Ψ: REVEALING TRUTH-DIRECTION SENSITIVITY LOSS UNDER SYCOPHANTIC PRESSURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models can retain decodable truth-related information while their answers change under user pressure. Assessing the robustness of these representations requires measuring how strongly answers respond to the same information. We introduce *Probing–Sensitivity Inspection* (PSI), a protocol that pairs a frozen linear probe with a two-sided intervention along the same fitted truth direction. On single-token true/false tasks with Qwen2.5 (7B, 14B, 32B) and Gemma-2-9B, PSI reveals three findings. First, *decodability–sensitivity dissociation*: probe accuracy remains at 1.000 across the tested prompt conditions and cue levels, while intervention responses vary and reverse under instructed inversion; Qwen2.5-7B then answers truthfully on 0.133 of items. Second, *sensitivity attenuation under preserved accuracy*: under honest system prompts, sensitivity declines with sycophantic pressure in both model families while truthful-answer accuracy remains high. At the first cue level on Qwen2.5-7B, the intervention effect falls by more than one fifth while accuracy is 0.933. Sycophantic cues produce larger reductions than length-matched neutral prompts. Third, *projection-based truth recovery*: under preference-conditioned prompts, amplifying each item's existing projection raises truthful-answer accuracy from 0.700 to 0.883 on Qwen2.5-7B and from 0.500 to 0.850 on Gemma-2-9B, with little change under norm-matched random directions. PSI reveals pressure-dependent changes in truth-direction sensitivity that remain hidden behind stable probe accuracy and strong answering performance.

1 INTRODUCTION

Prior work has shown that truth-related labels can be decoded from language-model activations and that interventions along fitted truth-related directions can alter model outputs (Azaria & Mitchell, 2023; Marks & Tegmark, 2024; Zou et al., 2023; Li et al., 2023). These findings motivate a joint assessment of two properties of a representation. Linear decodability asks whether truth-related information can be recovered from activations, whereas output sensitivity asks how strongly answer logits respond when those activations are displaced along the same direction. We ask how these two properties change together when prompts pressure a model toward a particular answer.

Sycophancy, the tendency to match a user's stated belief over the truthful answer, is a documented behaviour of models trained with human feedback (Sharma et al., 2024; Perez et al., 2023). Sycophancy provides a controlled setting in which a user's stated preference can compete with a truthful answer. We vary this pressure with a five-level cue ladder, from no cue to an authority-plus-threat cue, while an honest system prompt continues to request truthful answers. This design lets us compare truthful-answer accuracy, linear decodability, and truth-direction sensitivity as pressure increases.

We introduce *Probing–Sensitivity Inspection* (PSI), which pairs linear decodability with truth-direction sensitivity along the same fitted direction (Figure 1). PSI fits the direction, reads it with a frozen probe, and measures the answer-logit response to a two-sided intervention as prompt conditions and pressure vary. We evaluate three Qwen2.5 models (7B, 14B, 32B) and Gemma-2-9B on factual statements with single-token True/False answers. We use three prompt conditions: The honest condition supports the graded-pressure experiment; instructed inversion places the requested answer in direct conflict with the ground-truth label; and preference conditioning introduces explicit

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.

competition between truth and the user's stated preference. We fit a truth direction from statement-position residual-stream activations and evaluate on disjoint items both frozen-probe accuracy and output sensitivity to that direction, defined as the signed True-minus-False answer-logit response to a two-sided intervention.

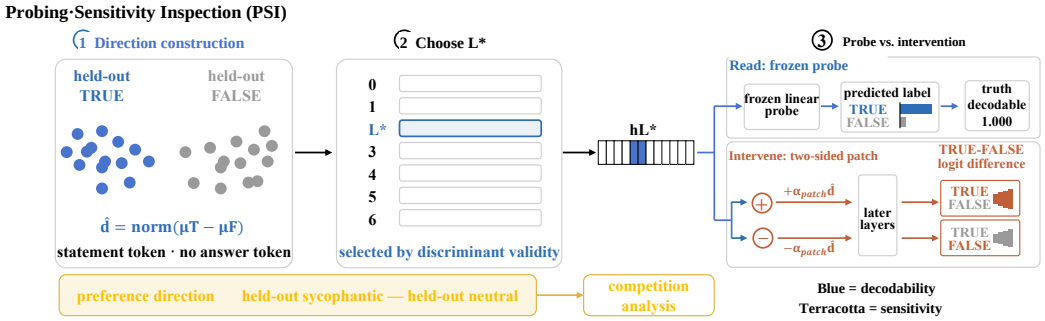


Figure 1: Probing–Sensitivity Inspection (PSI). The main pipeline fits the truth direction by difference of means at the statement position before any answer token, selects a layer on a held-out split, and measures the representation with a frozen probe and a two-sided intervention separately. The intervention adds and subtracts $\alpha\hat{d}$, and its effect is the change in the True-minus-False answer logit (Eq. 2); a preference direction fit from held-out stylistic contrasts supplies the second term of Eq. 5.

Our contributions are threefold: **I) Decodability–sensitivity dissociation.** Probe accuracy stays at 1.000 across the tested prompt conditions and cue levels. Under instructed inversion on Qwen2.5-7B, the model answers truthfully on only 0.133 of items, and the signed intervention effect reverses relative to the honest condition. We call this coexistence of stable decodability and changing intervention responses decodability–sensitivity dissociation. **II) Sensitivity attenuation under preserved accuracy.** Under an honest system prompt, sensitivity declines as sycophantic pressure increases. At the first cue level on Qwen2.5-7B, the intervention effect has already fallen by more than one fifth while truthful-answer accuracy is still 0.933 and probe accuracy remains 1.000. Gemma-2-9B also shows declining sensitivity across the cue ladder while truthful-answer accuracy stays near ceiling. Sycophantic cues produce larger sensitivity reductions than length-matched neutral prompts. **III) Projection-based truth recovery.** At the tested preference-conditioned cue levels, amplifying each item's existing projection onto the truth direction raises truthful-answer accuracy from 0.700 to 0.883 on Qwen2.5-7B and from 0.500 to 0.850 on Gemma-2-9B. Norm-matched random directions have little effect. This projection-based recovery establishes the fitted direction as a useful intervention target under preference conditioning.

