU-hours on Llama-3.2-3B. The layer-pair sweeps (Section 6.3) add approximately 13 GPU-hours across three models. Calibration runs across all models add an estimated further 10–20 GPU-hours not separately tabulated. Total documented compute is approximately 106 GPU-hours, with up to 20 GPU-hours additional calibration overhead. The LLM-as-a-judge inference uses Gemma-4-31B-it via the OpenRouter API at temperature

Method	α_{factor}	Lit.	Coh.	PPL ratio
baseline	—	0.12	1.00	1.00
mean-difference	-4.78	0.49	0.84	1.27
mean-difference	-2.39	0.52	0.95	1.10
mean-difference	+2.39	0.13	0.98	1.09
mean-difference	+4.78	0.13	0.97	1.14
probe	-2.39	0.37	0.95	1.06
probe	-4.78	0.34	0.79	1.15
random (avg.)	-4.78	0.25	0.86	1.17
random (avg.)	+4.78	0.26	0.91	1.10

Table 9: Behavioral effects and perplexity on Llama-3.2-3B at $L_s = 14$, $L_i = 2$. Literal rate (**Lit.**) and coherence rate (**Coh.**) are the judge-derived metrics defined in Section 5. Perplexity ratio (**PPL ratio**) is the token micro-perplexity of the steered generation under the unsteered model, normalized by the baseline perplexity (4.77). Random results are averaged over three independent unit-norm vectors.

Configuration	α_{factor}	rel	Literal	Coh.
Unsteered	—	0	0.116	1.000
Diagonal (14, 14)	-2.0	0.730	0.168	0.998
Diagonal (14, 14)	-3.75	1.369	0.179	0.992
Diagonal (14, 14)	-7.5	2.738	0.182	0.978
Diagonal (14, 14)	-15	5.476	0.360*	0.672
Diagonal (14, 14)	-30	10.952	0.406*	0.438
Cross-layer (14, 2)	-2.0	2.733	0.527	0.980

Table 10: Matched relative-strength check on Llama-3.2-3B. rel follows Equation 13; literal rates are over coherent generations. The two bolded rows are matched on rel to within 0.2% and on coherence to within 0.002. *: at these multipliers 32.8% and 56.2% of completions are labeled incoherent, so these two rates rest on a shrinking, selected subset (327 and 219 of 500).

0.0, $\text{max_tokens} = 700$; across all phases of the project, the judge was called approximately 358,700 times at a documented total cost of \$91.77.

Dataset splits. The IdioLink training pool contains 1,683 sentences across 22 idioms (629 figurative, 1,054 literal), drawn from the IdioLink train split after surface-form filtering (Appendix C). The cross-fold robustness analysis (Section 6.4) uses the union of the IdioLink train and validation idiom sets (32 idioms total) and partitions them into three disjoint folds of 11, 11, and 10 idioms. The IdioSteer evaluation benchmark contains 100 sentences across 20 idioms (five variants per idiom). The IdioSteer idiom set is disjoint from the IdioLink train and validation sets; this disjointness is asserted in code.

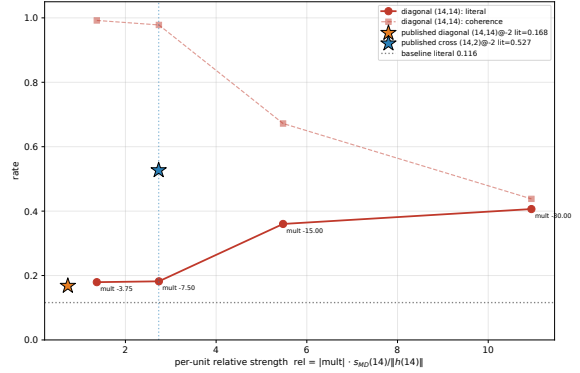


Figure 16: Literal and coherence rate for the same-layer diagonal ($L_s = L_i = 14$) on Llama-3.2-3B as a function of relative intervention strength. Stars mark the published cross-layer and diagonal operating points at $\alpha_{\text{factor}} = -2$; the dotted line marks the cross-layer relative strength.

