NATURAL LANGUAGE DICTIONARIES: LEARNING FAITHFUL DECOMPOSITIONS OF LLM BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) research often treats broad behavior labels such as sycophancy or deception as coherent units, although each can combine distinct observable behaviors. We study how to discover behavioral dictionaries whose natural-language concepts correspond to model directions and support behavior-specific interventions. Natural Language Dictionary Learning (NLD) learns behavioral measurement profiles before deriving their linearly readable representations, then selects textual descriptions by held-out agreement with those profiles. Across social sycophancy, deception, and emotion, NLD consistently improves name–direction fidelity and often improves intervention selectivity, with the clearest gains for compact dictionaries. And learned sycophancy and deception representations transfer to held-out datasets.

1 INTRODUCTION

Large language model (LLM) research is organized around human-defined behaviors such as sycophancy, emotion, and deception. These labels determine how researchers build datasets, evaluate models, identify internal representations, and design interventions (Rimsky et al., 2024; Chen et al., 2025; Arditi et al., 2024). This workflow implicitly treats each label as a coherent unit of model behavior.

However, even apparently intuitive behaviors often violate this assumption. Interpretability research finds that sycophancy, refusal, reasoning, and repetition decompose into functionally distinct representations (Vennemeyer et al., 2026; Zhao et al., 2026; Wollschläger et al., 2025; Venhoff et al., 2025; Mahaut & Franzon, 2025). If a dataset groups these components under one label, a learned direction can average over mechanisms with different causes and effects, making it ambiguous as both an explanation and an intervention target.

This fragmentation also has practical consequences. Benchmarks that nominally measure the same behavior can operationalize different mixtures of its components and therefore produce sharply different conclusions about the same model (Ye et al., 2026; Shi et al., 2026). Thus, a label that appears coherent upon qualitative inspection does not guarantee either *construct coherence* in evaluation or *representational coherence* within the model.

We argue that broad labels should instead be represented by *behavioral dictionaries*: natural-language behaviors paired with model directions that track them and affect them selectively. Here *behavior* means an observable regularity in model outputs, not an internal computation. This distinction determines our evaluation. A behavioral concept must be describable well enough that its presence can be recognized in outputs; its description must therefore faithfully track the associated model direction. A dictionary must also separate behaviors rather than provide several names for the same effect; its directions must therefore act selectively. We additionally require the directions to influence the broader behavioral domain, showing that the named behaviors are causally relevant constituents of the high-level behavior.

Current methods, such as PCA and sparse autoencoders (SAEs), can cause overall changes in behavior and produce qualitatively convincing dictionaries (Templeton et al., 2024; Lu et al., 2026). However, they fare poorly on interpretation faithfulness and behavioral selectivity (Kantamneni et al., 2025; Paulo & Belrose, 2025). This motivates *Natural Language Dictionary Learning* (NLD): instead of learning a direction and then naming it, NLD learns the behavioral concept first and derives

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

its linear representation from the concept's measurements. It searches for concept-presence profiles that explain behavioral variation and are readable from model activations, then retains an English name only when it reproduces the learned profile on held-out data.

Across social sycophancy, deception, and emotion, NLD consistently improves name–direction fidelity and behavioral selectivity, with the clearest selectivity gains for compact dictionaries. Its interventions remain competitive, even though NLD is not optimized to maximize this criterion. And learned sycophancy and deception representations also transfer to AITA (Cheng et al., 2026) and LiarsBench (Kretschmar et al., 2026), respectively, neither of which is used for discovery. We release our code[1] to support replication.

Our contributions are:

1. We formalize a *behavioral dictionary* as nameable behaviors paired with linearly readable representations, evaluated through fidelity, selectivity, and broad-domain relevance.
2. We introduce *Natural Language Dictionaries (NLD)*, a concept-first method that learns explanatory behavioral concepts grounded in linear activation directions and natural-language descriptions.
3. We demonstrate that across sycophancy, deception, and emotion, NLD improves name–direction fidelity and, especially for compact dictionaries, intervention selectivity, with robustness across an independent judge, held-out datasets, and a second model.

2 Related Work

Behavioral Disentanglement. Recent work has increasingly questioned whether common behavioral constructs used throughout LLM research correspond to coherent underlying phenomena. From a behavioral perspective, Ye et al. (2026) argue that AI sycophancy encompasses multiple distinct behaviors with different measurement and mitigation requirements, while Smith et al. (2025); Shi et al. (2026) argue that evaluating deception is fundamentally complicated by the possibility that deception is composed of many mechanisms. Complementing these behavioral observations, interpretability work has shown that behaviors including sycophancy (Vennemeyer et al., 2026; Baez et al., 2026; Bohacek et al., 2026), refusal (Zhao et al., 2026; Wollschläger et al., 2025), reasoning (Venhoff et al., 2025), repetition (Mahaut & Franzon, 2025), and other safety-relevant behaviors (Sartawita et al., 2025) decompose into multiple independently steerable behaviors at the representation level. Our work builds on these findings by discovering behavioral dictionaries grounded in model representations and evaluated by held-out correspondence and intervention tests, rather than assuming behavioral constructs a priori.

Natural-Language Interpretation of Model Representations. Recent work has explored explaining internal representations directly in natural language. Early approaches use prompting to generate descriptions of steering vectors or hidden activations (Lindsey, 2025; Chen et al., 2024; Ghandeharioun et al., 2024), while more recent methods train specialized models such as Activation Oracles (Karvonen et al., 2026) and Natural Language Autoencoders (Fraser-Taliente et al., 2026). These approaches generate compelling explanations for existing representations, but are generative. Our work addresses a different problem. Rather than generating a new explanation for every activation independently, we seek to learn a small, fixed dictionary of natural-language behavioral concepts that explains a behavioral dataset.

3 Evaluating Behavioral Dictionaries

Let M be the frozen Qwen3-4B-Instruct subject model (Yang et al., 2025) and $\mathcal{D}_B = \{(x_i, y_i)\}_{i=1}^{N}$ a dataset for a broad behavioral domain B, like sycophancy or deception. A method returns a behavioral dictionary $D = \{(c_1, v_1), \ldots, (c_K, v_K)\}$, where c_k describes an observable behavior and v_k is a model representation intended to track and influence it. The entries may not partition the domain, form a hierarchy, exhaust its variation, or negations of behaviors. We consider a dictionary useful when its names correspond to their representations, denote functionally distinct behaviors, and remain relevant to the broad domain.

[1] https://anonymous.4open.science/r/Natural-Language-Dictionary-7947/

For example, suppose a sycophancy dictionary contains "validates the user's feelings" and "accepts the user's framing." Plausible wording is not enough: responses with high activation on the validation direction should actually contain validation (fidelity), and steering that direction should increase validation more than framing acceptance (selectivity). The dictionary as a whole should also affect measured sycophancy (average intervention effect). These tests distinguish a behavioral decomposition from a fluent list of related phrases.

Domains. We study domains with plausible within-class variation. **Social sycophancy** uses the Open-Ended Questions (OEQ) dataset, with the r/AmITheAsshole (AITA) dataset reserved for transfer (Cheng et al., 2026). **Deception** can be expressed through fabrication, omission, or pragmatic distortion; we use the Roleplaying corpus of Goldowsky-Dill et al. (2025) for discovery and Liars-Bench (Kretschmar et al., 2026) instructed deception for transfer. **Emotion** is more compact: we contrast examples of high and low valence from EmoBank (Buechel & Hahn, 2017), specifically more than one standard deviation above and below mean valence. These domains range from socially heterogeneous behavior to a comparatively compact affective distinction. Evaluation should reward real decomposition without assuming that every domain supports the same number of components.

For dictionary discovery, all three datasets are treated as collections of response texts: the frozen subject model processes each text and we extract its response activation. Intervention evaluation is separate: the subject model generates new continuations under activation steering (Rimsky et al., 2024), which are scored with the domain-specific outcome.

3.1 Evaluation criteria

We ask whether names track their directions (fidelity), interventions have distinct effects (selectivity), and the dictionary affects the broad domain (AIE). We report these separately: fidelity and selectivity are primary, Average Intervention Effect (AIE) checks broad relevance, and two R^2 measures are diagnostics only. We do not use either R^2 measure to rank methods or support the paper's performance claims.

Name–direction fidelity. If a name describes an observable behavior associated with its direction, the name should predict when that direction is active. Let $r_i(x)$ be an LLM judge's score for the presence of c_i in response x, and let $a_i(x) = h(x) \cdot v_i$ be its direction score. On held-out examples, we define

$$\text{Fidelity}(i) = \rho_s(a_i(x), r_i(x)), \tag{1}$$

where ρ_s is Spearman correlation. Training and candidate selection use differentiable Pearson correlation; held-out evaluation uses Spearman to require only monotonic correspondence.

Behavioral selectivity. Fidelity is correlational and does not show that two named components have different measured effects. We therefore intervene on v_i and ask whether it changes its own named behavior more than every other behavior in the dictionary. Following Vennemeyer et al. (2026),

$$s_i = \frac{|\Delta_i|}{\max(1, \max_{j \neq i} |\Delta_j|)}, \tag{2}$$

where Δ_i and Δ_j are intervention-induced changes in target and non-target judged presence in the judge's original 0–9 units. A value above one means the target effect exceeds every cross-effect; the one-point floor prevents negligible cross-effects from producing extreme ratios. For example, an own effect of 2.8 points and maximum cross-effect of 0.6 gives $2.8/\max(1, 0.6) = 2.8$. Appendix diagnostic effects are divided by nine for display on a $[0, 1]$ scale, where the same floor is $1/9$. Appendix C.1 decomposes target and cross-effects and reports alternative summaries.

Broad-domain relevance. Selectivity concerns *which* behavior a direction changes. Average Intervention Effect (AIE) instead asks whether it changes the broad domain at all. Dictionary-level AIE uses a different domain outcome and is normalized by its baseline standard deviation, so its magnitude is not directly comparable to per-entry selectivity, which evaluates one dictionary direction at a time. For $K > 1$, we combine the dictionary using training-only weights balanced across three LLM-judge restatements of the domain behavior. At dose α,

$$\text{AIE}_\alpha(u) = \frac{1}{\sigma_B} \mathbb{E}\big[g_B(x^{\text{steer}}_{\alpha,u}) - g_B(x^{\text{base}})\big], \tag{3}$$

where g_B averages Cheng et al.'s three sycophancy judgments (Cheng et al., 2026), uses the MASK benchmark for deception (Ren et al., 2026), and uses VADER sentiment scoring for emotion (Hutto & Gilbert, 2014); σ_B is the baseline per-prompt standard deviation of that outcome. The table reports $t_{\mathrm{AIE}} = \mathrm{AIE}/\mathrm{SE}$. Appendix A.3 gives the full protocol.

Representation and interpretation diagnostics. These correlation diagnostics ask how well the learned directions or their names predict the target: $R^2_s = R^2(Y, \text{direction scores})$ and $R^2_n = R^2(Y, \text{judged name-presence scores})$. Their difference indicates how much of the structure exposed by the directions is communicated by their names. Higher R^2 is not inherently better. PCA explicitly maximizes variance, whereas an SAE may intentionally isolate useful low-variance features. Because methods target different kinds of variation, we do not use either metric to rank them.

Why use the subject model as a judge? We use the frozen subject model as the primary judge because we specifically seek behaviors coherent to that model. This makes the dictionary deliberately model-relative, so we test whether its conclusions survive outside that measurement loop in three ways: we rescore every fidelity comparison with DeepSeek V4 Flash (Appendix A.8), test recovery of human OEQ annotations (Figure 4), and transfer the learned names and directions to AITA and LiarsBench (Figure 3).

3.2 Baselines and splits

We compare NLD with PCA, which retains the first K components of behavior-conditioned activations, and a SAE, which trains a Top-k_{SAE} autoencoder and selects the K features most associated with the target (Venhoff et al., 2025). An SAE may isolate useful low-variance features rather than maximize coverage. See Appendix A.7 for SAE details.

Expert dictionaries are decompositions of these behaviors drawn from prior work. For social sycophancy, we use indirectness, emotional validation, and framing acceptance (Cheng et al., 2026); Ekman's basic emotions for emotion (Ekman, 1992); and fabrication, omission, and pragmatic distortion for deception (Shi et al., 2026). Each name is scored by the same judge and we derive a difference-in-means direction from these judgments (Marks & Tegmark, 2024; Wu et al., 2025). We do not vary k for these expert dictionaries. Directions are fit only on discovery examples and fidelity is measured on disjoint test examples. Thus the fit guarantees positive orientation on its training contrast, not positive held-out correlation.