Together, these findings establish PSI as a paired assessment of truth-related representations under pressure. Decodability identifies recoverable information, sensitivity tracks the answer response to that direction, and projection amplification tests its utility for improving truthful answers.

2 Related Work

Linear decodability of truth. Linear probes and contrast-based methods have been used to recover truth-related information from language-model activations with both supervised and unsupervised objectives (Azaria & Mitchell, 2023; Burns et al., 2023; Marks & Tegmark, 2024). Interventions along truth-related activation directions can alter how models treat true and false statements (Marks & Tegmark, 2024; Zou et al., 2023) and can improve truthful answering on TruthfulQA (Li et al., 2023; Lin et al., 2022). These studies establish the two operations on which our measurements build: reading truth-related information from an activation direction and intervening along that direction. A related line examines whether truth-related readouts generalise across negation, datasets, and tasks (Bürger et al., 2024; Orgad et al., 2025; Azizian et al., 2025). Our experiments study a different axis of variation: we use the same difference-of-means construction and compare linear decodability with output sensitivity across prompt conditions and cue levels.

Localization and intervention. The presence or localisation of information and the effectiveness of an intervention are separate questions. In model editing, causal-tracing localisation does not reliably

predict which layer is best to edit (Hase et al., 2023). Truth-related probes have also failed to generalise to negated statements in tested settings (Levinstein & Herrmann, 2025). Recent lie-detection studies examine probes across different deception settings: linear probes detect strategic deception in several tested settings (Goldowsky-Dill et al., 2025), while a benchmark covering diverse types of lies evaluates detectors in settings where the transcript alone is insufficient to determine whether the model lied (Kretschmar et al., 2025). Our paired measurement focuses on a different comparison within the same setting: what a frozen probe reads from a truth direction and how strongly the answer responds when that direction is perturbed.

Instructed deception and sycophancy. Sycophancy is documented behaviourally in models trained with human feedback (Sharma et al., 2024; Perez et al., 2023). Instructed dishonesty has been studied using probing and activation patching (Campbell et al., 2023), while other work examines representation shifts under deceptive instructions (Long et al., 2025). Causal interventions and contrastive activation steering have also been used to study lying in practical scenarios (Huan et al., 2025). Other related works take different output-focused approaches: Liang et al. (2025) estimate a belief proxy from first-token response probabilities and compare it with explicit claims to quantify indifference to truth, whereas Turpin et al. (2023) show that biasing prompt features can alter answers without being acknowledged in chain-of-thought explanations. Our contribution to this line is a measurement that compares decodability and output sensitivity along the same fitted direction across prompt conditions, with graded sycophantic pressure and length-matched neutral controls.

3 Probing–Sensitivity Inspection

PSI pairs two measurements on the same fitted truth direction: a frozen-probe readout of the label and the answer-logit response to a two-sided intervention. We compare this pair across prompt conditions and sycophantic cue levels, with direction-specificity and length controls. Projection amplification then tests whether the direction supports truthful-answer recovery.

Stimuli and direction. We use atomic factual statements s with ground-truth labels $y(s) \in \{\text{True}, \text{False}\}$, in the style of curated true/false sets (Azaria & Mitchell, 2023; Marks & Tegmark, 2024). At a residual-stream layer L^* we read the hidden state $h(s) \in \mathbb{R}^d$ at the statement position, before any answer token is produced. Following the difference-of-means construction (Marks & Tegmark, 2024), the truth direction is

$$d \;=\; \frac{1}{|\mathcal{T}|}\sum_{s\in\mathcal{T}} h(s) \;-\; \frac{1}{|\mathcal{F}|}\sum_{s\in\mathcal{F}} h(s), \qquad \hat{d} = d/\|d\|, \tag{1}$$

with $\mathcal{T}, \mathcal{F}$ the true and false statements of a fit split. Fitting at the statement position measures truth-related information before answer-token production. We select the layer L^* by the separation criterion below, using data held apart from direction fitting and evaluation.

Layer selection and data splits. We use three disjoint splits: a fit split for the truth and preference directions, a selection split for the layer, and an evaluation split for all reported sensitivities, dose curves and amplification accuracies. At each candidate layer, we score how well its difference-of-means direction separates the honest and inversion contrasts on the selection split, then choose the layer with the largest margin. The selection split is fixed before sensitivity estimation, and preference-conditioned data are reserved for evaluation. The preference direction is fit from stylistic contrasts in the fit split. The competition prediction in Eq. 5 is tested on the evaluation split.

Decodability. The first measure is whether the information is linearly present. A frozen linear probe on $\langle h(s), \hat{d}\rangle$ predicts $y(s)$, evaluated on the evaluation split.

Truth-direction sensitivity. The second measure is how much the answer moves when the residual stream is pushed along a direction. With the score defined as the logit difference between the True and False answer tokens, the signed effect of a direction d_x is

$$c_x \;=\; \text{score}\big(h + \alpha\, d_x\big) \;-\; \text{score}\big(h - \alpha\, d_x\big). \tag{2}$$

This finite-difference measurement around h quantifies the signed answer-logit response to displacement along d_x at the selected intervention scale. We write c_{truth} for the effect along the unit truth direction $\hat{d}$ and call this quantity truth-direction sensitivity. Large positive c_{truth} means the answer

moves with the truth direction, near zero means the answer barely responds to it, and negative means it moves against it.