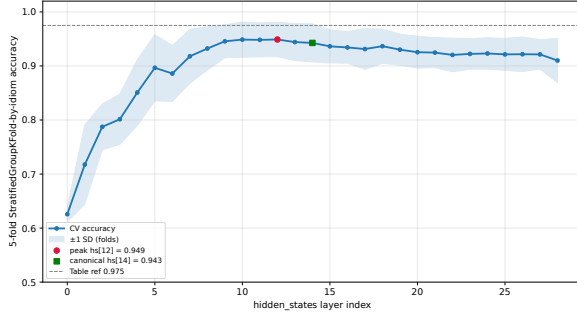
U Artifacts

Table 16 lists the datasets, software packages, and API providers used in this work. Model checkpoints are listed separately in Appendix S. Exact package versions are provided in the released repository.

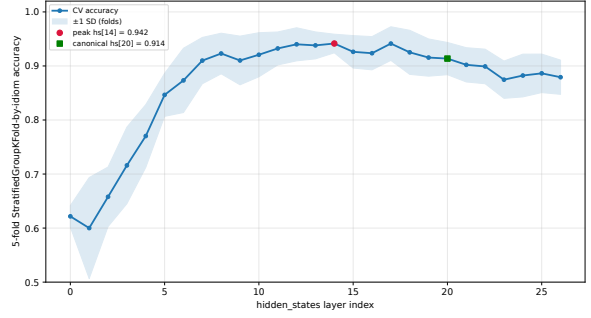
We use Hugging Face Transformers and PyTorch for model loading, activation extraction, and generation; scikit-learn for probe training and PCA analysis; and OpenRouter for LLM-as-a-judge API access. IdioSteer is newly created in this work and is released under a CC-BY-4.0 license. All external datasets, model checkpoints, and software artifacts are used for research purposes consistent with their licenses and terms of use.

V AI assistants

We used AI assistants (e.g., ChatGPT) to support code formatting, phrasing suggestions, and LaTeX styling during writing. Claude 3.5 Sonnet was used separately to propose candidate sentences during IdioSteer drafting, as described in Section 3 and Appendix B. All outputs were reviewed and edited by the authors. No content was directly generated or used without human verification.



(a) Llama-3.2-3B (28 layers, canonical $L_s = 14$).



(b) Gemma-2-2B (26 layers, canonical $L_s = 20$).

Figure 17: Per-layer decodability of the figurative-vs-literal distinction: 5-fold cross-validated probe accuracy, grouped by idiom so that test folds contain only unseen idioms, with ± 1 standard deviation across folds. Accuracy starts at the majority-class rate of the training pool, rises into a broad plateau in the mid-to-late network, and declines mildly at depth. Red markers indicate each model’s decodability maximum, green markers the calibrated source layer; the dashed line (panel a) is the stratified-CV accuracy from Table 2.