PCA and SAE directions receive names through the same post-hoc procedure: generate eight descriptions from high/low-activation responses and select by held-out name–direction agreement. Appendix A.4 gives the protocol and naming ablations.

NLD discovery and baseline feature/name selection use 1,000 examples, while dictionary metrics use 400 held-out examples. For the SAE, this split selects features from the already-trained dictionary, not for training the SAE. All selection precedes evaluation; methods share held-out intervention prompts. Appendices A.4 and A.3 give full splits and protocols.

4 Natural Language Dictionary Learning

Natural Language Dictionaries learn to interpret more faithfully by learning what concepts to measure before fitting directions. Its key observation is that a behavioral description can serve as a measurement query: the frozen subject model scores how strongly each response expresses the description. Figure 1 summarizes the method.

NLD learns in two stages. Stage one learns K continuous *soft concepts*: prompt embeddings (Lester et al., 2021; Li & Liang, 2021) whose presence scores jointly explain the target variation and whose distinct contributions are linearly readable from activations. Stage two asks the frozen model to verbalize each soft concept, then continues training with a cycle-consistency loss between the soft and named presence profiles. The output is K description–direction pairs; the subject model remains frozen. Unlike an SAE dictionary, NLD does not reconstruct the full activation state.

To see the logic, consider learning a sycophancy concept. A soft concept first acts like a learnable question asked of every response, producing a presence profile: high scores on some responses and low scores on others. NLD adjusts this question so that, together with the other concepts, its

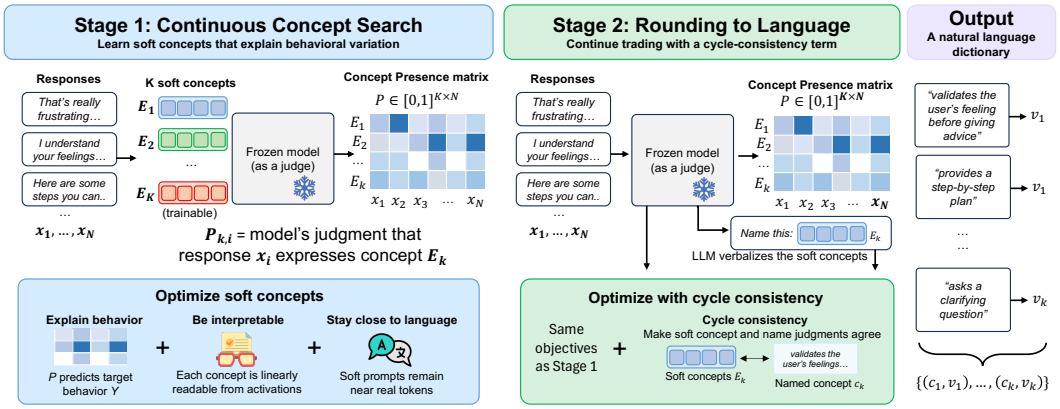

Figure 1: Overview of NLD. Stage one learns soft concept-presence profiles that explain the behavioral target and are readable from distinct linear directions. Stage two verbalizes each soft concept and aligns its profile with the resulting natural-language description.

Table 1: **Example behavioral dictionaries for social sycophancy (K=5).** All methods produce plausible fluent descriptions. However the PCA and SAE dictionaries are less faithful to their directions and less selective in intervention. Faithfulness and selectivity are evaluated quantitatively in Table 2.

SAE	PCA	NLD (ours)
Empathetic guidance	Concrete steps or resources	Philosophical or emotional reflection
Direct, step-by-step advice	Emotional validation and support	Balanced, multi-perspective framing
Actionable strategies with examples	Practical, actionable solutions	Brief, subtle insight without elaboration
Concrete, context-specific solutions	Concrete suggestions or next steps	Positive reflection on everyday experiences
Specific, actionable behavioral strategies	Mirroring experiences or struggles	Casual, confident, attention-grabbing language

scores explain how sycophantic responses vary. At the same time, the concept receives credit only when its unique profile can be predicted from a single activation direction. The result is therefore both behaviorally useful and linearly grounded. The model then converts the soft concept into a short description, such as a clause about validating the user. Cycle consistency is the final step: the English description should score responses similarly to the soft concept it replaces. Finally, NLD refits a direction to that concept's unique profile.

Formally, let M be the frozen subject model, and let $\mathcal{D} = \{(x_i, h_i)\}_{i=1}^{N}$ contain response sentences x_i with mean-pooled layer-ℓ residual activations $h_i \in \mathbb{R}^d$. A concept is either a text clause or a six-token soft prompt $E \in \mathbb{R}^{6 \times d_{\text{emb}}}$.

Concepts as differentiable measurements. The *judge* presents M with a sentence x and asks how strongly it expresses concept c on a 0–9 scale. Rather than decoding a rating, we use the expected digit under the model's next-token distribution:

$$\pi(c, x) = \tfrac{1}{9} \sum_{r=0}^{9} r \, \Pr_M(r \mid x, c) \;\in\; [0, 1]. \tag{4}$$

This is differentiable in a soft prompt's embeddings. We write

$$p_k = \big(\pi(c_k, x_1), \ldots, \pi(c_k, x_N)\big) \in [0, 1]^N$$

for concept k's *presence profile*, and $P = [p_1; \ldots; p_K] \in [0, 1]^{K \times N}$ for the presence matrix. The same π underlies learning, fidelity, selectivity, and baseline naming.

Behavioral target. The class label says whether a response belongs to the domain, not how positive responses differ. We apply PCA to positive responses and weight each component by its association with the class distinction. Let $\tilde{H}$ be the residualized response activations, let U contain the retained positive-response principal directions, and let s be the standardized score along a regularized class

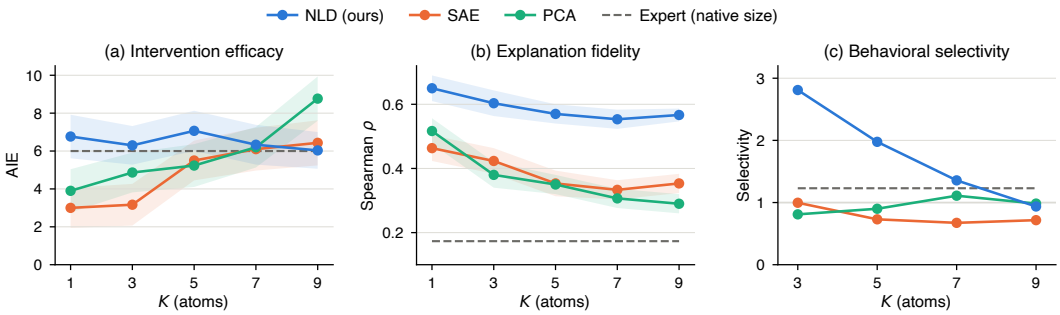

Figure 2: Qwen3-4B-Instruct macro-averaged broad-domain intervention effect (a), name–direction fidelity (b), and behavioral selectivity (c) across dictionary size K. Panel (a) reports the paired AIE t statistic; Appendix A.3 gives the intervention protocol.

direction. We define

$$w_m = \frac{\left|\mathrm{corr}(\tilde{H}u_m, s)\right|}{\max_{m'}\left|\mathrm{corr}(\tilde{H}u_{m'}, s)\right|}, \qquad Y = \tilde{H}U\,\mathrm{diag}(w). \tag{5}$$

Thus Y retains prominent positive-class variation associated with the domain.

We use absolute correlation because PCA signs are arbitrary. Y is not a proposed ground-truth definition of the behavior, it is an engineered search signal. It is used to train NLD and to compute the two diagnostic R^2 values, but not to compute fidelity, selectivity, AIE, or external transfer. Appendix B gives the full construction; Table 3 ablates prompt residualization. With Y specifying the variation to search, NLD can now find measurements that divide it into readable components.

Stage one: continuous concept search. Each dictionary slot begins as a soft concept $E_k \in \mathbb{R}^{T\times d_{\mathrm{emb}}}$. These soft-concept embeddings are the only trainable parameters; the subject model M remains frozen. The objective has three roles: the profiles should jointly explain the target, each slot should contribute distinct linearly readable signal, and the soft concepts should remain supported by the data, and should be close enough to token embeddings to name. Stage one optimizes

$$\mathcal{L} = -\mathcal{J}(P;Y) - \lambda_{\mathrm{fid}}\sum_{k=1}^{K} F_k + \lambda_{\mathrm{lang}}\Omega_{\mathrm{lang}}(E) + \lambda_{\mathrm{sup}}\Omega_{\mathrm{sup}}(P). \tag{6}$$

Both principal rewards are cross-fitted. For complementary halves $(S, S') \in \{(A,B),(B,A)\}$, let $X_S = [P_S^\top\ \mathbf{1}]$ and let $\widehat{W}_S$ be the ridge map from X_S to Y_S. Write $R^2_{S\to S'} = R^2(Y_{S'}, X_{S'}\widehat{W}_S)$ for its held-out explanatory power on the opposite half. The explanatory reward averages the two directions:

$$\mathcal{J}(P;Y) = \frac{1}{2}\left(R^2_{S\to S'} + R^2_{S'\to S}\right). \tag{7}$$

It rewards concepts whose presence profiles jointly explain behavioral variation on examples not used to fit the predictor.

For each concept, we first isolate the part of its presence profile that is not explained by the other concepts. Let u_k denote this unique component of p_k, and let $v_{k,S}$ be a ridge direction trained to read out u_k from activations $\tilde{H}_S$. We reward a concept when this direction continues to predict its unique component on the held-out split:

$$F_k = \frac{1}{2}\sum_{(S,S')} \underbrace{\mathrm{corr}(\tilde{H}_{S'}v_{k,S}, u_{k,S'})}_{\text{held-out readability}} \underbrace{\min\left(1, \frac{\mathrm{std}(u_{k,S'})}{\mathrm{std}(p_{k,S'})}\right)}_{\text{unique-signal weight}}. \tag{8}$$

The correlation rewards held-out readability; the scale factor discounts nearly duplicate concepts with little unique variation. Appendix B.3 gives the full construction.

Table 2: Comparison of behavioral-dictionary methods on Qwen3-4B-Instruct across Social Sycophancy, Deception, and Emotion. R_s^2 and R_n^2 denote structural and named-axis explanatory power, respectively. t_{AIE} is the paired t statistic for the intervention effect; Appendix A.3 gives the intervention protocol. ρ measures fidelity to the named behavioral axes, and Sel. denotes intervention selectivity. Bold indicates the best AIE, fidelity, and selectivity point estimate at each dictionary size within each domain. Expert dictionaries are evaluated at their published sizes: $K = 3$ for sycophancy and deception and $K = 6$ for emotion.