Validation and working scale. We compare the truth direction with a norm-matched random direction and a shuffled-label direction over five seeds. These two controls test whether the response is specific to the fitted truth direction. At the working scale, its effect is between 10.2 and 12.1 times the larger null effect across the four models. We report intervention effects at $\alpha^* = 1.0$, the selected working scale from the sweep in Table 1. Across the sweep, the truth-direction effect is about ten times the larger null effect. The table reports accuracy after intervention and $p(\text{True})$ at each scale, including the accuracy drop to 0.500 at $\alpha = 2.00$.

Table 1: Null comparison and scale selection on Qwen2.5-7B (Yang et al., 2024). Ratio is the truth-direction effect over the larger of the two nulls.

α	acc	$p(\text{True})$	c_{truth}	larger null	ratio
0.05	0.933	0.500	+0.642	0.058	11.1
0.25	0.933	0.517	+3.219	0.303	10.6
0.50	0.917	0.533	+6.472	0.621	10.4
1.00	0.850	0.567	+13.09	1.284	10.2
2.00	0.500	0.717	+26.33	2.611	10.1

Two scale-free readouts. Normalising by the honest no-cue reference gives normalized signed sensitivity,

$$\rho \;=\; \frac{c_x}{c_{\text{truth}}^{\text{honest, cue=0}}}, \qquad \rho \approx +1 \text{ (tracks truth)}, \quad \rho \approx 0 \text{ (unresponsive)}, \quad \rho < 0 \text{ (moves against)}, \tag{3}$$

and comparing a preference direction against the truth direction gives

$$\kappa \;=\; \frac{|c_{\text{pref}}|}{|c_{\text{truth}}|}, \qquad \kappa > 1 \;\Rightarrow\; \text{the answer is more sensitive to the cue than to the truth.} \tag{4}$$

The preference direction d_{pref} is fit from stylistic contrasts in the fit split; c_{pref} denotes the intervention effect along that direction. Direction fitting and answer evaluation use disjoint splits.

Prompt conditions and pressure. Three system prompts define the conditions: honest (answer truthfully), instructed inversion (state the negation, elicited with few-shot demonstrations), and preference-conditioned (follow a stated user preference). Each preference-conditioned item carries a user preference assigned independently of the ground truth, making the binary answer an observable expression of that preference. Section 8 relates this operational setting to truth-indifferent answering (Frankfurt, 2005). Sycophantic pressure is a five-level cue ladder from no cue to an authority-plus-threat cue. Inversion is measured without the cue, since the cue overrides the explicit instruction to invert, which we report as a result in Section 7.

Inclusion criterion. We establish the intended prompt conditions through a behavioural screen before comparing their sensitivities. Before any sensitivity is computed for a model, a behavioural criterion (Appendix A) requires that honest answers track truth, preference-conditioned answers track the preference, and instructed inversion inverts at the no-cue level, each by an exact test at $\alpha = 0.05$. Four of eight instruction-tuned models tested meet it.

Competition prediction. To predict emitted answers from independently estimated intervention effects, we form

$$\text{net} \;=\; c_{\text{truth}} \cdot \text{sign}(\text{truth}) \;+\; c_{\text{pref}} \cdot \text{sign}(\text{pref}), \qquad \widehat{\text{answer}} \;=\; \mathbb{1}[\text{net} > 0], \tag{5}$$

and evaluate it out of sample. Both terms are estimated from interventions, with direction fitting and prediction evaluated on disjoint splits.

Length control. A stronger cue is also a longer prompt. We compare each cue against a length-matched neutral prompt and attribute the reduction in sensitivity,

$$\Delta_{\text{cue}} = c_{\text{truth}}^{\text{cue=0}} - c_{\text{truth}}^{\text{cue}}, \quad \Delta_{\text{len}} = c_{\text{truth}}^{\text{cue=0}} - c_{\text{truth}}^{\text{placebo}}, \quad \text{share} = \frac{\Delta_{\text{cue}} - \Delta_{\text{len}}}{\Delta_{\text{cue}}}. \tag{6}$$

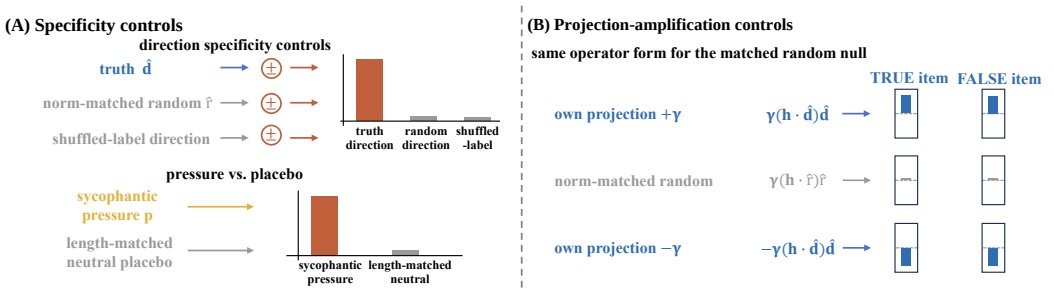


Figure 2: Specificity and projection-amplification controls. (A) The truth direction is compared against a norm-matched random direction and a shuffled-label direction, and the sycophantic cue against a length-matched neutral prompt (Eq. 6). (B) Amplifying an item's own projection (Eq. 7) is compared against a matched random projection and negative amplification on true and false items separately.