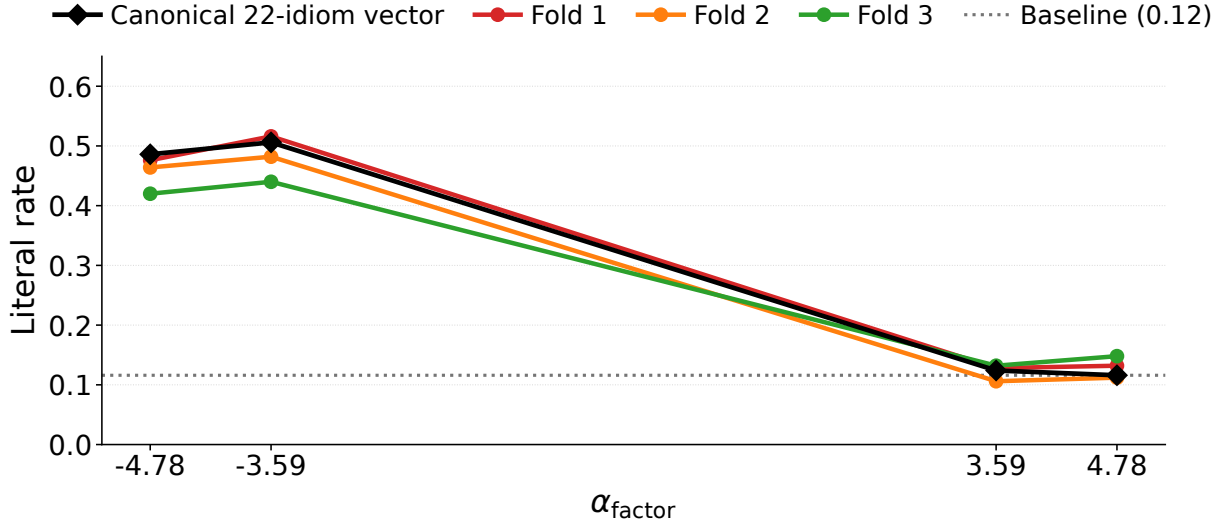


Figure 18: Cross-fold robustness on Llama-3.2-3B: literal rate vs α_{factor} for the canonical 22-idiom vector and three fold-derived vectors trained on disjoint 10–11-idiom subsets, all evaluated at $\alpha_{\text{factor}} \in \{\pm 3.59, \pm 4.78\}$. Fold curves track the canonical curve closely at the tested points.

Idiom	Llama-3.2-3B		Llama-3.1-8B		Gemma-2-2B		Gemma-3-4B		Gemma-3-12B	
	base	steer	base	steer	base	steer	base	steer	base	steer
shut the door on	0.12	0.68	0.08	0.68	0.16	0.84	0.20	0.72	0.28	0.68
have a ball	0.12	0.88	0.16	0.84	0.24	0.92	0.40	0.64	0.20	0.52
in the pink	0.00	0.64	0.12	0.44	0.20	0.64	0.00	0.56	0.20	0.56
pull someone’s leg	0.00	0.52	0.00	0.68	0.04	0.80	0.04	0.32	0.08	0.16
spill the beans	0.04	0.60	0.00	0.44	0.04	0.36	0.00	0.52	0.00	0.36
lie low	0.00	0.64	0.00	0.72	0.00	0.64	0.00	0.00	0.00	0.16
break the ice	0.08	0.60	0.20	0.72	0.16	0.68	0.16	0.32	0.00	0.36
see red	0.08	0.52	0.04	0.56	0.08	0.68	0.00	0.32	0.12	0.20
take a bow	0.04	0.40	0.08	0.44	0.04	0.16	0.00	0.56	0.08	0.40
on ice	0.04	0.48	0.08	0.32	0.04	0.68	0.00	0.28	0.00	0.08
in hot water	0.04	0.32	0.08	0.52	0.00	0.44	0.08	0.16	0.04	0.40
in stitches	0.12	0.56	0.08	0.24	0.12	0.60	0.20	0.52	0.24	0.40
green light	0.36	0.72	0.32	0.64	0.36	0.68	0.32	0.52	0.28	0.60
at the crossroads	0.24	0.60	0.16	0.36	0.32	0.60	0.20	0.40	0.20	0.48
out of the blue	0.00	0.36	0.00	0.44	0.00	0.36	0.00	0.04	0.00	0.08
rough edges	0.20	0.48	0.24	0.52	0.20	0.84	0.36	0.32	0.28	0.40
behind bars	0.28	0.36	0.16	0.44	0.16	0.44	0.16	0.28	0.20	0.32
up in the air	0.12	0.24	0.08	0.32	0.08	0.24	0.12	0.28	0.12	0.16
under the sun	0.04	0.16	0.04	0.28	0.08	0.24	0.00	0.04	0.00	0.08
in the fast lane	0.40	0.60	0.56	0.64	0.36	0.64	0.36	0.44	0.56	0.44

Table 11: Per-idiom literal rate at each model’s headline mean-difference additive operating point (“steer”) and unsteered baseline (“base”), over 25 generations per idiom per condition. Idioms are ordered by mean gain over baseline across the five models. 97 of the 100 cells improve on baseline; every idiom improves on at least four models.

