Static Unit-Scale Bias Steering Transfers Poorly to a Reasoning-Distilled LLM

Anonymous Author(s)
Affiliation
Address
email

Abstract

Activation steering is a leading technique for controlling LLM behavior, but its reliability in reasoning-distilled checkpoints is unclear. We study a common intervention class—static continuous addition of a pre-generation linear-probe direction—and find that a cognitive-bias direction that controls an instruction-tuned model does not transfer cleanly to a reasoning-distilled sibling under the same protocol. We construct a 470-item contrastive benchmark spanning 11 bias categories (base-rate neglect, conjunction fallacy, framing, and others) and compare matched-architecture pairs (Llama-3.1-8B-Instruct vs. R1-Distill-Llama-8B, OLMo-3 7B/32B Instruct vs. Think) and Qwen-3-8B's thinking toggle. Behavioral lure rates are scorer- and benchmark-scope dependent: under final-answer rescoring of preserved 470-item raw-response artifacts, R1-Distill has lower overall lure than Llama (25.5% vs. 33.2%) but remains highly vulnerable on base-rate and conjunction items; on the full 470-item OLMo-32B scale run, lure falls from 19.6% to 0.4%. Lure suppression is not treated as correctness; accuracy and other-response rates are analyzed jointly. The main result characterizing the scoped dissociation is that probe-direction steering on the three vulnerable categories produces a monotonic 37.5pp dose-response in Llama (lure rates span 31% at α=+5 to 69% at α=$-$5, zero incoherent outputs), while the original static continuous intervention in R1-Distill, applied across four candidate layers including its probe peak, yields only small non-monotonic lure-rate fluctuations (5.0pp full-sweep range at L31), not a stable dose-response; uncalibrated final-answer-span P0/T0 diagnostics likewise show no endpoint effect but have near-zero prompt-prefill KL. Appendix diagnostics additionally show that, in the available OLMo-family 32B comparison, the larger Instruct checkpoint has higher scored lure rate than 7B (14.9%$\rightarrow$19.6%) while the Think checkpoint remains near-zero (0.4%). Diagnostic analyses show that Qwen-3-8B's hard think/no-think template induces non-transferring `P0` geometries and that within-CoT linear separability is non-stationary, but these diagnostics are not treated as evidence for a template-invariant semantic axis or causal CoT stages. These results support a narrow claim: a static unit-scale probe direction that steers Llama can fail to provide comparable behavioral control in R1-Distill-Llama-8B even when the same bias distinction remains linearly decodable. They do not rule out calibrated dynamic, broader multi-layer, SAE-feature, or nonlinear interventions.

1 Introduction

Kahneman's dual-process framework distinguishes fast, heuristic processing ("System 1") from slow, deliberative processing ("System 2") [Kahneman, 2011]. While the binary framing is an acknowledged oversimplification (modern accounts treat this as a *continuum* of processing intensity

Submitted to ICML 2026 Mechanistic Interpretability Workshop. Do not distribute.

[Melnikoff and Bargh, 2018, Evans and Stanovich, 2013, De Neys, 2023]), the core observation that cognitive systems flexibly allocate effort, with measurable consequences for bias susceptibility, remains robust. De Neys' conflict detection paradigm [De Neys, 2012, 2017] demonstrates that humans often *detect* conflicts between heuristic and normative responses even when they fail to override the heuristic, a dissociation between conflict *monitoring* and conflict *resolution* with clear neural correlates in the anterior cingulate cortex [Botvinick et al., 2001].

LLMs exhibit analogous patterns: instruction-tuned models reproduce human-like errors on conflict items while succeeding on matched controls [Hagendorff et al., 2023], and reasoning-trained models reduce many such lures [DeepSeek-AI, 2025], though this reduction is bias-specific rather than universal [Kim et al., 2025]. A growing literature documents these parallels [Coda-Forno et al., 2025, Ziabari et al., 2025], and Huang et al. [2026] showed that bias-relevant directions are probeable and steerable. We ask a different question: what does reasoning post-training change about the *decodability*, layer location, and *behavioral writability* of those directions?