Projection amplification. Projection amplification tests whether strengthening the fitted representation improves truthful answers under preference conditioning. We scale each item's own projection onto the truth direction,

$$h \leftarrow h + \gamma\,(h \cdot \hat{d})\,\hat{d}. \tag{7}$$

For unit $\hat{d}$, the operator rescales each item's projection from $h \cdot \hat{d}$ to $(1+\gamma)(h \cdot \hat{d})$. The update is therefore determined by the projection already present in that item's activation.

4 Decodability–Sensitivity Dissociation

Probe accuracy remains stable across conditions. Under the inversion prompt the frozen probe recovers the ground-truth label at accuracy 1.000 while the model's own answers are correct on 0.133 of items (Figure 3). Probe accuracy also remains 1.000 under honest and preference-conditioned prompts and at every tested cue level. The three conditions share ceiling-level decodability, providing a common reference for comparing their intervention responses.

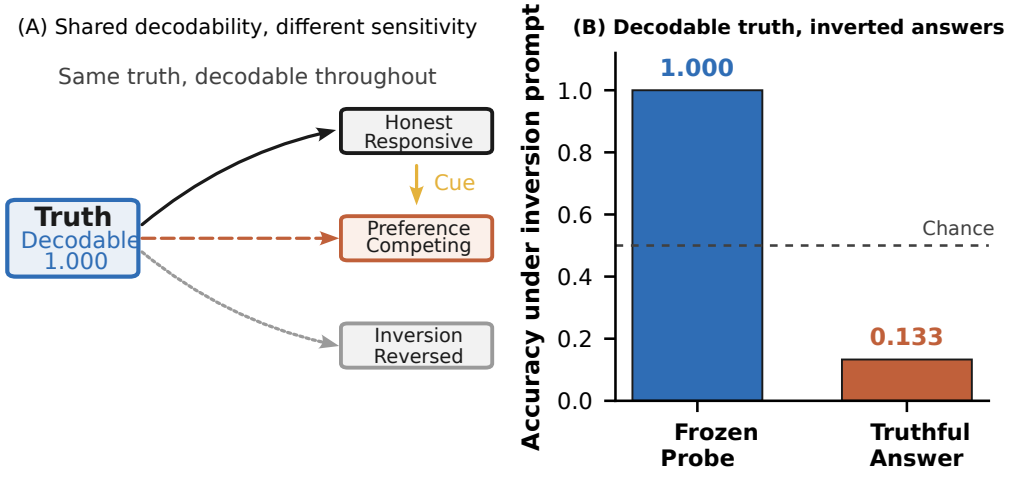


Figure 3: Probe accuracy and answer accuracy under instructed inversion. (A) The truth direction is decodable in all three conditions; what varies is the sensitivity of the answer to intervention along it. (B) On Qwen2.5-7B (Yang et al., 2024) the probe recovers the label at 1.000 while the model's answers are correct on 0.133 of items.

Sensitivity separates them. What differs is how much the answer moves under intervention (Figure 4). At $\alpha^* = 1.0$, the raw effects follow honest > preference-conditioned > 0 > instructed inversion in both model families (Table 2). On Qwen2.5-7B the honest condition gives $\rho = +1.00$ with $\kappa = 0.18$, inversion gives $\rho = -0.37$, and the preference-conditioned condition falls between them, close enough to the honest value that the answer still moves with the truth direction on most items

even while the cue is present. Per item, the honest and inversion conditions separate at 1.000 and the preference-conditioned and inversion conditions at 0.992, while honest and preference-conditioned overlap at 0.683. Separability is computed leave-one-out, so each value is a per-item rate.

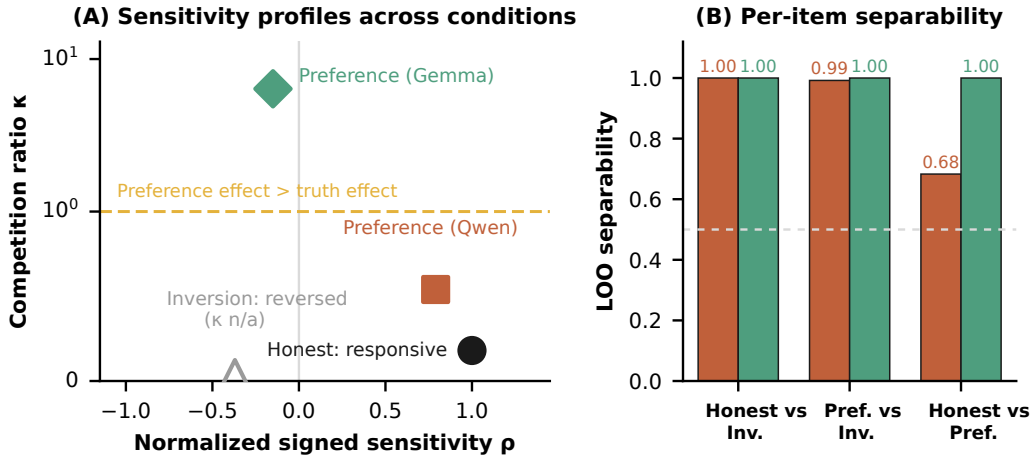


Figure 4: Normalized signed sensitivity and per-item separability. (A) Normalized signed sensitivity ρ against competition ratio κ: the honest condition sits at $\rho = +1.00$, inversion at $\rho = -0.37$, and the preference-conditioned condition at $\rho = +0.79$ on Qwen2.5-7B (Yang et al., 2024) and $\rho = -0.15$, κ=6.38 on Gemma-2-9B (Gemma Team, 2024). (B) Per-item separability, Qwen2.5-7B in terracotta (■) and Gemma-2-9B in green (■); the honest and preference-conditioned conditions overlap on Qwen because the answer still moves with the truth direction on most items.