Model	$\cos(v_{MD}, v_{conc})$		top-1 dim. share		unrelated contrast
	raw	whitened	v_{MD}	v_{conc}	cos with v_{MD} (raw)
Llama-3.2-3B	−0.394	−0.274	2.4%	2.8%	+0.28
Llama-3.1-8B	−0.324	−0.206	3.6%	5.5%	+0.27
Gemma-2-2B	−0.341	−0.166	8.3%	5.1%	+0.31
Gemma-3-4B	−0.697	−0.228	83.5%	39.3%	+0.79
Gemma-3-12B	−0.894	−0.219	95.0%	77.0%	+0.73

Table 12: Geometry of the idiom direction (v_{MD} , positive pole *figurative*) relative to the non-idiomatic concreteness vector (v_{conc} , positive pole *concrete*); a negative cosine indicates that figurative aligns with abstract. “top-1 dim. share” is the fraction of a vector’s squared norm carried by its single largest coordinate. The “unrelated contrast” column reports the raw cosine between v_{MD} and a semantically unrelated mean-difference vector (IdioLink subject categories, same activations): on the Gemma-3 models *any* two mean-difference vectors appear strongly aligned, so raw cosines are not comparable across models. On those models the dominant dimension is the *same* in both vectors (443 and 2339 respectively), so zeroing a single coordinate collapses the cosine to −0.39/−0.37; on the other models the top dimensions differ and no single coordinate drives the cosine. Whitened cosines are modest and uniform (−0.17 to −0.27) across all five models.

Model	v_{conc} (Brysbaert)		alt. construction		
	acc	acc _{whit}	split-half	acc	acc _{whit}
Llama-3.2-3B	0.997	0.997	0.980	0.946	0.942
Llama-3.1-8B	0.996	0.997	0.978	0.952	0.942
Gemma-2-2B	0.999	0.995	0.976	0.958	0.950
Gemma-3-4B	0.642	0.995	0.999	0.662	0.910
Gemma-3-12B	0.630	0.991	1.000	0.700	0.894

Table 13: Validation of the concreteness vectors. “acc” is held-out concrete-vs-abstract separability by projection onto the direction, under word-level splits; subscript “whit” uses per-dimension whitened activations. Split-half cosines for v_{conc} are 0.970, 0.968, 0.977, 0.981, 0.990 (top to bottom). Raw separability fails on the Gemma-3 models and recovers under whitening (see text).