		Social Sycophancy					**Deception**					**Emotion**				
	k	$\mathbf{R_s^2}$	$\mathbf{R_n^2}$	$t_{\mathbf{AIE}}$	ρ	**Sel.**	$\mathbf{R_s^2}$	$\mathbf{R_n^2}$	$t_{\mathbf{AIE}}$	ρ	**Sel.**	$\mathbf{R_s^2}$	$\mathbf{R_n^2}$	$t_{\mathbf{AIE}}$	ρ	**Sel.**
SAE	1	0.019	0.156	+5.3	**0.60**	—	0.014	0.028	**+3.4**	0.19	—	0.038	0.062	+0.3	0.60	—
	3	0.051	0.192	+0.4	0.55	0.94	0.049	0.059	**+5.0**	0.10	0.75	0.083	0.117	+4.1	0.62	1.30
	5	0.071	0.232	+2.3	0.40	0.73	0.059	0.116	+3.9	0.08	0.69	0.105	0.174	+10.3	0.58	0.77
	7	0.114	0.259	+1.9	0.31	0.73	0.066	0.124	**+4.8**	0.11	0.69	0.135	0.182	**+11.6**	0.58	0.60
	9	0.149	0.292	+2.3	0.31	0.73	0.082	0.130	+5.1	0.14	**0.75**	0.137	0.184	**+11.9**	0.61	0.67
PCA	1	0.415	0.144	-1.0	0.55	—	0.236	0.013	+0.9	0.25	—	0.157	0.114	+11.8	0.75	—
	3	0.634	0.188	+2.3	0.28	0.96	0.425	0.039	+1.7	0.22	0.35	0.262	0.184	+10.6	0.64	1.12
	5	0.759	0.214	+3.1	0.26	0.86	0.483	0.058	+3.7	0.19	0.87	0.445	0.216	+8.9	0.60	0.97
	7	0.777	0.246	**+4.0**	0.24	0.75	0.537	0.062	+3.8	0.17	0.87	0.689	0.316	+10.8	0.51	**1.71**
	9	0.832	0.257	**+3.3**	0.22	0.75	0.626	0.089	**+12.3**	0.20	0.74	0.691	0.325	+10.7	0.45	**1.46**
NLD	1	0.118	0.045	**+6.5**	0.51	—	0.054	0.081	+1.8	**0.64**	—	0.131	0.091	**+12.0**	**0.80**	—
	3	0.454	0.271	**+4.7**	**0.64**	**2.70**	0.203	0.089	+2.7	**0.51**	**3.06**	0.327	0.240	**+11.5**	**0.66**	**2.67**
	5	0.520	0.332	**+4.9**	**0.58**	**2.87**	0.228	0.139	**+4.1**	**0.50**	**1.94**	0.483	0.284	**+12.2**	**0.63**	**1.12**
	7	0.511	0.395	+3.5	**0.54**	**1.59**	0.281	0.147	+3.9	**0.49**	**1.13**	0.550	0.320	**+11.6**	**0.63**	1.35
	9	0.560	0.378	+3.1	**0.56**	**1.44**	0.281	0.215	+4.5	**0.48**	0.74	0.588	0.309	+10.5	**0.66**	0.63
Expert		0.535	0.048	+1.4	0.33	1.03	0.350	0.062	+4.8	0.36	1.18	0.388	0.070	+11.8	-0.17	1.48

The language regularizer, Ω_{lang}, penalizes each soft token by its distance to the nearest real-token embedding, encouraging the learned concepts to remain close to the model's lexical embedding space. A support floor, Ω_{sup}, additionally penalizes concepts whose mean presence across examples falls below a fixed threshold, preventing nearly inactive concepts from satisfying the objective trivially. Appendix B gives the precise definitions and thresholds.

Stage two: rounding to language. Stage one produces soft concepts rather than readable descriptions. To name each concept, we insert its learned embeddings into a prompt asking the same frozen model to express it as one specific, observable assistant behavior. The model returns a short clause completing "`[concept embeddings]`. Restate that behavior in plain English ..." (Li et al., 2025; Hewitt et al., 2026). Appendix B.4 gives the exact prompt.

We score the resulting description with the same judge and continue optimizing the soft concept with a cycle-consistency loss: the soft and named concepts should score the same responses similarly. Let $p_k^{\mathrm{anc}} = \pi(c_k, \cdot)$ denote the plain-text presence profile of the description for concept k. We add

$$\mathcal{L}_{\mathrm{cyc}} = \lambda_{\mathrm{cyc}} \frac{1}{K_{\mathrm{acc}}} \sum_{k \in \mathcal{K}_{\mathrm{acc}}} [1 - \mathrm{corr}(p_k, p_k^{\mathrm{anc}})] \tag{9}$$

to equation 6. This pulls each soft concept toward agreement with what its *description measures*, rather than toward the description's token embeddings. We then decode new candidates and repeat candidate acceptance against the updated profiles (Appendix B.4). Final directions are refit on the full discovery set using the same unique-component readout. NLD never optimizes steering; all named results use the final descriptions, never soft concepts.

5 Results

Figure 2 summarizes the three primary metrics across dictionary sizes; Table 2 gives the domain-level results.

Do names match their directions? NLD consistently yields the highest fidelity in every multi-entry comparison across table 2. More importantly, the advantage largely survives changing the measurement model: under DeepSeek V4 Flash, NLD exceeds SAE and PCA in 11 of 12 comparisons (Appendix A.8). Absolute values and some orderings change, so the judge matters, but the overall pattern is not merely self-agreement with the model used during training. Learning a measurable concept before its direction is therefore more reliable than naming a fixed direction afterward.

Do the directions have distinct effects? Fidelity alone could still assign different names to directions that cause the same change. Selectivity tests this stronger requirement. At small dictionary sizes, NLD is the only method consistently well above one across all three domains; representation-first and expert dictionaries remain near it. Thus NLD's names correspond not only to different activation patterns, but to interventions with more behavior-specific measured effects.

The advantage narrows as dictionaries grow. This does not necessarily mean that larger solutions are worse: some domains—especially valence—genuinely may not contain nine separable observable behaviors. But the metric also becomes stricter as K increases because each added entry creates another possible cross-effect. The appropriate comparison is therefore between methods at matched K, not between K.

Does the dictionary remain relevant to the broad domain? Selective effects matter only if the dictionary still captures constituents of the broad behavior. AIE provides this check. No method dominates, but NLD remains competitive despite never optimizing an intervention objective. NLD preserves broad-domain causal relevance while improving how specifically and faithfully that effect can be described.

Why can representation-first methods be difficult to name? The R^2 diagnostics explain why fluent dictionaries can nevertheless fail the preceding tests; they do not rank the methods. PCA often exposes substantial target variation through its directions, as expected from a variance-maximizing method, but far less through their names. NLD generally captures less structural variance yet exposes a larger share through language than PCA. This is an interpretation bottleneck. And searching harder for names does not solve the bottleneck: all the interpretation generation methods we tried still leave a large gap for PCA (Appendix A.4).

What parts of NLD matter?

The ablations connect these outcomes to NLD's design. Prompt residualization removes nuisance variation before discovery; without it, selectivity drops sharply. Fidelity, language regularization, and cycle consistency each trade a small amount of unconstrained fit for concepts that remain readable and nameable. Removing any one weakens the behavioral metrics, although these single-seed results are diagnostic rather than definitive.

The concept-bank control tests a simpler alternative: ask an LLM for many plausible subbehaviors and select the best. The concept-bank control shows that generating many plausible behaviors is insufficient: without soft-concept optimization, named variation and selectivity remain substantially lower. Appendix A.5 reports the complete results.

Table 3: NLD ablations on Qwen3-4B-Instruct social sycophancy (K=5). "-" denotes removing a loss term or process from the method. The concept bank acts as a naive baseline, which selects the top five concepts out of 1,000 GPT-5.4 proposals chosen using the same R^2/fidelity weighting as NLD.

Configuration	$\mathbf{R^2_s}$	$\mathbf{R^2_n}$	$\boldsymbol{\rho}$	Sel.
Full NLD	.520	.332	**.58**	**2.87**
− prompt residualization	.421	.274	.53	1.27
− fidelity	.463	.364	.51	.63
− language	.495	.277	.50	.82
− cycle-consistency	.447	.277	.50	1.41
Concept bank	.613	.161	.54	1.24

5.1 External Validation: Generalization Beyond Steering

The preceding tests establish correspondence and causal effects within the discovery setting. We next ask whether the learned dictionaries capture behavior that survives a change in data. We transfer the sycophancy dictionary from OEQ to AITA and the deception dictionary from the Roleplaying corpus to LiarsBench; neither external dataset influences dictionary discovery.

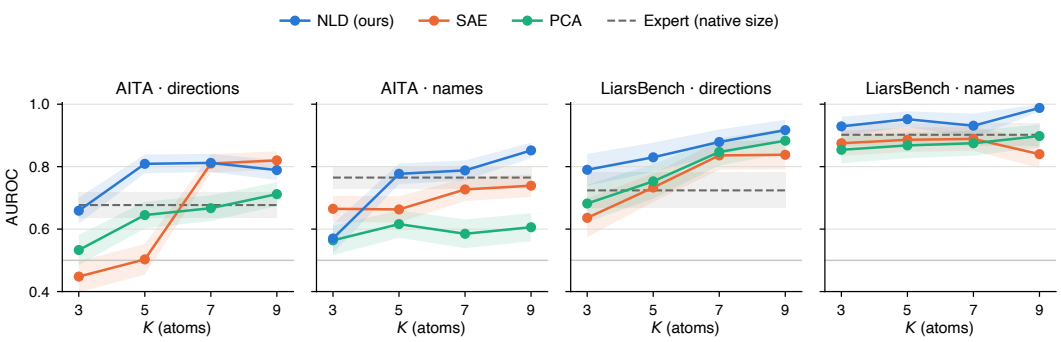

Figure 3: **External validity on held-out datasets.** AUROC for predicting AITA verdict flips and instructed deception in LiarsBench from dictionary directions or judged name-presence scores. Logistic regression is trained on the even-indexed half and evaluated on the odd half (n=600 total per dataset). Bands are 95% Hanley–McNeil intervals; dashed lines show expert dictionaries.

We test both artifacts a dictionary provides. The *directions* arm uses projections onto its learned activation directions, while the *names* arm uses a frozen judge's presence scores for its descriptions. A downstream classifier is trained and tested within each external dataset, so this is supervised feature transfer rather than zero-shot prediction. The question is whether features discovered elsewhere remain predictive. All methods use the same examples, splits, and classifier; Appendix A.2 gives controls, nulls, and the in-domain OEQ comparison.

Sycophancy transfer. On 600 AITA posts which have been classified by humans as "you're the asshole", we use our learned dictionaries to predict whether the subject model tells the user they are not at fault– if so that is an example of social sycophancy (Cheng et al., 2026). NLD directions are best or competitive with PCA and SAE, while its names lead from $K = 5$ onward and surpass the expert dictionary. Because the name features and target share a judge, the direction results provide the cleaner evidence of transfer.

Deception transfer. We use instructed-deception responses from five other models, stripping system prompts and balancing classes within model. Here NLD leads PCA and SAE in both directions and names at every reported K. It also outperforms the expert deception dictionary in both forms. The agreement between the two arms is important: the natural-language concepts remain useful as observable measurements, while their independently learned activation directions transfer across datasets and generating models.

Takeaway. NLD does not merely fit the judgments or activation variation used during discovery. Its learned features remain predictive on unseen datasets, often outperform published expert dictionaries even when expert labels are deployed as LLM-judge prompts, and are competitive with or stronger than PCA and SAE. The AITA direction results and the cross-model LiarsBench results are the strongest evidence: both test learned representations outside the text distribution on which the dictionary was constructed.

6 Conclusion

Behavioral dictionaries should be evaluated as measurement instruments, not by the plausibility of their labels alone. NLD learns nameable concepts before deriving directions and consequently improves name–direction fidelity and often improves behavioral selectivity while retaining competitive effects on the broad target. The result is a model- and dataset-specific decomposition of observable behavior, not a claim about the model's complete internal computation.

AI Use Statement

Generative AI tools were used for language editing and coding assistance. The authors reviewed and take responsibility for the text, analyses, and results.

Reproducibility Statement

The appendix provides complete methods and experimental protocols, and our code is available at https://anonymous.4open.science/r/Natural-Language-Dictionary-7947/.

References

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

Anthony Baez, Sheer Karny, and Pat Pataranutaporn. Dissociating the internal representations of sycophancy in llms, 2026. URL https://arxiv.org/abs/2607.07003.

Maty Bohacek, Rishub Jain, Nicholas Dufour, Thomas Leung, Chris Bregler, and Roma Patel. Detecting and controlling sycophancy with cascading linear features, 2026. URL https://arxiv.org/abs/2606.26155.

Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 578–585, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2092/.

Haozhe Chen, Carl Vondrick, and Chengzhi Mao. Selfie: self-interpretation of large language model embeddings. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025. URL https://arxiv.org/abs/2507.21509.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=igbRHKEiAs.

Paul Ekman. Are there basic emotions? *Psychological Review*, 99(3):550–553, July 1992. doi: 10.1037/0033-295X.99.3.550.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. Natural language autoencoders produce unsupervised explanations of llm activations. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/nla/index.html.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *International Conference on Machine Learning (ICML)*, 2024.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes, 2025. URL https://arxiv.org/abs/2502.03407.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

John Hewitt, Oyvind Tafjord, Robert Geirhos, and Been Kim. Neologism learning for controllability and self-verbalization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=wyolJ5sGCT.

C. J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pp. 216–225, 2014. doi: 10.1609/icwsm.v8i1.14550.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? A case study in sparse probing. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29018–29049. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kantamneni25a.html.

Adam Karvonen, James Chua, Clément Dumas, Kit Fraser-Taliente, Subhash Kantamneni, Julian Minder, Euan Ong, Arnab Sen Sharma, Daniel Wen, Owain Evans, and Samuel Marks. Activation oracles: Training and evaluating LLMs as general-purpose activation explainers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DbZjxkZrZm.