Out-of-sample competition prediction. On Qwen2.5-7B, the rule in Eq. 5 combines independently estimated intervention effects to predict preference-conditioned answers at 0.942 out of sample against a truth-only baseline of 0.808, a gain of +0.133 from the preference term, and adds 0.008 on the honest condition. On items where the answer contradicts the truth, the preference term is the larger of the two with probability 0.913, a probability difference of +0.789 relative to items answered truthfully. The fraction of items on which the preference term exceeds the truth term rises from 0.275 at the first cue level to 0.567 at the second. On Gemma-2-9B that fraction is 1.000 at every cue level, so every evaluated item is preference-dominated. Both sensitivities carry information about the answer, and their relative size tracks which way the answer goes.

Cue and prompt contribute separately. Against a no-cue reference of $c_{\text{truth}} = +33.67$, the cue accounts for 57% of the reduction in the honest condition at the first cue level and the preference prompt for 43%, and 54% to 46% at the second. Here the system prompt specifies the requested behaviour and the user-side cue supplies the pressure. The larger attributed share comes from the cue. This comparison complements the overlap between honest and preference-conditioned responses in Figure 4B. As pressure rises the balance shifts: κ moves from 0.54 at the first cue level to 1.25 at the second, crossing the point where the answer is more sensitive to the cue than to the truth (Appendix B).

5 Sensitivity Attenuation under Preserved Accuracy

Sensitivity under increasing pressure. Under an honest system prompt we raise the sycophantic cue and measure the intervention effect at each level. The effect declines with pressure on both families. Attributing the decline against length-matched neutral prompts (Eq. 6), the cue-specific share of sensitivity reduction is 98%, 81%, 68% and 61% across the four cue levels on Qwen2.5-7B, and 57% and 54% at the first and fourth levels on Gemma-2-9B. Longer cues carry more text, so some of the decline is attributable to length alone, and that term accounts for a growing share as the cue lengthens; the cue content nonetheless remains the larger of the two at every level we measure.

Accuracy over the same range. Truthful-answer accuracy under the honest prompt falls from 0.933 to 0.733 on Qwen2.5-7B across the same four levels and stays near ceiling on Gemma-2-9B, moving from 1.000 to 0.967. The two curves have different shapes (Figure 5). At the first cue level,

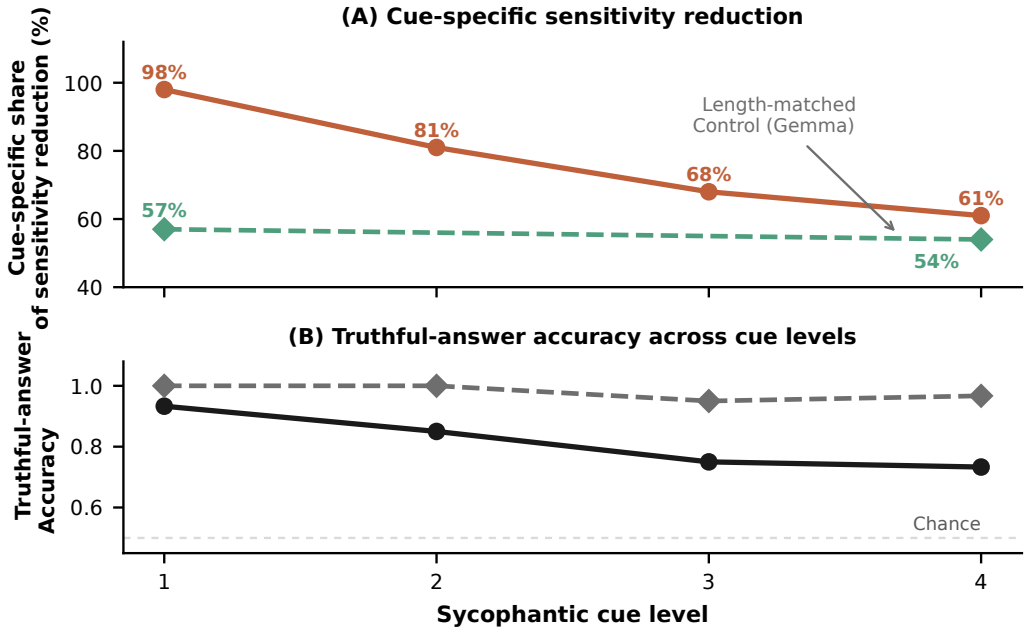


Figure 5: Truth-direction sensitivity and truthful-answer accuracy by cue level. (A) Cue-specific share of sensitivity reduction, measured against length-matched neutral prompts; Qwen2.5 (Yang et al., 2024) in terracotta (■), Gemma-2 (Gemma Team, 2024) in green (■). The share falls as the lengthening cue contributes more, while cue content stays the larger term. (B) Truthful-answer accuracy over the same cue levels, Qwen2.5-7B in black (■) and Gemma-2-9B in grey (■); the dashed line marks chance.

Qwen2.5-7B still answers truthfully at 0.933 and follows the stated preference at chance level, while the intervention effect has already fallen by more than a fifth. Its accuracy declines at higher cue levels. On Gemma-2-9B the divergence is starker still, since accuracy barely moves at all across the ladder while the intervention effect declines throughout.

Reading the two together. PSI compares the ordering of changes along the pressure ladder: at cue levels where truthful-answer accuracy remains high, the intervention effect has already changed measurably, and it continues to change over levels where accuracy moves little. This ordering is sensitivity attenuation under preserved accuracy: the directional response changes at cue levels where truthful-answer accuracy remains high.

What the probe reports over the same range. Decodability is 1.000 at every cue level and in every condition. Across this stable probe readout, sensitivity varies by a factor of several. PSI exposes that pressure-dependent response while tracking the continued availability of the decoded information.