Yet behavior alone cannot distinguish genuine processing-mode differences from surface mimicry: chain-of-thought traces are often unfaithful, with only $\sim$2.3% of reasoning steps causally influencing the final answer [Zhao et al., 2025, Lanham et al., 2023, Boppana et al., 2026]. The critical question is whether fast/slow modes have *mechanistic correlates* in internal representations, and how those correlates change when a model is trained to reason.

Recent work on activation steering [Turner et al., 2023, Panickssery et al., 2024, Cox et al., 2026] has shown that targeted interventions on internal representations can modulate behavior, with pre-CoT probes achieving $>$0.9 AUC and answer-flipping rates above 50% on instruction-tuned models [Cox et al., 2026]. This raises a natural question for the dual-process setting: if linear probes identify a direction in activation space that separates S1-like from S2-like processing, does a standard static intervention along that direction modulate bias susceptibility? If so, the representational findings move from correlational to intervention-based evidence, providing evidence that the probe direction captures more than an epiphenomenal surface-feature correlation under the tested protocol.

We address this question with a natural experiment: Llama-3.1-8B-Instruct and DeepSeek-R1-Distill-Llama-8B share identical architectures (32 layers, 32 attention heads, 4096 hidden dimensions) but differ in reasoning distillation training. Representational differences are therefore less likely to be driven by architecture alone. This controls architecture, not the full intervention: distillation data, CoT formatting, optimization, and answer style also change. We make six contributions:

1. A **contrastive cognitive bias benchmark** of 470 items (235 matched conflict/control pairs) across 11 categories spanning four heuristic families (representativeness, cognitive reflection, decision framing, and loss/availability), with novel isomorphs to resist memorization.
2. **Behavioral validation** showing that scored lure rates are model-, scorer-, and benchmark-scope dependent: final-answer rescoring of preserved raw-response artifacts gives lower overall lure for R1-Distill than Llama (25.5% vs. 33.2%) but high residual vulnerability on base-rate and conjunction items; OLMo-3-7B shows a stronger Instruct–Think drop (14.9% $\rightarrow$ 0.9%). We report accuracy and other-response rates so lure suppression is not equated with correctness.
3. **Probing analysis** showing high conflict/control separability in both models (Llama AUC = 0.974; R1-Distill AUC = 0.930), with cross-prediction, text baselines, and control tasks used as checks against purely surface-level explanations.
4. **Held-in evidence via static probe-direction steering**: in Llama, steering along the probe weight vector produces a 37.5pp bidirectional behavioral swing; in R1-Distill, the corresponding static `P0` probe direction is linearly decodable (AUC = 0.930) but does not show a stable dose-response at any of four tested layers (L14/L25/L28/L31) under continuous 2048-token intervention. A post-hoc `P0`-only diagnostic on the same 80 conflict items leaves R1 verdicts unchanged at $\alpha \in \{-5, 0, +5\}$, but this diagnostic is uncalibrated and near-zero-KL.
5. **Diagnostic analyses** showing two limits on simple probe narratives: Qwen 3-8B's hard think/no-think template yields non-transferring `P0` geometries, and within-CoT probe AUC is non-stationary over a generated trace.

Appendix diagnostics also compare available **OLMo 32B** checkpoints, finding higher scored lure rates for the larger Instruct checkpoint in this family while the Think checkpoint remains near-zero

on scored lures. Exploratory SAE and attention entropy analyses are reported as diagnostics rather than primary evidence.

2 Related Work

Dual-process cognition in LLMs. Hagendorff et al. [2023] first documented human-like cognitive biases in LLMs, finding that instruction-tuned models reproduce representativeness and anchoring effects that vanish in later ChatGPT versions. Coda-Forno et al. [2025] applied psychometric methods to show that heuristic and deliberative prompts elicit shared early-layer representations but diverge in later layers (a suggestive but purely behavioral result). Ziabari et al. [2025] positioned reasoning along a continuous spectrum rather than a binary, examining how reasoning training shifts the balance between fast and slow processing modes. Brady et al. [2025] review the dual-process lens for LLM decision-making in *Nature Reviews Psychology*, noting that LLMs mimic both System 1- and System 2-like responses but through mechanisms distinct from human cognition. Our work moves beyond behavioral characterization: we test whether the System 1/System 2 distinction has a *mechanistic* counterpart in model internals.