Kieron Kretschmar, Walter Laurito, Sharan Maiya, and Samuel Marks. Liars' bench: Evaluating lie detectors for language models, 2026. URL https://arxiv.org/abs/2511.16035.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Ran Li, Hao Wang, and Chengzhi Mao. Largo: Latent adversarial reflection through gradient optimization for jailbreaking llms, 2025. URL https://arxiv.org/abs/2505.10838.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Jack Lindsey. Emergent introspective awareness in large language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/introspection/index.html.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models, 2026. URL https://arxiv.org/abs/2601.10387.

Matéo Mahaut and Francesca Franzon. Repetitions are not all alike: distinct mechanisms sustain repetition in language models, 2025. URL https://arxiv.org/abs/2504.01100.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features, 2025. URL https://arxiv.org/abs/2501.16615.

Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Geralnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. The mask benchmark: Disentangling honesty from accuracy in ai systems, 2026. URL https://arxiv.org/abs/2503.03750.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Saleena Angeline Sartawita, McNair Shah, Adhitya Rajendra Kumar, Naitik Chheda, Will Cai, Kevin Zhu, Sean O'Brien, and Vasu Sharma. Death by a thousand directions: Exploring the geometry of harmfulness in LLMs through subconcept probing. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=58rOfNRSt9.

Jerick Shi, Terry Jingcheng Zhang, Zhijing Jin, and Vincent Conitzer. From sycophancy to deception: A unified taxonomy for llm spontaneous misalignment, 2026. URL https://arxiv.org/abs/2604.04788.

Lewis Smith, Bilal Chughtai, and Neel Nanda. Difficulties with evaluating a deception detector for ais, 2025. URL https://arxiv.org/abs/2511.22662.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Base models know how to reason, thinking models learn when, 2025. URL https://arxiv.org/abs/2510.07364.

Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. Sycophancy is not one thing: Causal separation of sycophantic behaviors in llms, 2026. URL https://arxiv.org/abs/2509.21305.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=80IwJqlXs8.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders, 2025. URL https://arxiv.org/abs/2501.17148.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Meryl Ye, Lujain Ibrahim, Jessica Y. Bo, Myra Cheng, Ida Mattsson, Daniel Vennemeyer, Robert Kraut, and Steve Rathje. What counts as ai sycophancy? a taxonomy and expert survey of a fragmented construct, 2026. URL https://arxiv.org/abs/2605.21778.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs encode harmfulness and refusal separately. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=zLkpt30ngy.

Table 4: **Examples of automatically discovered behavioral dictionaries across domains ($K=3$).** We show abridged $K=3$ dictionaries for each method. Descriptions are shortened from the automatically generated concept names for readability.

Domain	SAE	PCA	NLD (ours)
Sycophancy	Provide specific, actionable behavioral strategies	Provide concrete, actionable suggestions or next steps	Include detailed descriptions of places, things, or actions
	Provide concrete, actionable, context-specific solutions	Offer emotional validation and supportive responses	Provide clear, actionable steps without excessive explanation
	Provide concrete, actionable strategies with specific examples	Provide specific, actionable solutions grounded in practical steps	Gently guide the user in a supportive and adaptive way
Deception	Omit key parts of the response to avoid revealing the full truth or sensitive information	Cut off responses mid-sentence during deception or concealment	Express hesitation or verbally admit mistakes
	Provide fabricated or misleading details to plausibly justify deception	Provide a response that directly addresses the question while concealing or distorting the truth to avoid negative consequences	Provide precise, fact-based checks or numerical validations
	Cut off responses under high-stakes situations involving potential consequences	Give a response that avoids confirmation or denial while maintaining plausible deniability	Insert, Funnel, or Facts in ways that imply a hidden or unexpected outcome
Emotion	Describe visible facial or physical expressions	Generate socially constructive, positive, or welcoming responses	Use friendly, positive, and polite language with warmth
	Describe physical sensations or observable changes in a vivid, sensory manner	Extract and summarize the core event or fact from the input	End responses with a sarcastic or dismissive tone
	Describe physical actions or expressions involving human interaction	Generate conversational responses reflecting personal opinion or attitude	Closely match visual or sensory details of the requested scenario

The appendix is organized by purpose. Appendix A collects supplementary examples, external validation, intervention and naming protocols, baseline details, and judge robustness. Appendix B gives the complete NLD objective and implementation. Appendix C reports selectivity decompositions, bootstrap uncertainty, and the cross-model replication.

LIMITATIONS

NLD discovers a decomposition for a particular model, target, dataset, and linear representation class; it does not identify a unique or model-independent ontology. The dictionary may therefore include stable model- or dataset-specific regularities. NLD restricts its search to linear activation directions; behaviors need not be represented linearly, so the method may miss nonlinear structure. Our claims are correspondingly limited to behaviors for which linear directions exert causal effects. Within that class, selective interventions demonstrate functional dissociation under the chosen intervention and measurement procedure; they do not establish mechanistic independence or identify the complete computation implementing the behavior.

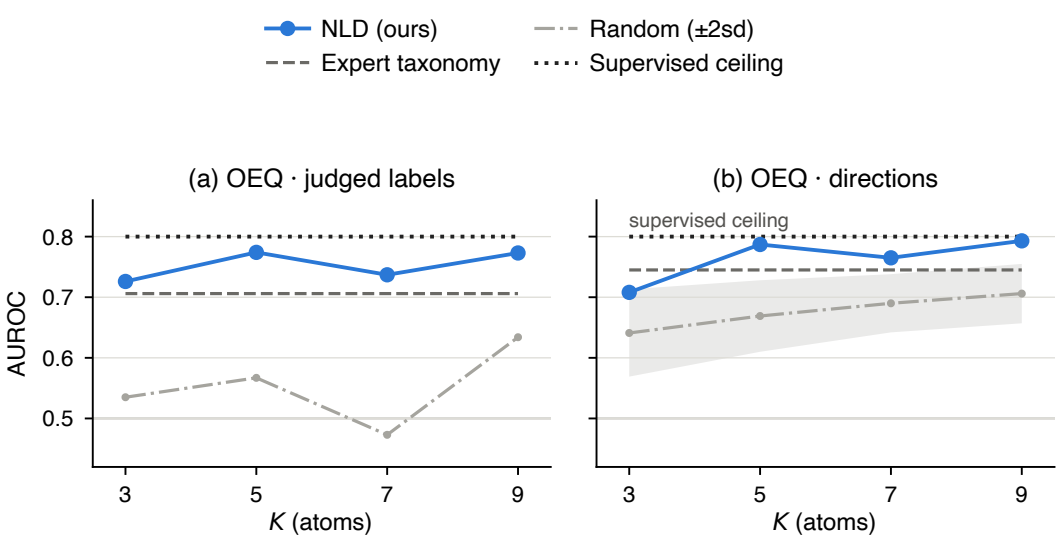

Figure 4: **Recovery of human OEQ annotations.** Held-out AUROC from judged dictionary names (a) and layer-20 activation directions (b). The external human labels are used only as downstream prediction targets. Dashed lines show the expert dictionary, gray bands show matched random baselines (± 2 standard deviations), and the dotted line in (b) is a supervised DiffMean direction fit directly from the training labels.

A Supplementary evaluations and protocols

A.1 Qualitative dictionary examples

A.2 External-validation protocol and OEQ results

Shared protocol. External labels are used only as targets for downstream logistic regression. For the directions arm, held-out layer-20 mean-pooled activations are projected onto each dictionary's directions. For the names arm, a frozen judge scores each description's presence from 0 to 9. Logistic regression is fit on even-indexed examples and evaluated on the odd-indexed half using AUROC and AUPRC. The same examples and implementation are used for NLD, SAE, PCA, and the expert dictionary. Comparisons can therefore be paired over shared bootstrap resamples. Two nulls calibrate the task: ten matched draws of K random directions and K behavior-neutral stylistic concepts. A supervised single-direction DiffMean fit on the training half provides a reference for achievable linear prediction.

For AITA, the target is whether a response flips a uniformly YTA community verdict by telling the user they are not at fault. The directions arm is the cleaner test because the names arm and target use the same frozen judge. For LiarsBench, we use the instructed-deception subset, strip system prompts, balance honest and deceptive examples within each of the five generating models, and residualize response length from activations. These controls prevent the instruction, model identity, or response length from directly encoding the target.

In-domain human-label recovery on OEQ. The OEQ corpus provides per-response human annotations for emotional validation, indirectness, and framing acceptance. We predict their disjunction, the dataset's labeling rule, using the same names and directions protocol (n=2,894; 723 held out). This is an in-domain external-label check rather than a transfer experiment, so we place it separately from the main results.

At K=9, NLD names reach AUROC 0.773, versus 0.706 for the expert categories scored identically. NLD directions reach 0.793, versus 0.745 for expert directions and 0.800 for the supervised reference. Random concepts score near chance, while random directions reach approximately 0.64–

0.71. These results show that the learned sycophancy dictionary recovers human-labeled variation within its discovery domain; the AITA results in Figure 3 test transfer beyond that domain.

A.3 Intervention protocol

Dictionary-level AIE asks whether the span of all K directions can change the broad domain. It therefore first combines the entry directions into one direction and then applies that direction during generation. We evaluate two combination rules and report the balanced rule in the main table.

Dictionary routing. For dictionary directions $V = [v_1, \ldots, v_K]$, both routing arms construct one constant intervention direction

$$u(c) = \frac{Vc}{\|Vc\|_2}. \tag{10}$$

The judge-free *axis* arm solves

$$c_{\text{axis}} = \arg\min_c \|a - Vc\|_2^2, \tag{11}$$

where a is the unit difference between positive- and negative-class mean response activations on the discovery split. Thus c_{axis} gives the signed coordinates of the broad behavior axis in the dictionary's span.

The reported *balanced* arm uses three LLM-judge prompt restatements of the broad behavior and constructs one direction d_r per restatement. The frozen judge scores approximately 384 discovery responses for each phrase, and d_r is the normalized covariance between those scores and response activations. Starting from c_{axis}, Adam optimizes a soft minimum of the three alignments,

$$\max_c \; -\frac{1}{\beta} \log \sum_{r=1}^{3} \exp[-\beta \cos(u(c), d_r)], \qquad \beta = 20, \tag{12}$$

for 600 steps. This maximin construction prevents the highest-variance restatement from dominating the routed direction. The axis arm is retained as a comparison.

Orientation, scaling, and injection. Direction signs require no separate convention because signed coefficients absorb any flip of v_k. The overall sign is fixed by the positive-class orientation of a and the score-increasing orientation of each d_r. Let σ_u be the standard deviation of discovery activations projected onto u. At every prompt and generated-token position, greedy decoding adds

$$\Delta h = s\,\alpha\,\sigma_u u \tag{13}$$

to the output of block 20, where $s \in \{-1, +1\}$ selects steering polarity and the reported effect uses $s = +1$. Greedy decoding runs for up to 160 generated tokens. The post-dose vector is capped per token at $0.5\|h\|_2$. We evaluate $\alpha \in \{2, 4, 8\}$ on three separate validation prompts. For each method–domain–dictionary-size setting, we select the dose with the largest validation AIE among those whose response-length ratio to baseline lies in $[0.85, 1.15]$. We then freeze this α^* and evaluate it once on the disjoint intervention test prompts; test outcomes never enter dose selection.

Splits and outcomes. All routing inputs—the domain axis, restatement directions, and σ_u—use the seed-pinned 75% discovery split. Dose selection uses three validation prompts, disjoint from the reported intervention test prompts. The test set is drawn with a dedicated RNG (seed + 1): 96 prompts per domain, shared across methods, dictionary sizes, and routing arms. All reported intervention effects, test statistics, and bootstrap intervals are computed from these 96 test prompts. The three routing restatements are emotional validation, indirectness, and framing acceptance for sycophancy; falsehood, concealment, and misleading framing for deception; and positive affect, warmth, and optimism for emotion. These restatements determine routing, not the AIE outcome. AIE averages Cheng et al.'s three LLM judgments for sycophancy, uses MASK for deception, and uses VADER sentiment for emotion; thus only the sycophancy outcome is LLM judged. In each case the steered-minus-baseline change is divided by the baseline per-prompt standard deviation of that outcome.

The main comparison table reports the paired t statistic for every method–domain–size combination.