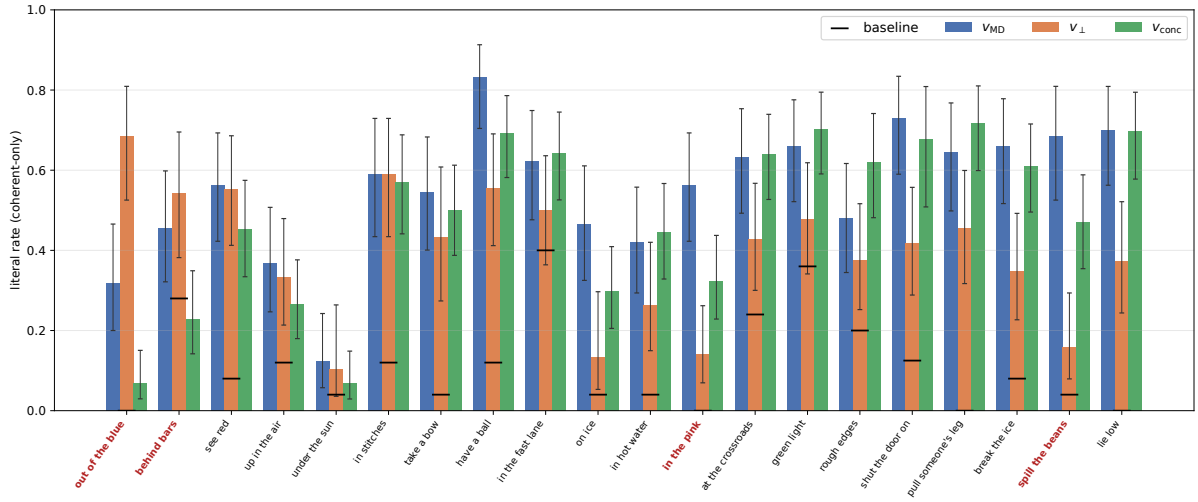


Figure 19: Per-idiom literal rates on Llama-3.2-3B for the idiom vector (v_{MD}), its concreteness-ablated component ($v_{\perp}$), and the pure concreteness vector (v_{conc}), pooled over the plateau coefficients with Wilson 95% intervals; rates over coherent generations. Dashes mark the unsteered baseline per idiom. Idioms are sorted by $v_{conc} - v_{\perp}$: the ablated idiom vector dominates on the left (*out of the blue*, *behind bars*), where the literal reading is not a more concrete scene, and concreteness steering dominates on the right (*lie low*, *spill the beans*), where it is. Idioms discussed in Section 6.4 and Appendix J.1 are highlighted.

Condition	α	Literal	Corrected	Coherence
Unsteered	—	0.116	—	0.998
v_{MD}	−10	0.518	0.535	0.956
v_{conc}	+10	0.432	0.370	0.928
$v_{\perp}$	−10	0.344	0.320	0.898
Random (additive)	—	0.259	—	0.938

Table 14: Literal and coherence rates on IdioSteer (Llama-3.2-3B, raw additive coefficient α ; all vectors unit-norm, so intervention magnitudes are comparable). Rates follow the main-paper convention (all generations in the denominator). “Corrected” applies the human-audit confusion matrices (Appendix J.1) to the judged label distribution at the same operating point; pooling over the plateau coefficients instead gives 0.362 for v_{conc} and 0.309 for $v_{\perp}$, which are the more precisely estimated values. Note the sign convention: v_{conc} steers toward literal at *positive* α , since its positive pole is concrete. The random row is the mean over the literal-direction half-grid ($\alpha \in \{-5, -10, -15, -20\}$) of three random unit vectors (12 conditions, 6,000 generations), rather than a single operating point; its literal and coherence rates use the same all-generations denominator as the rows above.

Model	Checkpoint	License	# Param.
Llama-3.2-3B	meta-llama/Llama-3.2-3B	Llama 3.2 Community License	3B
Llama-3.1-8B	meta-llama/Llama-3.1-8B	Llama 3.1 Community License	8B
Gemma-2-2B	google/gemma-2-2b	Gemma Terms of Use	2B
Gemma-3-4B	google/gemma-3-4b-pt	Gemma Terms of Use	4B
Gemma-3-12B	google/gemma-3-12b-pt	Gemma Terms of Use	12B

Checkpoints correspond to the public pretrained/base variants used for continuation generation and activation extraction. Instruction-tuned variants are not used in the main steering experiments.

Table 15: Base-model checkpoints used in the steering experiments.

Artifact	Type	License / Terms	Usage
IdioSteer	Dataset	CC-BY-4.0	Evaluation benchmark introduced in this work.
MAGPIE ^a	Dataset	CC-BY-4.0	Source idiom inventory for selecting IdioSteer idioms.
IdioLink ^b	Dataset / Code repository	Apache-2.0	Training pool used to construct steering vectors.
HuggingFace Transformers ^c	Software library	Apache-2.0	Model loading, activation extraction, and generation.
PyTorch ^d	Software library	BSD-style license	Tensor computation, model execution, and forward hooks.
scikit-learn ^e	Software library	BSD 3-Clause	Logistic-regression probes, PCA analysis, and cross-validation.
OpenRouter ^f	API provider	Proprietary terms	API access for LLM-as-a-judge evaluation.

^a <https://github.com/hslh/magpie-corpus>

^b <https://github.com/Intellexus-DSI/IdioLink>

^c <https://github.com/huggingface/transformers>

^d <https://github.com/pytorch/pytorch>

^e <https://github.com/scikit-learn/scikit-learn>

^f <https://openrouter.ai/terms>

Table 16: Datasets, software packages, and API providers used in this work. Model checkpoints are reported separately in Table 15.