Cognitive bias benchmarks and probing. Malberg et al. [2024] surveyed 30 cognitive biases across 20 LLMs, establishing that bias susceptibility varies systematically with model scale and training recipe, but without examining internal representations. Most directly relevant, Huang et al. [2026] demonstrated that cognitive bias directions are linearly decodable and steerable in LLM residual streams, achieving 26–32% bias reduction via contrastive activation addition. We differ from CogBias in three respects: (i) architecturally matched base–reasoning model pairs that hold architecture fixed while varying the post-training recipe, (ii) Hewitt–Liang control tasks and cross-domain transfer tests as probe validity checks [Hewitt and Liang, 2019], and (iii) a different causal question: whether the discriminative probe boundary itself is bidirectionally writable, in addition to comparing against a CogBias-style mean-difference vector.

Activation steering and representation engineering. Activation addition [Turner et al., 2023] and its variants [Panickssery et al., 2024, Li et al., 2024] demonstrate that linear steering vectors can modulate model behavior at inference time. CogBias [Huang et al., 2026] applied this to cognitive bias (26–32% reduction) using mean-difference vectors. Probe-direction steering asks whether the learned discriminative boundary is behaviorally writable; mean-difference activation addition remains the stronger one-sided mitigation baseline in our Llama control. Pre-CoT probing and steering work shows that hidden states before generated reasoning can encode answers or downstream reasoning success [Cox et al., 2026, Afzal et al., 2025, Zhang et al., 2025]; our narrower question is whether a pre-generation conflict/control direction remains behaviorally writable during reasoning-trained generation. For SAE analysis, we adopt Ma et al.'s falsification protocol [Ma et al., 2026] as a mandatory filter, since 45–90% of purported "reasoning features" are spurious token-level artifacts [Méloux et al., 2025].

Conflict detection in cognitive science. Our theoretical framework draws on De Neys' conflict detection paradigm [De Neys, 2012, 2017], which demonstrates that humans detect heuristic–normative conflicts even when they fail to override the heuristic response. Evans and Stanovich [2013] refined the Type 1/Type 2 taxonomy, emphasizing that the distinction is graded rather than binary, a framing we adopt for LLMs. Botvinick et al. [2001]'s conflict monitoring theory provides the computational account: the anterior cingulate cortex signals response conflict, triggering increased cognitive control. We test the LLM analogue of this account: do models show internal signatures of conflict detection (elevated attention entropy, reduced probe confidence) even when producing heuristic-driven responses?

3 Benchmark

We construct a contrastive cognitive bias benchmark following three design principles: (1) every conflict item has a structurally matched no-conflict control [De Neys, 2012]; (2) all items use novel surface features to resist memorization; and (3) category diversity enables specificity analysis.

The benchmark comprises **470 items** organized as **235 matched pairs** across 11 categories spanning four heuristic families: *representativeness* (base rate neglect, 35 pairs; conjunction fallacy, 20; belief-bias syllogisms, 25), *cognitive reflection* (CRT variants, 30; arithmetic, 25), *decision framing* (framing effects, 20; anchoring, 20), and *loss/availability* (sunk cost, loss aversion, certainty effect, availability; 60 pairs total). Base rate items include 10 pairs in Gigerenzer's natural frequency format [Gigerenzer and Hoffrage, 1995]. Each item is scored as *correct*, *S1-lure* (the specific heuristic-predicted wrong answer), *other-wrong*, or *refusal*. Classic problems (e.g., the original bat-and-ball) serve only as contamination baselines.

Item counts and analysis subsets. Our benchmark contains 470 items across 11 categories. Primary probe analyses use the 3 vulnerable categories (160 items). Cross-prediction uses all 7 original categories (330 items; 165 pairs). The full 11-category set includes 4 additional categories added to test domain generality.