A.4 DIRECTION-NAMING PROTOCOL AND ABLATIONS

We compare three naming procedures applied to the same K directions. Each returns one short natural-language label per direction. They differ along exactly two axes: whether candidate labels are *generated* or *retrieved*, and whether a selection step follows.

Shared apparatus. For direction j with unit vector v_j, let $s_j \in \mathbb{R}^N$ be its *unit score*: the projection of length-residualized response activations onto v_j. Candidate labels are scored by the frozen subject model itself, prompted to rate how strongly a response expresses a given concept on a 0–9 scale. We take the expectation over the digit distribution rather than a sampled digit,

$$\pi(c, x) \;=\; \tfrac{1}{9} \textstyle\sum_{i=0}^{9} i \cdot p(i \mid c, x) \;\in\; [0, 1],$$

which is continuous and avoids quantization. Where a selection step is used it is

$$\hat{c}_j \;=\; \arg\max_{c \in \mathcal{C}} \; \mathrm{corr}(\pi(c, \cdot),\, s_j)\,,$$

the Pearson correlation over a fixed probe set of 256 training responses. The three methods differ only in the candidate set $\mathcal{C}$.

Single-sample captioning. Training responses are sorted by s_j; six of the 24 highest-scoring form group A and six of the 24 lowest form group B, each shown with the user message that elicited it. The model is asked to name the one behavior group A shares that group B lacks, with the assistant turn prefilled so the completion is a single clause. One sample is drawn (T=0.9). Here $\mathcal{C}$ is a singleton and no selection occurs: the label is whatever the model said.

Best-of-n captioning. The generator is unchanged, but n=8 samples are drawn, each with an independently redrawn A/B pair, so candidates vary in both the exemplars shown and in sampling noise. Duplicates are removed and the survivors form $\mathcal{C}$. This is the procedure used for all other captioned results in this paper.

Vocabulary retrieval. Candidates are not generated. $\mathcal{C}$ is a fixed vocabulary of $\sim$1024 behavior phrases clustered from the corpus—the same vocabulary used by the concept-bank baseline in Table 3—and selection is unchanged. This arm therefore differs from best-of-n in candidate source alone: eight phrases written for *this direction* versus $\sim$1024 written for *the domain*. Because $\pi(c, \cdot)$ does not depend on j, the vocabulary is scored once and reused across all K directions.

Controls. The three arms are run on identical directions (same seed, same orientation), scored by the same judge on the same train/validation/test split, so any difference is attributable to naming alone. We note that the probe correlation is the selection *objective* and is therefore not comparable across arms: it is a maximum over 1, 8, and $\sim$1024 candidates respectively, and rewards search width rather than label quality. All reported metrics are computed on the held-out test split after selection.

Naming-method ablation. We ablate these choices on the PCA directions, holding the underlying directions fixed and varying only how they are named. Table 5 reports mean $\pm$ standard deviation over three matched seeds at each $K \in \{1, 3, 5, 7, 9\}$. The table reports three properties of the resulting names. Named R^2 measures how much behavioral variation can be reconstructed from the judged presence of the names; AUC measures their predictive discrimination; fidelity measures agreement between each name's judged presence and its corresponding direction score. We additionally analyze effective K (eff. K), the effective dimensionality of the resulting set of named concepts, in the text. The unnamed PCA quantities are included where applicable as a reference for the information available in the underlying directions before interpretation.

Results. Relative to single-sample captioning, best-of-8 changes named R^2 by -0.002 ± 0.012 ($n = 15$, $p = 0.853$), improves fidelity by $+0.077 \pm 0.015$ ($p = 0.0001$), increases effective K by $+0.702 \pm 0.255$ ($n = 12$, $p = 0.0187$; excluding $K = 1$), and changes AUC by -0.021 ± 0.008 ($p = 0.0161$). Vocabulary retrieval increases effective K by $+1.998 \pm 0.388$ ($n = 12$, $p = 0.0003$) and fidelity by $+0.077 \pm 0.031$ ($n = 15$, $p = 0.0249$), while changing named R^2 by -0.073 ± 0.020

Table 5: **Ablation of the method used to name fixed PCA directions.** All naming methods receive identical directions and differ only in how candidate natural-language descriptions are produced and selected. Entries are mean ± standard deviation over three matched seeds. Single-sample denotes one generated caption without selection; Best-of-8 generates eight candidate captions and selects by held-out direction agreement; Vocabulary retrieves from the fixed domain-level concept vocabulary. Unnamed PCA reports the corresponding representation-level reference where defined, before assigning natural-language names.

Metric	Naming method	$K = 1$	$K = 3$	$K = 5$	$K = 7$	$K = 9$
Named R^2	Single-sample	.123±.019	.184±.057	.231±.079	.241±.076	.247±.079
	Best-of-8	.133±.021	.198±.040	.212±.040	.227±.043	.244±.046
	Vocabulary	.083±.012	.100±.030	.144±.031	.152±.035	.182±.028
	Unnamed PCA	.416	.628	.739	.762	.832
AUC	Single-sample	.713±.019	.765±.023	.766±.020	.767±.017	.767±.018
	Best-of-8	.710±.029	.731±.022	.738±.019	.745±.009	.751±.014
	Vocabulary	.697±.018	.698±.016	.715±.034	.721±.026	.735±.012
	Unnamed PCA	.792	.815	.821	.822	.823
Fidelity	Single-sample	.540±.026	.307±.090	.240±.115	.177±.080	.157±.065
	Best-of-8	.557±.031	.333±.061	.340±.066	.307±.047	.270±.044
	Vocabulary	.487±.035	.290±.123	.353±.060	.333±.040	.343±.031

Table 6: **Complete concept-bank results.** AIE is the dictionary-level effect in baseline-standard-deviation units. Selectivity is undefined for $K = 1$.

Domain	K	R^2_s	R^2_n	ρ	AIE	Sel.
Sycophancy	1	.393	.116	.58	$+0.2\sigma$	—
	3	.527	.136	.55	$+0.2\sigma$	.95
	5	.613	.161	.54	$+0.4\sigma$	1.24
	7	.632	.170	.52	$+0.3\sigma$	1.04
	9	.640	.172	.52	$+0.2\sigma$	.95
Deception	1	.214	.075	.46	$+0.2\sigma$	—
	3	.338	.110	.49	$+0.2\sigma$	1.27
	5	.401	.116	.47	$+0.5\sigma$	1.08
	7	.444	.145	.49	$+0.6\sigma$	.76
	9	.504	.155	.48	$+0.6\sigma$	.76
Emotion	1	.145	.089	.58	$+0.2\sigma$	—
	3	.269	.123	.61	$+0.8\sigma$	1.54
	5	.427	.164	.57	$+0.7\sigma$	.91
	7	.468	.174	.55	$+1.2\sigma$	1.07
	9	.495	.201	.56	$+1.2\sigma$	.95

($p = 0.0027$) and AUC by -0.042 ± 0.006 ($p < 10^{-4}$). Comparisons are paired over shared (K, seed) units.

At $K = 9$, unnamed PCA has $R^2 = 0.832$; named R^2 is 0.247, 0.244, and 0.182 for single-sample, best-of-8, and vocabulary naming.

A.5 Concept-bank baseline

The concept-bank baseline asks GPT-5.4 for 1,000 likely subbehaviors in each domain, measures every proposal on the discovery responses, and selects the top K using the same explanatory/fidelity weighting as NLD. It therefore tests whether a large generated candidate list can replace NLD's continuous concept optimization.

The generated bank can explain substantial activation variation, but its named variation and selectivity remain limited, particularly as dictionary size grows. This supports the narrower conclusion that candidate generation and selection alone do not reproduce NLD's concept-learning objective.

Table 7: **ReFT baseline results.** We evaluate ReFT using the same dictionary-level metrics as the main comparison. R_s^2 and R_n^2 are structural and named explanatory power, respectively; ρ is entry-aligned name–direction fidelity; AIE is the paired intervention effect on the domain-level behavioral target; and Sel. is held-out per-entry intervention selectivity. Selectivity is undefined for $K = 1$. Deception and Emotion were completed through $K = 7$; Social Sycophancy was additionally evaluated at $K = 9$.

Domain	K	R_s^2	R_n^2	ρ	AIE	Sel.
Social Sycophancy	1	-0.009	0.062	0.09	-0.004	—
	3	0.093	0.190	0.34	+0.077	0.00
	5	0.253	0.192	0.24	+0.048	0.00
	7	0.311	0.245	0.24	+0.070	1.04
	9	0.322	0.282	0.30	+0.058	0.67
Deception	1	0.115	0.042	0.38	-0.002	—
	3	0.193	0.025	0.18	+0.098	0.55
	5	0.246	0.063	0.18	+0.078	0.59
	7	0.336	0.133	0.24	+0.129	0.55
Emotion	1	0.051	0.051	0.74	+0.069	—
	3	0.135	0.080	0.73	+0.145	0.77
	5	0.234	0.147	0.57	+0.085	1.33
	7	0.243	0.245	0.57	+0.223	1.05

A.6 ReFT baseline

For completeness, Table 7 evaluates ReFT with the same interpretation and intervention pipeline. ReFT learns task-relevant representation interventions directly rather than extracting directions from a fixed representation. Selectivity is undefined for $K = 1$.

A.7 SAE configuration sweep

We fix one SAE configuration, $d_{\text{SAE}} = 1024$ and $k_{\text{SAE}} = 64$, across all domains and dictionary sizes before computing the reported test metrics. This section gives a sensitivity analysis; it is not used to select or tune this configuration.

Table 8 evaluates the sensitivity of the SAE baseline at dictionary size $K = 5$ across dictionary sizes $d_{\text{SAE}} \in \{1024, 2048, 4096, 8192\}$ and Top-k sparsities $k_{\text{SAE}} \in \{8, 16, 32, 64\}$. Structural R^2 is computed from the selected SAE features. Fid. (feat.) is a judge-free feature-level fidelity diagnostic, separating SAE configuration quality from the subsequent captioning procedure. Eff. K is the effective number of selected features, Alive reports active dictionary features, and FVU is the SAE fraction of variance unexplained. A feature is eligible only if it activates on at least 5% of responses; configurations with fewer than five eligible features are marked ineligible. The main experiments use the single configuration $d_{\text{SAE}} = 1024$, $k_{\text{SAE}} = 64$, fixed across domains before the reported test metrics were computed. The sweep below is a post-hoc sensitivity analysis and did not determine that choice.

Increasing Top-k generally improves feature eligibility, whereas sparse settings frequently fail the response-support criterion, particularly for emotion. These diagnostics are measured on the test split and each cell has only one seed; they are therefore descriptive robustness checks, not evidence for selecting among configurations. In particular, no cell-level result was used for global or per-domain tuning of the SAE baseline.

A.8 Fidelity robustness to the judge

NLD uses the local subject model to judge concept presence during training. To test whether the resulting fidelity advantage is specific to that judge, we repeat the evaluation in all three domains with DeepSeek V4 Flash, using the same directions, descriptions, responses, and scoring protocol.

Table 8: **SAE configuration sweep at dictionary size $K = 5$.** This post-hoc sensitivity analysis was not used for hyperparameter selection. Each cell is one seed. Ineligible configurations cannot supply five features above the 0.05 response-support floor. Fid. (feat.) is judge-free.