6 Projection-based truth recovery

Effect of amplification. Under preference-conditioned prompts we scale each item's own projection onto the truth direction (Eq. 7) over the amplification settings shown in Figure 6. Truthful-answer accuracy rises from 0.500 to 0.850 on Gemma-2-9B at the fourth cue level and from 0.700 to 0.883 on Qwen2.5-7B at the second (Figure 6).

Specificity controls. A norm-matched random direction gives gains of +0.017 on Gemma-2-9B and +0.033 on Qwen2.5-7B; neither is statistically significant. True and false items both improve. On Gemma-2-9B at the fourth cue level, true-item accuracy rises from 0.500 to 0.833 and false-item accuracy from 0.500 to 0.867. The improvement spans both answer classes, supporting recovery through the item-specific projection. Negative amplification lowers accuracy, matching the sign of the operator.

Recovery under preference conditioning. The amplification experiments test the practical utility of the fitted direction under preference-conditioned prompts, where the user cue competes with the

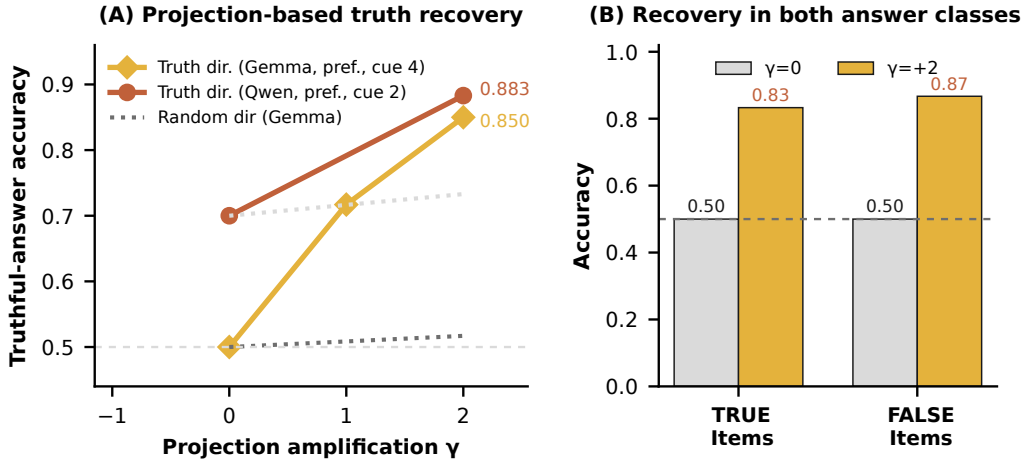


Figure 6: Accuracy under item-specific projection amplification. (A) Truthful-answer accuracy against amplification γ for the truth direction and a norm-matched random direction, on Qwen2.5 (Yang et al., 2024) and Gemma-2 (Gemma Team, 2024). (B) True and false items on Gemma-2-9B at the fourth cue level, at $\gamma = 0$ and $\gamma = +2$.

truthful answer. Their gains establish projection-based truth recovery: an item-specific intervention improves truthful answers even as sensitivity to the direction varies across prompts. Together with the matched random and negative-amplification controls, the result links recovery to the direction and the sign of the projection update.

7 Results across model sizes and families

Across Qwen scales. The pattern is stable across the three Qwen2.5 sizes (Table 2). Preference-conditioned sensitivity retains 76 to 80% of the honest value at 7B, 14B and 32B, with ρ between $+0.76$ and $+0.80$: the answer stays substantially responsive to the truth direction while also responding to the cue.

Across families. Gemma-2-9B shows the same behavioural signature with a different sensitivity profile. Preference-conditioned ρ is -0.15 and κ is 6.38. The answer is far more responsive to the cue than to the truth direction, and the preference term in Eq. 5 dominates on every evaluated item. PSI therefore identifies different sensitivity profiles across the two measured model families.

Models meeting the inclusion criteria. The behavioural screen identifies four instruction-tuned models for the three-condition comparison: three Qwen2.5 sizes and Gemma-2-9B. Among the five tested non-Qwen models, Gemma-2-9B meets the inversion criterion at the no-cue level. Appendix A reports the screen for Mistral-7B, Phi-3-mini, OLMo-2-7B and Llama-3.1-8B alongside the included models.

Cue overrides the inversion instruction. Presenting the sycophantic cue to a model under the inversion prompt produces answers that follow the cue rather than the instruction. We therefore measure inversion at the no-cue level throughout. This override shows that the user-side preference cue can redirect answers even under an explicit instruction to state falsehoods.

Table 2: Signed intervention effects by model at $\alpha^* = 1.0$. Retention is the preference-conditioned effect over the honest effect, and ρ_{pref} is the preference-conditioned normalized signed sensitivity. Models: Qwen2.5 (Yang et al., 2024), Gemma-2 (Gemma Team, 2024).

model	honest	preference	inversion	retention	ρ_{pref}
Qwen2.5-7B	$+13.09$	$+10.14$	-6.28	77%	$+0.79$
Qwen2.5-14B	$+11.84$	$+9.02$	-5.71	76%	$+0.76$
Qwen2.5-32B	$+13.91$	$+11.16$	-6.94	80%	$+0.80$
Gemma-2-9B	$+10.72$	$+6.79$	-5.44	63%	-0.15

8 DISCUSSION

PSI characterizes truth-related representations through three connected contributions. Decodability–sensitivity dissociation identifies stable probe accuracy alongside changing intervention responses. Sensitivity attenuation under preserved accuracy identifies pressure levels at which the response to the truth direction has weakened while truthful answering remains strong. Projection-based truth recovery demonstrates that amplifying the existing representation can improve answers under preference conditioning. These results connect decodable information, output response and intervention utility.