4 Methods

4.1 Models

Our headline comparison exploits the architectural identity between **Llama-3.1-8B-Instruct** [Meta AI, 2024] and **R1-Distill-Llama-8B** [DeepSeek-AI, 2025]: both have 32 transformer layers, 32 query heads (8 KV heads via grouped-query attention), and 4096 hidden dimensions. They therefore control architecture but not the full training intervention: R1-Distill-Llama-8B comes from a different post-training pipeline, including distillation from a reasoning-trained teacher, CoT formatting, objective choices, and answer-style changes. We additionally evaluate **Qwen 3-8B** in both thinking (`/think`) and non-thinking (`/no_think`) modes, providing a same-weights comparison whose hard chat-template switch changes both reasoning behavior and the `P0` positional context. For cross-architecture replication, we evaluate **OLMo-3-7B-Instruct** and **OLMo-3-7B-Think** [Team OLMo, 2025], which share the same OLMo base but differ in reasoning training. For scale analysis, we additionally evaluate **OLMo-3.1-32B-Instruct** and **OLMo-3-32B-Think** (64 layers, 4.5$\times$ the parameters of the 7B variant), enabling a within-OLMo scale comparison on identical benchmark items; because the 32B Instruct and Think checkpoints use different OLMo release lines, this is not a pure size or training-objective ablation.

Prompting and decoding. All models use greedy decoding (`do_sample=False`, `temperature=1.0`) with no system prompt. For Qwen 3-8B, we use the hard mode switch (`enable_thinking=True/False` in `apply_chat_template`). In no-think mode, this appends an empty `<think>\n\n</think>\n\n` block after the assistant turn marker, so the token sequences *diverge* after that marker: think-mode `P0` is the newline after `assistant`, while no-think-mode `P0` follows the injected `</think>` close tag. Cross-mode probe comparisons are therefore template-conditioned geometry tests, not clean same-position tests of an invariant semantic conflict axis. This is distinct from Qwen's soft switch (`/think`, `/no_think` tokens). For R1-Distill, we parse any generated `<think>...</think>` span, and conservatively treat an open but unclosed thinking span as no final answer. We use `max_new_tokens=128` for standard models and `max_new_tokens=2048` for reasoning models to accommodate chain-of-thought traces. We acknowledge this departs from vendor-recommended settings (e.g., DeepSeek recommends `temperature=0.5-0.7`); Appendix E reports multi-seed robustness analysis with sampled decoding (`temperature=0.7`) showing qualitatively consistent results. For R1 steering at the later tested layers (L25/L28/L31), the "other" response rate is 0% under this parser; the earlier L14 intervention is less fluent, with 30–31.25% "other" responses.

4.2 Activation Extraction

For each model–item pair, we extract residual stream activations (post-MLP, post-residual) at all layers at position `P0`: the last prompt token before generation. This "pre-decision" position captures the internal state from which the answer is generated. Forward passes use the requested model precision (bfloat16 where applicable), but HDF5 residual caches store bfloat16 requests as float16 because h5py has no native bfloat16 dtype. For within-CoT probing (Section 5.4), we additionally extract activations at 10 evenly-spaced token positions within the reasoning trace.

4.3 Linear Probes with Hewitt–Liang Controls

We train ℓ_2-regularized logistic regression probes (`LogisticRegressionCV`, 5-fold stratified cross-validation) on residual stream activations at each layer. The probe target is binary: *conflict* vs. *no-conflict* condition, restricted to the three vulnerable categories (base rate neglect, conjunction fallacy, syllogisms) where the standard model shows non-trivial lure rates.

Controls. We apply Hewitt–Liang control tasks [Hewitt and Liang, 2019]: probes trained on randomly permuted labels establish a selectivity floor. Selectivity (true AUC minus control AUC) below 5 percentage points indicates the signal reflects probe capacity, not representational structure. We report ROC-AUC as the primary metric with bootstrap 95% confidence intervals (1000 resamples).

Lower-vulnerability categories. We separately probe four categories that were lower-vulnerability in the original source runs (CRT, arithmetic, framing, anchoring) as a *negative control*: if probes achieve high AUC on categories where there is minimal behavioral difference, the signal is confounded by surface features rather than processing mode.