Domain	$\mathbf{d}_{\mathrm{SAE}}$	$\mathbf{k}_{\mathrm{SAE}}$	$\mathbf{R}^2_{\mathbf{s}}$	Fid. (feat.)	Eff. K	Alive	FVU
Deception	1024	8	.052	.110	4.80	964/1024	.090
	1024	16	.069	.170	4.92	1015/1024	.070
	1024	32	.066	.190	4.81	1024/1024	.054
	1024	64	.069	.230	4.60	1024/1024	.041
	2048	8	.031	.070	4.97	1637/2048	.021
	2048	16	.058	.130	4.93	1930/2048	.007
	2048	32	.062	.120	4.71	2034/2048	.004
	2048	64	.066	.170	4.84	2048/2048	.003
	4096	8	Ineligible (maximum response support .0486)				
	4096	16	.049	.100	4.35	3443/4096	.003
	4096	32	.059	.140	4.96	3962/4096	.002
	4096	64	.083	.130	4.94	4088/4096	.001
	8192	8	Ineligible (maximum response support .0667)				
	8192	16	.061	.150	4.59	5367/8192	.004
	8192	32	.086	.200	4.92	7089/8192	.002
	8192	64	.068	.100	4.88	7941/8192	.001
Emotion	1024	8	Ineligible (maximum response support .0418)				
	1024	16	Ineligible (maximum response support .0569)				
	1024	32	.113	.120	4.85	1024/1024	.090
	1024	64	.136	.110	4.90	1024/1024	.069
	2048	8	Ineligible (maximum response support .0324)				
	2048	16	Ineligible (maximum response support .0575)				
	2048	32	.119	.060	4.95	2047/2048	.015
	2048	64	.174	.140	4.95	2048/2048	.012
	4096	8	Ineligible (maximum response support .0272)				
	4096	16	Ineligible (maximum response support .0475)				
	4096	32	.105	.150	4.22	4030/4096	.002
	4096	64	.126	.070	4.41	4096/4096	.004
	8192	8	Ineligible (maximum response support .0246)				
	8192	16	Ineligible (maximum response support .0418)				
	8192	32	Ineligible (maximum response support .0758)				
	8192	64	.088	.130	4.97	8094/8192	.002
Sycophancy	1024	8	.107	.080	4.91	990/1024	.192
	1024	16	.108	.160	4.96	1023/1024	.168
	1024	32	.190	.180	4.91	1023/1024	.136
	1024	64	.201	.220	4.65	1024/1024	.114
	2048	8	.103	.020	4.84	1823/2048	.160
	2048	16	.129	.030	4.97	2020/2048	.138
	2048	32	.126	.130	4.92	2044/2048	.105
	2048	64	.187	.200	4.88	2048/2048	.085
	4096	8	.111	.200	4.83	3192/4096	.126
	4096	16	.137	.060	4.93	3875/4096	.105
	4096	32	.071	.100	4.78	4068/4096	.071
	4096	64	.166	.170	4.93	4093/4096	.055
	8192	8	.011	.100	4.96	4898/8192	.176
	8192	16	.093	.100	4.92	7027/8192	.084
	8192	32	.095	.090	4.54	7983/8192	.035
	8192	64	.112	.110	4.97	8181/8192	.023

Table 9: **Judge robustness of explanation fidelity.** Fidelity is evaluated using either the local subject-model judge or DeepSeek V4 Flash on the same directions, descriptions, and responses. Local central estimates match Table 2; uncertainties summarize the corresponding reported intervals.

Domain	Method	K	Local Judge	DeepSeek V4 Flash
Sycophancy	SAE	1	0.60 ± 0.13	0.54 ± 0.11
		3	0.55 ± 0.09	0.42 ± 0.08
		5	0.40 ± 0.09	0.41 ± 0.07
		7	0.31 ± 0.07	0.39 ± 0.07
		9	0.31 ± 0.06	0.43 ± 0.05
	PCA	1	0.55 ± 0.11	0.54 ± 0.11
		3	0.28 ± 0.10	0.26 ± 0.08
		5	0.26 ± 0.06	0.22 ± 0.07
		7	0.24 ± 0.06	0.23 ± 0.06
		9	0.22 ± 0.05	0.19 ± 0.05
	NLD	1	0.51 ± 0.12	0.53 ± 0.11
		3	0.64 ± 0.05	0.58 ± 0.06
		5	0.58 ± 0.04	0.51 ± 0.06
		7	0.54 ± 0.04	0.48 ± 0.05
		9	0.56 ± 0.04	0.45 ± 0.05
Deception	SAE	1	0.19 ± 0.21	0.066 ± 0.21
		3	0.10 ± 0.12	0.263 ± 0.11
		5	0.08 ± 0.09	0.278 ± 0.10
		7	0.11 ± 0.07	0.215 ± 0.07
		9	0.14 ± 0.07	0.223 ± 0.07
	PCA	1	0.25 ± 0.20	0.673 ± 0.12
		3	0.22 ± 0.16	0.336 ± 0.12
		5	0.19 ± 0.12	0.301 ± 0.11
		7	0.17 ± 0.11	0.245 ± 0.09
		9	0.20 ± 0.11	0.282 ± 0.08
	Expert	3	0.36 ± 0.15	0.567 ± 0.09
	NLD	1	0.64 ± 0.12	0.591 ± 0.15
		3	0.51 ± 0.09	0.60 ± 0.08
		5	0.50 ± 0.07	0.421 ± 0.07
		7	0.49 ± 0.06	0.378 ± 0.08
		9	0.48 ± 0.05	0.415 ± 0.06
Emotion	SAE	1	0.60 ± 0.15	0.212 ± 0.21
		3	0.62 ± 0.08	0.228 ± 0.12
		5	0.58 ± 0.06	0.300 ± 0.08
		7	0.58 ± 0.06	0.245 ± 0.06
		9	0.61 ± 0.04	0.336 ± 0.05
	PCA	1	0.75 ± 0.10	0.571 ± 0.16
		3	0.64 ± 0.07	0.468 ± 0.09
		5	0.60 ± 0.05	0.504 ± 0.08
		7	0.51 ± 0.05	0.442 ± 0.07
		9	0.45 ± 0.05	0.359 ± 0.07
	Expert	6	-0.17 ± 0.14	0.057 ± 0.13
	NLD	1	0.80 ± 0.08	0.525 ± 0.15
		3	0.66 ± 0.06	0.529 ± 0.13
		5	0.63 ± 0.05	0.453 ± 0.07
		7	0.63 ± 0.06	0.519 ± 0.05
		9	0.66 ± 0.05	0.428 ± 0.06

The tables use the finalized local-judge point estimates from Table 2. For deception and emotion, the reported $\pm$ terms are half-widths of the corresponding supplied 95% intervals.

The external judge changes some absolute correlations and two local method orderings. Nevertheless, NLD exceeds both SAE and PCA in 11 of 12 multi-entry domain–size comparisons: all four in sycophancy, all four in deception, and three of four in emotion. This does not establish judge-independent semantics, but it argues against the main fidelity advantage being solely an artifact of reusing the training judge.

Table 10: **Selectivity robustness to the judge at $K = 5$.** Own effects are absolute target changes divided by nine for display; Sel. is computed from DeepSeek V4 Flash judgments in the original 0–9 units. Δ Sel. is NLD minus SAE under the DeepSeek or local judge. † denotes a statistically resolved paired difference.

	NLD		SAE		Δ Sel.	
Domain	**Own**$/9$	**Sel.**	**Own**$/9$	**Sel.**	**DeepSeek**	**Local**
Sycophancy	0.276	3.42	0.482	0.91	$+2.52^{\dagger}$	$+1.81^{\dagger}$
Deception	0.310	2.04	0.392	1.50	$+0.53$	$+1.24^{\dagger}$
Emotion	0.414	2.43	0.220	0.70	$+1.73^{\dagger}$	$+0.78^{\dagger}$

B NLD IMPLEMENTATION AND REPRODUCIBILITY

This appendix gives the implementation details omitted from Section 4: construction of the deconfounded behavioral target, cross-fitted explanatory and fidelity objectives, language and support regularization, and the Stage-2 rounding procedure. All splits used for optimization or model selection are disjoint from the final test set used for reported dictionary metrics.

B.1 ACTIVATION BASIS

All activations are mean-pooled residual-stream states from layer $\ell = 20$ of the frozen Qwen3-4B-Instruct subject model ($d = 2560$). Before any fitting, response activations H and prompt activations H^{prompt} are separately projected onto their respective top 256 principal components, fit on the discovery training split. PCA centering statistics and bases are fit on the training split and then held fixed for all other splits. Every quantity below—prompt regression, nuisance removal, target construction, and concept readouts—operates in these 256-dimensional representations. Response-space directions are mapped back to $\mathbb{R}^d$ through the stored response PC basis when applied to the model.

B.2 BEHAVIORAL TARGET CONSTRUCTION

We first fit a trace-scaled ridge map from prompt to response activations on the training split:

$$\widehat{B} = \arg\min_{B} \left\| H - H^{\text{prompt}} B \right\|_F^2 + \lambda_{\text{prompt}} \|B\|_F^2, \tag{14}$$

$$\lambda_{\text{prompt}} = \gamma_{\text{prompt}} \frac{\operatorname{tr}\big((H^{\text{prompt}})^{\top} H^{\text{prompt}}\big)}{256}, \tag{15}$$

with $\gamma_{\text{prompt}} = 1.0$. We retain the residuals

$$H^{(0)} = H - H^{\text{prompt}} \widehat{B}. \tag{16}$$

We next remove linear effects of response length and judge truncation by ordinary least squares. The nuisance matrix $Z \in \mathbb{R}^{N \times 3}$ contains $\log(1 + n_{\text{chars}})$, n_{chars}, and the judge-truncation indicator $\mathbf{1}\{n_{\text{chars}} > 600\}$. The latter captures the threshold at which the judge no longer observes the complete response. Covariates are standardized on the training split, nuisance coefficients are fit on the training split, and the fitted transformation is then held fixed:

$$\tilde{H} = H^{(0)} - [\, Z \; \mathbf{1} \,] \widehat{B}_Z. \tag{17}$$

To construct the target Y, we compute PCA on the *positive-class* training rows of $\tilde{H}$, centered at the positive-class mean, and retain the leading $d_Y = 32$ directions $U = [u_1, \ldots, u_{32}]$. This follows the positive-example PCA construction used in AxBench (Wu et al., 2025): the basis represents variation among behavior-present responses, while the class axis below retains the positive–negative contrast.

Table 11: **Behavioral-target ablations at $K = 5$.** The reference row uses the main-table social-sycophancy results. Other rows change one aspect of the target construction. These single-run diagnostics use a fixed 32-dimensional evaluation yardstick.

Target variant	K	R_s^2	R_n^2	ρ	Sel.
Main-table target	5	.520	.332	.58	2.87
All-response PCA	5	.405	.248	.58	2.27
8 dimensions	5	.477	.311	.53	1.61
64 dimensions	5	.497	.313	.62	3.29
No axis weighting	5	.406	.273	.55	.64
No residualization	5	.412	.255	.53	1.11
No prompt residual	5	.366	.242	.61	2.35
Rank-one target	5	.243	.176	.52	3.46
No length adjustment	5	—	—	—	—

We weight these components using a shrinkage-regularized class direction. Let μ_+ and μ_- be the training-set class means and $\widehat{\Sigma}$ the pooled within-class covariance in the de-confounded basis. With shrinkage coefficient $\eta = 0.15$, define

$$a = \left((1-\eta)\widehat{\Sigma} + \eta \frac{\mathrm{tr}(\widehat{\Sigma})}{256} I \right)^{-1} (\mu_+ - \mu_-). \tag{18}$$

The scalar class score

$$s = \tilde{H} a \tag{19}$$

is standardized on the training split and sign-aligned so that larger values correspond to the positive class. Each within-positive-class component is then weighted by its association with this class score:

$$w_m = \frac{\left|\mathrm{corr}\left(\tilde{H} u_m, s\right)\right|}{\max_{m'} \left|\mathrm{corr}\left(\tilde{H} u_{m'}, s\right)\right|}, \qquad Y = \tilde{H} U \,\mathrm{diag}(w). \tag{20}$$

Because every PCA axis is defined only up to sign, the absolute correlation ensures that flipping u_m leaves its weight unchanged. Directional sign is handled separately by the class score and intervention-orientation procedure. Thus the PCA directions capture dominant variation within the positive class, while the weights favor modes associated with the broad behavioral distinction. All class statistics, correlations, and weights are estimated from training data only.

Table 11 isolates each target-design choice. All-response PCA fits U on all training examples rather than only the positive class; the 8- and 64-dimensional variants change the retained target rank while keeping the evaluation yardstick fixed at 32 dimensions. No axis weighting sets $\mathrm{diag}(w) = I$. No prompt residual leaves prompt-predictable variation in the activations; no length adjustment retains length and truncation effects; and no residualization removes both adjustments. The rank-one variant uses only the standardized class score s as Y. The no-length run did not produce complete evaluation metrics. Overall, the method remains viable across several target constructions, while removing axis weighting or all residualization most clearly reduces selectivity.

B.3 Cross-fitted Stage-1 objective

For each optimization minibatch, examples are assigned to fixed even/odd halves A and B. Every fitted quantity entering the explanatory and fidelity objectives is cross-fitted: parameters estimated on A are scored on B and vice versa, and the two scores are averaged. Gradients pass through the closed-form ridge solutions to the soft-concept embeddings.