What each measurement tells the reader. Probe accuracy measures whether a fitted linear readout can recover the truth label. Truth-direction sensitivity measures the change in the True-minus-False answer logit under a two-sided displacement along that direction. Truthful-answer accuracy measures the model's actual decisions. The first two measurements share a direction but examine different properties: recoverable information and output response. Reporting all three makes their agreement or divergence visible. For example, at the first cue level on Qwen2.5-7B, probe accuracy is 1.000 and truthful-answer accuracy is 0.933, while the intervention effect has already fallen by more than one fifth. The sensitivity measurement makes this change visible alongside both accuracy readouts.

Reading pressure effects. The honest cue ladder keeps the system's request for truthful answers fixed while increasing user pressure. This separates the pressure comparison from the comparison between honest, inversion and preference-conditioned system prompts. Length-matched neutral prompts provide a reference for the extra text introduced by stronger cues. The larger reduction under sycophantic cues links the additional attenuation to cue content. The comparison concerns where changes occur along the cue ladder; sensitivity is an answer-logit difference, whereas accuracy is a fraction of correct answers. PSI thus identifies changed directional responses alongside accurate decisions.

Interpreting sensitivity and recovery. The two-sided intervention measures the response to an imposed displacement at the specified scale. Measuring that direction's contribution during an unperturbed forward pass requires a different operation, such as projection ablation comparing $\mathrm{score}(h)$ with $\mathrm{score}(h - (h \cdot \hat{d})\hat{d})$. Recovery answers a further question: whether strengthening each item's existing projection improves its answer. The recovery evidence is obtained under preference conditioning; the attenuation evidence is obtained along the honest cue ladder. Testing amplification along that same honest ladder would directly connect these two results.

Why amplify each item's projection? The update in Eq. 7 depends on both the sign and magnitude of the projection already present in the activation. Positive amplification increases its magnitude while preserving its sign. A constant vector addition instead shifts all items in the same direction, irrespective of their initial projections. This tests whether strengthening the decoded representation benefits both answer classes. On Gemma-2-9B, true-item accuracy rises from 0.500 to 0.833 and false-item accuracy from 0.500 to 0.867. Matched random directions produce much smaller gains, and negative amplification lowers accuracy. These controls support the usefulness of the fitted projection and the importance of the update's sign.

Evaluation resolution. Several accuracy conditions contain approximately sixty evaluation items, giving accuracy increments of about 0.017. Confidence intervals use item-level bootstrap resampling and describe uncertainty across evaluation items. Repeated forward passes, and comparisons between an unmodified pass and a pass after restoring patched activations, have maximum per-item differences of 0.000 at the reported precision. The accompanying scripts record condition-specific counts, selected layers and bootstrap units. These details distinguish the resolution of the evaluation set from numerical repeatability and make small accuracy differences easier to interpret.

Scope and use of PSI. We evaluate single-token true/false judgments on four instruction-tuned models from two families that meet the behavioural inclusion criteria. Preference conditioning models competition between a user's preference and the truthful answer, an operational setting related to truth-indifferent behaviour (Frankfurt, 2005). For broader tasks, PSI offers three assessment questions: is truth-related information decodable, does the output response change under pressure, and can projection amplification improve truthful decisions?

AI USE STATEMENT

We used generative AI tools as coding and writing assistants: to help implement and debug analysis and plotting scripts, and to draft and revise prose. The experimental design, the choice of controls, and the interpretation of results are the authors' own. AI-assisted code was reviewed by the authors and its outputs checked against our experiment records; AI-drafted text was edited and verified by the authors. The authors take responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

All magnitudes are reported at a single intervention scale $\alpha^* = 1.0$, selected as described in Section 3. The direction construction, the decodability and sensitivity measures, the two scale-free readouts, the competition prediction, the length attribution and the amplification operator are defined by equation in Section 3. The two-null comparison is reported in Table 1. The behavioural inclusion criteria, per-cue readouts and eight-model screen are given in Appendices A to C, and the layer-selection rule and split protocol in Section 3, with evaluation-count and bootstrap details in Section 8. Models are public (Qwen2.5 at 7B, 14B and 32B; Gemma-2-9B). The analysis uses inference, probing and activation patching. Directions are fit by difference of means over a few hundred contrastive prompts on a held-out split. Appendix D records the design decisions behind the final protocol.

ETHICS STATEMENT

This work studies truth-related representations under user pressure. Projection amplification increases truthful answering in the tested preference-conditioned settings. The cue prompts are standard sycophancy elicitations used to vary pressure in a controlled way. PSI supports evaluations that jointly measure decodability and truth-direction sensitivity. We use public models and factual true/false stimuli, with no human subjects and no personal data. Directional interventions can also be applied to user-preference directions, a potential misuse relevant to this white-box method. The recovery experiments target truthful answering.

REFERENCES

Marah Abdin et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*, 2024. URL https://arxiv.org/abs/2404.14219.

Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Waiss Azizian, Michael Kirchhof, Eugene Ndiaye, Louis Bethune, Michal Klein, Pierre Ablin, and Marco Cuturi. The Geometries of Truth Are Orthogonal Across Tasks. *arXiv preprint arXiv:2506.08572*, 2025. URL https://arxiv.org/abs/2506.08572.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is Universal: Robust Detection of Lies in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2407.12831. arXiv:2407.12831.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

James Campbell, Richard Ren, and Phillip Guo. Localizing Lying in Llama: Understanding Instructed Dishonesty on True-False Questions Through Prompting, Probing, and Patching. *arXiv preprint arXiv:2311.15131*, 2023. URL https://arxiv.org/abs/2311.15131.

Harry G. Frankfurt. *On Bullshit*. Princeton University Press, 2005.

Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting Strategic Deception Using Linear Probes. *arXiv preprint arXiv:2502.03407*, 2025. URL https://arxiv.org/abs/2502.03407.

Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2301.04213. arXiv:2301.04213.

Haoran Huan, Mihir Prabhudesai, Mengning Wu, Shantanu Jaiswal, and Deepak Pathak. Can LLMs Lie? Investigation beyond Hallucination. *arXiv preprint arXiv:2509.03518*, 2025. URL https://arxiv.org/abs/2509.03518.

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

Kieron Kretschmar, Walter Laurito, Sharan Maiya, and Samuel Marks. Liars' Bench: Evaluating Lie Detectors for Language Models. *arXiv preprint arXiv:2511.16035*, 2025. URL https://arxiv.org/abs/2511.16035.

Benjamin A. Levinstein and Daniel A. Herrmann. Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks. *Philosophical Studies*, 182:1539–1565, 2025. doi: 10.1007/s11098-023-02094-3.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.03341. arXiv:2306.03341.

Kaiqu Liang, Haimin Hu, Xuandong Zhao, Dawn Song, Thomas L. Griffiths, and Jaime Fernández Fisac. Machine Bullshit: Characterizing the Emergent Disregard for Truth in Large Language Models. *arXiv preprint arXiv:2507.07484*, 2025. URL https://arxiv.org/abs/2507.07484.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://arxiv.org/abs/2109.07958. arXiv:2109.07958.

Xianxuan Long, Yao Fu, Runchao Li, Mu Sheng, Haotian Yu, Xiaotian Han, and Pan Li. When Truthful Representations Flip Under Deceptive Instructions? *arXiv preprint arXiv:2507.22149*, 2025. URL https://arxiv.org/abs/2507.22149.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2310.06824.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.02707. arXiv:2410.02707.

Ethan Perez, Sam Ringer, Kamile Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations (ICLR)*, 2024.

Team OLMo et al. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2024. URL https://arxiv.org/abs/2501.00656.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A Behavioural inclusion criteria and model screen

Before any sensitivity is computed for a model, the prompts must induce the conditions they name. At a candidate cue level we require honest truth accuracy of at least 0.8, preference-conditioned answers that track the stated preference, and instructed inversion that inverts at the no-cue level, each by an exact test at $\alpha = 0.05$. Table 3 reports the screen over the eight instruction-tuned models tested. Three Qwen2.5 sizes and Gemma-2-9B meet the criteria. Mistral-7B, Phi-3-mini, OLMo-2-7B and Llama-3.1-8B fail the inversion criterion under this elicitation protocol; the three-condition comparison uses the four models that pass the screen.

Table 3: Model screen. The column reports inversion accuracy under the inversion prompt at the no-cue level, which is where we measure inversion throughout; a value near 0.5 means the prompt does not induce the condition. Models: Qwen2.5 (Yang et al., 2024), Gemma-2 (Gemma Team, 2024), Mistral (Jiang et al., 2023), Phi-3 (Abdin et al., 2024), OLMo-2 (Team OLMo et al., 2024), Llama-3.1 (Grattafiori et al., 2024).

model	inversion at cue 0	meets criteria
Qwen2.5-7B	0.875	yes
Qwen2.5-14B	0.850	yes
Qwen2.5-32B	0.867	yes
Gemma-2-9B	0.733	yes
Mistral-7B	0.508	no
Phi-3-mini	0.067	no
OLMo-2-7B	0.500	no
Llama-3.1-8B	0.500	no

B Per-cue readouts

Table 4 gives ρ and κ across cue levels for the two families. On Qwen2.5-7B the preference-conditioned κ crosses one between the first and second cue levels. On Gemma-2-9B the preference-conditioned ρ is near zero from the first cue level and κ is large throughout, so the answer responds mainly to the cue.

On Gemma-2-9B every evaluated item is preference-dominated; with no remaining comparison group, the enrichment statistic is undefined. The item-level competition results are reported in Section 4; Table 4 provides the corresponding per-cue readouts.

Table 4: Normalized signed sensitivity ρ and competition ratio κ by cue level for the honest and preference-conditioned prompts. Inversion is measured at the no-cue level and quoted once in Table 2.

model	cue	ρ honest	κ honest	ρ pref.	κ pref.
Qwen2.5-7B	1	$+1.00$	0.18	$+0.79$	0.54
Qwen2.5-7B	2	$+1.00$	0.15	$+0.61$	1.25
Gemma-2-9B	1	$+1.00$	0.01	-0.15	6.38
Gemma-2-9B	2	$+1.00$	0.05	-0.11	10.73
Gemma-2-9B	3	$+1.00$	0.21	-0.05	32.17
Gemma-2-9B	4	$+1.00$	0.02	-0.08	20.09

C NULL COMPARISONS AND SCALE SELECTION

The intervention-scale sweep and null comparisons are reported in Section 3 and Table 1. Additional length-control and scoring checks follow.

Length attribution in absolute terms. The shares in Section 5 are attributions of an absolute reduction. At the first cue level on Qwen2.5-7B a length-matched neutral prompt reduces the effect by $+0.134$ (CI $[+0.01, +0.27]$) while the sycophantic cue of the same length reduces it by $+7.206$ (CI $[+5.76, +8.64]$); at the second cue level the neutral cost is unchanged and the cue cost is $+10.509$ (CI $[+8.33, +12.71]$).

Reproducibility of the scoring. Evaluation granularity and repeated-pass checks are reported in Section 8.

D ADDITIONAL PROTOCOL DETAILS

The layer-selection and three-split protocol are described in Section 3. Evaluation granularity and bootstrap details are described in Section 8.

Section 8 explains the interpretation of sensitivity, projection amplification and evaluation resolution.

The experiments retain the chat context in which user cues operate, so the three prompt conditions are compared within the same interaction format. Length-matched neutral prompts in Eq. 6 quantify the length contribution, while the additional reduction under sycophantic cues identifies the cue-specific component.