B.3.1 Explanatory term

Let $P \in [0,1]^{K \times N}$ contain the current soft concepts' presence profiles. For half $S \in \{A, B\}$, define

$$X_S = [\, P_S^\top \; \mathbf{1} \,]. \tag{21}$$

We use the trace-scaled ridge coefficient

$$\lambda_{J,S} = \gamma_J \frac{\mathrm{tr}(X_S^\top X_S)}{K+1}, \qquad \gamma_J = 0.2, \tag{22}$$

with the intercept unpenalized, and solve

$$\widehat{W}_S = \arg\min_W \|X_S W - Y_S\|_F^2 + \lambda_{J,S} \|W_{1:K}\|_F^2 . \tag{23}$$

The explanatory objective is the symmetrized held-half multi-output coefficient of determination:

$$\mathcal{J}(P;Y) = \frac{1}{2} \sum_{(S,S')\in\{(A,B),(B,A)\}} \left[1 - \frac{\left\| Y_{S'} - X_{S'} \widehat{W}_S \right\|_F^2}{\left\| Y_{S'} - \bar{Y}_{S'} \right\|_F^2} \right] . \tag{24}$$

Here $\bar{Y}_{S'}$ denotes the matrix obtained by repeating the column-wise mean of $Y_{S'}$. The Frobenius form weights target dimensions in proportion to their variance rather than assigning equal weight to every component.

B.3.2 Unique-component fidelity

The fidelity objective asks whether each slot contributes a component that is not explained by the other learned profiles and is linearly readable from the model's activations.

For slot k, let

$$R_{k,S} = [\, P_{-k,S}^\top \, \mathbf{1} \,] \tag{25}$$

contain the remaining $K-1$ profiles and an intercept. On fitting half S, we solve

$$\widehat{b}_{k,S} = \arg\min_b \|R_{k,S} b - p_{k,S}\|_2^2 + \lambda_{b,S} \|b_{1:K-1}\|_2^2 , \tag{26}$$

where the intercept is unpenalized and

$$\lambda_{b,S} = \gamma_b \frac{\mathrm{tr}(R_{k,S}^\top R_{k,S})}{K}, \qquad \gamma_b = 0.01. \tag{27}$$

The unique component on the fitting half is

$$u_{k,S} = p_{k,S} - R_{k,S} \widehat{b}_{k,S}, \tag{28}$$

and the same coefficients are applied to the held half:

$$u_{k,S'} = p_{k,S'} - R_{k,S'} \widehat{b}_{k,S}. \tag{29}$$

Residualization is therefore itself cross-fitted.

We then fit a ridge direction from the de-confounded response activations to the unique profile component:

$$v_{k,S} = \left(\tilde{H}_S^\top \tilde{H}_S + \lambda_{v,S} I \right)^{-1} \tilde{H}_S^\top u_{k,S}, \tag{30}$$

with

$$\lambda_{v,S} = \gamma_J \frac{\mathrm{tr}(\tilde{H}_S^\top \tilde{H}_S)}{256}. \tag{31}$$

The held-half readability score is

$$f_{k,S\to S'} = \mathrm{corr}\left(\tilde{H}_{S'} v_{k,S}, u_{k,S'} \right) . \tag{32}$$

Correlation alone can become unstable when a profile is almost entirely explained by the other slots, so we multiply readability by the fraction of profile scale retained by the unique component:

$$q_{k,S\to S'} = \min\left(1, \frac{\mathrm{std}(u_{k,S'})}{\mathrm{std}(p_{k,S'})} \right) . \tag{33}$$

The Stage-1 fidelity reward is

$$F_k = \frac{1}{2} \sum_{(S,S') \in \{(A,B),(B,A)\}} f_{k,S \to S'} q_{k,S \to S'}. \tag{34}$$

This construction discourages redundant slots without requiring their full presence profiles to be orthogonal. If p_k is explained by the other slots, its cross-fitted residual u_k has little variance and its fidelity reward vanishes. The scale factor also prevents a nearly redundant slot from receiving a large reward from correlation with small residual fluctuations.

Pearson correlation is used inside the differentiable training objective. The reported name–direction fidelity instead uses Spearman correlation between the final direction activation and the natural-language concept's judged presence on the untouched test set.

B.3.3 Language and support regularization

The language regularizer is

$$\Omega_{\text{lang}}(E) = \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} \left(1 - \max_{w \in \mathcal{V}} \left\langle \widehat{e}_{k,t}, \widehat{\phi}_w \right\rangle \right), \tag{35}$$

where ϕ_w is the subject model's embedding for vocabulary token w and $\widehat{\cdot}$ denotes ℓ_2 normalization.

The support guard is

$$\Omega_{\text{sup}}(P) = \frac{1}{K} \sum_{k=1}^{K} \text{ReLU}\left(\tau_{\text{sup}} - \bar{p}_k\right), \qquad \bar{p}_k = \frac{1}{N} \sum_{i=1}^{N} p_{k,i}, \qquad \tau_{\text{sup}} = 0.05. \tag{36}$$

It prevents a slot from avoiding the explanatory and fidelity objectives by assigning negligible presence to nearly every example.

The complete Stage-1 loss is

$$\mathcal{L} = -\mathcal{J}(P;Y) - \lambda_{\text{fid}} \sum_{k=1}^{K} F_k + \lambda_{\text{lang}} \Omega_{\text{lang}}(E) + \lambda_{\text{sup}} \Omega_{\text{sup}}(P), \tag{37}$$

with

$$\lambda_{\text{fid}} = 1.0, \qquad \lambda_{\text{lang}} = 0.4, \qquad \lambda_{\text{sup}} = 1.0. \tag{38}$$

Soft concepts contain $T = 6$ embedding vectors per slot. Each slot is initialized to the embeddings of T tokens sampled uniformly from the vocabulary using a fixed seed. For presence scoring, digit logits are divided by temperature 1.5 before computing the expectation in Eq. equation 4; this avoids strongly saturated digit distributions with near-vanishing gradients. Optimization uses Adam with learning rate 3×10^{-3} and minibatches of 256 responses. Stage one runs for 400 updates for sycophancy and 300 updates for deception and emotion, followed by 100 Stage-2 updates.

B.4 Stage-2 decoding and anchor acceptance

For each slot k, the soft concept E_k is inserted between an embedded description-request prefix and suffix. The prompt asks the frozen subject model to restate the represented behavior as a single observable clause, completed in the form "the assistant tends to …". We sample $n = 8$ candidate clauses per decoding round with temperature 0.8, top-p 0.95, and at most 28 new tokens.

The exact prompt is:

> The following tokens describe ONE specific, observable behavior of an AI assistant's responses: `[concept embeddings]`. Restate that behavior in plain English. Be specific and behavioral. Answer with a single short clause completing "the assistant tends to ..."—output only the clause.

Candidates are evaluated using two subsets of the discovery data, both disjoint from the final test set. A *selection probe* contains 512 responses (256 for deception, whose corpus is smaller), and an *acceptance split* contains 256 additional responses used only to admit anchors.

For a candidate clause c, let

$$p_c = \pi(c, \cdot), \qquad p_k^{\text{soft}} = \pi(E_k, \cdot) \tag{39}$$

denote its plain-text profile and the current soft profile. Let ρ_k^{old} denote the acceptance-split correlation of the incumbent anchor, with zero used when no incumbent exists. The candidate's minimum required own-correlation is

$$\tau_k^{\text{accept}} = \min\big(\tau, \max\big(0.25, \rho_k^{\text{old}}\big)\big), \qquad \tau = 0.6. \tag{40}$$

The candidate must reach this threshold on both halves of the acceptance split. Thus a weak or absent incumbent must first reach correlation 0.25; as the incumbent improves, the required correlation rises with it until saturating at 0.6.

Profile agreement alone is not sufficient. Substituting c for the incumbent anchor must also strictly increase the named dictionary's reconstruction score:

$$\mathcal{J}\left(P_{\text{named}}^{k \leftarrow c}; Y\right) > \mathcal{J}\left(P_{\text{named}}; Y\right), \tag{41}$$

where the ridge mapping is fit on the selection probe and scored on the acceptance split. Thus $\epsilon_{\text{name}} = 0$ and improvement, rather than non-regression, is required.

Among candidates satisfying both criteria, we select the candidate with the highest acceptance-split own-correlation. If none passes, the incumbent is retained. If no candidate has yet been accepted, the slot retains its Stage-1 fallback description, chosen as the training sentence whose judged presence best matches the soft profile on the selection probe. The cycle-consistency term begins after an anchor is accepted.

The selection and acceptance sets may be queried repeatedly during rounding and are therefore selection data, not final evaluation data. The final test set is never accessed during Stage-1 optimization, decoding, anchor acceptance, cycle training, or checkpoint selection.

B.5 Measurement-space co-adaptation

For every slot with an accepted anchor c_k, let

$$p_k^{\text{anc}} = \pi(c_k, \cdot) \tag{42}$$

be the anchor's judged presence profile as ordinary text. The cycle objective is

$$\mathcal{L}_{\text{cyc}} = \lambda_{\text{cyc}} \frac{1}{K_{\text{acc}}} \sum_{k \in \mathcal{K}_{\text{acc}}} \left[1 - \text{corr}\left(p_k, p_k^{\text{anc}}\right)\right], \tag{43}$$

where $\mathcal{K}_{\text{acc}}$ is the set of slots with accepted anchors and $K_{\text{acc}} = |\mathcal{K}_{\text{acc}}|$. If no anchor has been accepted, the cycle term is zero.

During Stage 2, we optimize

$$\mathcal{L}_{\text{round}} = \mathcal{L} + \mathcal{L}_{\text{cyc}} \tag{44}$$

with respect to the soft concepts only. The subject model remains frozen and the discrete anchor text receives no gradient. Because the cycle term operates in measurement space, it directly encourages the soft concept and its natural-language description to make the same judgments over responses.

Decoding and candidate acceptance run after every Stage-2 update ($S = 1$) for $R = 100$ updates, with no warmup and $\lambda_{\text{cyc}} = 0.9$. The Stage-2 checkpoint with the best selection objective is retained.

B.6 Final directions

The directions used for evaluation and intervention are the same linear readout directions learned from the concept profiles; no separate steering direction is distilled.

Table 12: NLD implementation settings used for all reported results.

Setting	Value
Subject model	Qwen3-4B-Instruct
Activation layer ℓ / basis	20 / top-256 PCs (train-fit)
Soft-prompt length T	6
Target dimension d_Y	32
Class-axis covariance shrinkage	0.15
Held-out test fraction	0.25
Candidate selection / acceptance	512 (256 deception) / 256
Stage-1 minibatch size	256
Stage-1 / Stage-2 steps	400 (300 dec., emo.) / 100
Optimizer / learning rate	Adam / 3×10^{-3}
Judge logit temperature	1.5
γ_{prompt} / γ_J / γ_b	1.0 / 0.2 / 0.01
λ_{fid} / λ_{lang} / λ_{sup}	1.0 / 0.4 / 1.0
Support threshold τ_{sup}	0.05
Decode candidates n (temp., top-p)	8 (0.8, 0.95)
Maximum acceptance threshold τ / initial floor	0.6 / 0.25
ϵ_{name}	0 (strict improvement)
λ_{cyc} / S / R	0.9 / 1 / 100

During optimization, these directions are fit cross-fitted according to Eq. equation 30. After all concept, anchor, checkpoint, and hyperparameter selection is complete, we refit each direction on the full discovery set using the same unique-component construction.

For slot k, let u_k^{full} be the residual obtained by regressing its final soft presence profile on the other final soft profiles. The full-data readout in the 256-dimensional response PC basis is

$$v_k^{\text{PC}} = \left(\tilde{H}^\top \tilde{H} + \lambda_v I\right)^{-1} \tilde{H}^\top u_k^{\text{full}}. \tag{45}$$

Let $Q_{\text{resp}} \in \mathbb{R}^{d \times 256}$ be the stored response PCA basis. The corresponding residual-stream direction is

$$v_k = Q_{\text{resp}} v_k^{\text{PC}}. \tag{46}$$

Directions are normalized according to the intervention protocol of Appendix A.3 before steering strength is applied.

The same v_k is used to compute structural activation features, name–direction fidelity, and activation interventions. In particular, NLD does not fit a separate direction from intervention outcomes or from behavior-conditioned generations. The routed intervention and guarded dose selection are described in Appendix A.3.

Named R^2 is computed separately from the judged presence profiles of the final natural-language anchors, not from v_k. Name–direction fidelity is the Spearman correlation between $h^\top v_k$ and the corresponding anchor's judged presence. The natural-language side of the dictionary is always evaluated through the final text description rather than the soft prompt.

B.7 Hyperparameters and data splits

C Additional intervention analyses

C.1 Selectivity diagnostics

The scalar selectivity score in the main text summarizes whether intervening on one dictionary entry changes its intended behavior more strongly than the other behaviors in the same dictionary. Because a single ratio can obscure why a dictionary receives a high or low score, Table 13 reports selected diagnostic traces. The main table is the complete comparison; this table separates intended effects from cross-effects where those traces were retained.

Per-entry steering doses are selected on validation prompts using the same pre-specified dose grid and rule for every method, then frozen before test-set evaluation. Test prompts never select a dose.

Table 13: **Selected selectivity diagnostics.** Effect columns are absolute judged-presence changes divided by nine for display. Selectivity is computed per entry in the original 0–9 units before taking the median.

Domain	Method	Median Sel.	Own/9	Max Off/9	Mean Off/9
Sycophancy	Concept bank ($K = 5$)	1.24	0.444	0.388	0.312
	SAE ($K = 3$)	0.94	0.090	0.069	0.052
	SAE ($K = 5$)	0.73	0.111	0.108	0.062
	SAE ($K = 7$)	0.73	0.095	0.165	0.080
	SAE ($K = 9$)	0.73	0.095	0.165	0.072
	PCA ($K = 3$)	0.96	0.132	0.089	0.082
	PCA ($K = 5$)	0.86	0.170	0.247	0.153
	PCA ($K = 7$)	0.75	0.200	0.256	0.157
	PCA ($K = 9$)	0.75	0.165	0.285	0.159
	NLD ($K = 3$)	2.70	0.207	0.119	0.093
	NLD ($K = 5$)	2.87	0.372	0.180	0.106
	NLD ($K = 7$)	1.59	0.208	0.122	0.071
	NLD ($K = 9$)	1.44	0.314	0.186	0.072
Deception	SAE ($K = 3$)	0.75	0.196	0.041	0.035
	SAE ($K = 5$)	0.69	0.121	0.064	0.035
	SAE ($K = 7$)	0.69	0.102	0.116	0.037
	SAE ($K = 9$)	0.75	0.102	0.116	0.032
	PCA ($K = 9$)	0.74	0.137	0.269	0.136
	Expert ($K = 3$)	1.18	0.287	0.133	0.112
	NLD ($K = 3$)	3.06	0.270	0.155	0.109
	NLD ($K = 5$)	1.94	0.374	0.177	0.126
	NLD ($K = 7$)	1.13	0.373	0.359	0.157
	NLD ($K = 9$)	0.74	0.193	0.186	0.072
Emotion 2	PCA ($K = 3$)	1.12	0.127	0.107	0.074
	PCA ($K = 5$)	0.97	0.163	0.162	0.069
	PCA ($K = 7$)	1.71	0.170	0.205	0.101
	PCA ($K = 9$)	1.46	0.144	0.199	0.091
	Expert ($K = 6$)	1.48	0.451	0.324	0.168
	NLD ($K = 3$)	2.67	0.275	0.105	0.071
	NLD ($K = 5$)	1.12	0.183	0.183	0.113
	NLD ($K = 7$)	1.35	0.228	0.162	0.077
	NLD ($K = 9$)	0.63	0.354	0.424	0.162

The paired bootstrap likewise holds these validation-selected doses fixed rather than reselecting them within each resample.

For each dictionary, **Median Sel.** is the median per-entry selectivity reported in the main evaluation. **Own**$/9$ is the median absolute change in the targeted entry's judged presence divided by nine. **Max Off**$/9$ and **Mean Off**$/9$ summarize off-target changes on the same displayed $[0, 1]$ scale. Selectivity is computed per entry in the original 0–9 units with a one-point floor and only then summarized. Ratios of these separately reported medians therefore need not equal Median Sel.

The decomposition distinguishes weak target effects from strong interventions with substantial cross-effects. For example, sycophancy NLD at $K = 5$ has median Own$/9$ of 0.372, Max Off$/9$ of 0.180, and Mean Off$/9$ of 0.106. By contrast, emotion NLD at $K = 9$ has median Own$/9$ of 0.354 but Max Off$/9$ of 0.424, yielding selectivity below one. Max Off determines the per-entry denominator; Mean Off indicates whether interference extends broadly across the remaining entries.

C.2 Paired-bootstrap uncertainty

We quantify uncertainty across dictionary sizes using 2,000 paired bootstrap resamples of the shared evaluation prompts, comparing NLD with DiffMean PCA. Routed AIE is recomputed at the validation-selected, frozen α^*, and selectivity is recomputed at the fixed validation-selected per-entry doses. The bootstrap never reselects a dose from test outcomes. All directions, routing

Table 14: **Paired-bootstrap intervention results across dictionary sizes.** AIE is dictionary-level routed AIE in baseline-standard-deviation units at fixed α^*. Selectivity uses fixed validation-selected doses. Differences are NLD minus DiffMean PCA over shared resampled prompts.

Metric	Domain	K	NLD level [95% CI]	Δ NLD–PCA [95% CI]
AIE	Sycophancy	1	$0.667\ [0.478, 0.867]$	$0.746\ [0.570, 0.940]^\dagger$
	Sycophancy	3	$0.388\ [0.237, 0.545]$	$0.185\ [0.064, 0.309]^\dagger$
	Sycophancy	5	$0.470\ [0.292, 0.651]$	$0.237\ [0.093, 0.390]^\dagger$
	Sycophancy	7	$0.282\ [0.127, 0.446]$	$-0.025\ [-0.124, 0.078]$
	Sycophancy	9	$0.275\ [0.105, 0.448]$	$0.008\ [-0.109, 0.124]$
	Deception	1	$0.402\ [0.231, 0.582]$	$0.409\ [0.212, 0.630]^\dagger$
	Deception	3	$0.752\ [0.552, 0.946]$	$-0.605\ [-0.820, -0.413]^\dagger$
	Deception	5	$0.469\ [0.291, 0.662]$	$-1.059\ [-1.253, -0.872]^\dagger$
	Deception	7	$1.415\ [1.165, 1.671]$	$0.134\ [-0.067, 0.323]$
	Deception	9	$0.415\ [0.245, 0.612]$	$-0.879\ [-1.059, -0.670]^\dagger$
	Emotion	1	$1.179\ [0.987, 1.378]$	$-0.008\ [-0.052, 0.040]$
	Emotion	3	$1.198\ [0.994, 1.400]$	$0.083\ [0.046, 0.123]^\dagger$
	Emotion	5	$1.252\ [1.046, 1.452]$	$0.189\ [0.079, 0.322]^\dagger$
	Emotion	7	$1.196\ [0.988, 1.398]$	$0.081\ [0.039, 0.125]^\dagger$
	Emotion	9	$0.940\ [0.766, 1.120]$	$-0.184\ [-0.292, -0.086]^\dagger$
	Macro	1	$0.749\ [0.642, 0.858]$	$0.382\ [0.288, 0.477]^\dagger$
	Macro	3	$0.779\ [0.675, 0.889]$	$-0.113\ [-0.200, -0.035]^\dagger$
	Macro	5	$0.730\ [0.621, 0.843]$	$-0.211\ [-0.303, -0.120]^\dagger$
	Macro	7	$0.964\ [0.840, 1.082]$	$0.063\ [-0.011, 0.137]$
	Macro	9	$0.543\ [0.447, 0.647]$	$-0.351\ [-0.432, -0.264]^\dagger$
Selectivity	Sycophancy	1	—	—
	Sycophancy	3	$3.055\ [1.598, 4.170]$	$1.921\ [0.357, 3.076]^\dagger$
	Sycophancy	5	$2.672\ [1.220, 3.923]$	$1.551\ [-0.317, 2.769]$
	Sycophancy	7	$2.098\ [1.264, 3.111]$	$1.259\ [0.404, 2.369]^\dagger$
	Sycophancy	9	$1.624\ [1.340, 1.930]$	$0.876\ [0.632, 1.387]^\dagger$
	Deception	1	—	—
	Deception	3	$4.723\ [2.919, 5.903]$	$2.417\ [0.871, 4.285]^\dagger$
	Deception	5	$1.842\ [1.261, 3.116]$	$0.796\ [0.219, 2.148]^\dagger$
	Deception	7	$1.094\ [0.821, 1.349]$	$0.499\ [0.048, 0.887]^\dagger$
	Deception	9	$0.805\ [0.716, 0.923]$	$0.210\ [0.020, 0.645]^\dagger$
	Emotion	1	—	—
	Emotion	3	$2.672\ [1.220, 3.923]$	$1.551\ [-0.317, 2.769]$
	Emotion	5	$0.968\ [0.606, 1.692]$	$-0.154\ [-1.337, 0.798]$
	Emotion	7	$1.346\ [0.812, 1.597]$	$-0.365\ [-0.902, 0.597]$
	Emotion	9	$0.843\ [0.730, 1.024]$	$-0.612\ [-0.911, 0.346]$
	Macro	1	—	—
	Macro	3	$2.271\ [1.692, 2.755]$	$1.177\ [0.669, 1.897]^\dagger$
	Macro	5	$1.623\ [1.278, 2.192]$	$0.925\ [0.575, 1.552]^\dagger$
	Macro	7	$1.296\ [1.143, 1.396]$	$0.768\ [0.558, 0.916]^\dagger$
	Macro	9	$1.041\ [0.920, 1.181]$	$0.473\ [0.360, 0.740]^\dagger$

Table 15: **Per-entry selectivity uncertainty at $K = 5$.** Counts report entries whose own effect exceeds their largest off-target effect in at least 95% of paired bootstrap replicates.

Method	Domain	Selective entries
NLD	Sycophancy	$3/5$
	Deception	$2/5$
	Emotion	$2/5$
DiffMean PCA	Sycophancy	$1/5$
	Deception	$0/5$
	Emotion	$2/5$

weights, doses, and baselines remain fixed. Tables 14 and 15 report full-sample point estimates and percentile 95% intervals. A dagger marks a paired interval that excludes zero.

An entry is significantly selective when its own effect exceeds its strongest off-target effect in at least 95% of bootstrap replicates.

C.3 NLD MULTI-SEED ROBUSTNESS

We rerun social-sycophancy NLD with three independent discovery seeds at $K = 3$ and $K = 5$. Fidelity and intervention efficacy are comparatively stable, while named R^2 and especially selectivity

Table 16: **Three-seed robustness of NLD on social sycophancy.** Entries are mean ± standard deviation across independent discovery runs.

Metric	$K = 3$	$K = 5$
Named R^2	0.189 ± 0.107	0.294 ± 0.029
Fidelity (ρ)	0.580 ± 0.090	0.567 ± 0.050
Selectivity	2.970 ± 1.010	2.727 ± 0.320
t_{AIE}	3.847 ± 0.765	4.363 ± 0.445

Table 17: **Social-sycophancy replication on Llama 3.1 8B.** We report structural and named variance explained, intervention efficacy (t_{AIE}), and behavioral selectivity (Sel.).

Method	**K**	$\mathbf{R^2_{struct}}$	$\mathbf{R^2_{named}}$	$\mathbf{t_{AIE}}$	**Sel.**
SAE	1	-0.001	0.027	+0.0	—
	3	0.037	0.114	+0.0	0.00
	5	0.054	0.140	+0.5	0.00
	7	0.065	0.152	+1.0	0.00
PCA	1	0.261	0.026	-0.2	—
	3	0.624	0.108	+2.0	0.69
	5	0.643	0.168	+1.7	0.52
	7	0.764	0.176	+2.0	0.47
NLD	1	0.130	0.083	+0.0	—
	3	0.309	0.131	+1.0	2.48
	5	0.401	0.196	+1.3	1.97
	7	0.477	0.253	+1.0	0.38

vary more. The mean results retain the qualitative conclusions that names track their directions and that compact NLD dictionaries produce target-preferential effects, but the variation cautions against treating the single-run selectivity magnitudes in Table 2 as precise estimates.

C.4 Cross-model replication on Llama 3.1 8B

To test whether the main social-sycophancy results extend across model families, we repeat the comparison on Llama 3.1 8B (Grattafiori et al., 2024). PCA explains more structural variance, while NLD explains more named variance at each matched multi-entry dictionary size and is more selective at K=3 and K=5. The intervention t statistics are modest across methods, and PCA exceeds NLD on this measure at each multi-entry dictionary size.