4.4 Probe-Direction Steering

To test behavioral control by a static linear probe direction, we extract the weight vector $\mathbf{w}$ from the trained logistic regression probe at the peak layer and use it as a steering direction. At inference time, we add $\alpha \cdot \hat{\mathbf{w}}$ (where $\hat{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|$) to the residual stream at the intervention layer, sweeping positive values toward S2-like and negative values toward S1-like. The coefficient α is a unit-direction residual multiplier, not a perturbation calibrated to equal KL shift or activation-norm change across models or layers; matched α sweeps should therefore be read as a fixed-protocol comparison rather than a matched-strength comparison. We evaluate the steered model on the 80 vulnerable-category conflict items and report lure-rate changes; specificity controls on matched non-conflict items appear in Appendix H.

5 Results

5.1 Behavioral Validation

Table 1 reports lure rates (proportion of S1-lure responses on conflict items). For Llama/R1/Qwen, we use the preserved 470-item raw-response artifacts and rescore only the final-answer span after any generated thinking trace; this avoids counting lure mentions inside reasoning text. OLMo rows use the original full-run verdict artifacts because raw responses were not cached. The compact table shows selected categories; full per-category rates appear in Appendix G.

Four findings emerge. First, behavioral vulnerability is not a simple reasoning/non-reasoning binary. Under final-answer rescoring of preserved 470-item raw responses, R1-Distill reduces overall lure relative to Llama (25.5% vs. 33.2%) but remains highly vulnerable on base rate (62.9%) and conjunction items (80.0%). This replaces the older compact seven-category source-run estimate, whose raw responses were not fully available for rescoring. Second, vulnerability is *category-specific*: base rate neglect and conjunction fallacy are consistently susceptible, whereas several other categories are model- and scorer-sensitive rather than uniformly immune. Third, different models show distinct profiles (for example, Qwen no-think has 95.0% conjunction lure but 0.0% syllogism lure), suggesting model-specific rather than universal heuristic behavior. Fourth, Qwen's within-model thinking toggle reduces overall lure from 31.1% to 7.7% under strict final-answer rescoring, but mostly by shifting unfinished traces into the 'other' category (56.6%) rather than by pure accuracy gain; conjunction remains vulnerable at 30.0%. Lure suppression is not automatically normative reasoning: "other" answers are counted as errors and reported in Appendix G. We therefore interpret lure, accuracy, and other-response rates jointly rather than treating lure-rate changes as sufficient evidence of reasoning success.

Cross-architecture replication. OLMo-3-7B-Instruct/Think [Team OLMo, 2025] independently show a stronger Instruct–Think drop: 14.9% $\rightarrow$ 0.9% lure rate on the full 470-item run. Probing shows the same mechanistic direction: OLMo Instruct AUC = 0.996 at L24, OLMo Think AUC = 0.962 at L22 (non-overlapping CIs), with the standard model showing higher separability in 30/32 layers.

Table 1: Behavioral results on conflict items. Bolded entries indicate vulnerable categories (>10% lure). Llama/R1/Qwen entries are final-answer rescoring of preserved 470-item raw-response artifacts; OLMo entries are the original 470-item verdict artifacts. "Overall" is a micro-average over all 235 conflict items. Behavioral rates are used as validation and sensitivity evidence, not as the causal steering result.

	Llama	R1-Distill	Qwen 3-8B		OLMo	OLMo
Category	**8B-Inst.**	**Llama-8B**	**no think**	**think**	**3-7B-Inst.**	**3-7B-Think**
Base rate	**88.6**	**62.9**	**71.4**	2.9	**46**	0
Conjunction	**55.0**	**80.0**	**95.0**	**30.0**	**50**	0
Syllogism	**32.0**	**20.0**	0.0	0.0	0	4
CRT variants	10.0	**16.7**	**13.3**	0.0	3	3
Arithmetic	0.0	0.0	0.0	0.0	0	0
Framing	**45.0**	**35.0**	**45.0**	5.0	0	0
Anchoring	10.0	5.0	**15.0**	0.0	5	0
Overall lure	**33.2**	**25.5**	**31.1**	7.7	**14.9**	**0.9**

	Llama	R1	Qwen-NT	Qwen-T	OLMo-I	OLMo-T
Overall Corr	59.6	68.1	59.1	35.7	75	88
Overall Lure	33.2	25.5	31.1	7.7	14.9	0.9
Overall Other	7.2	6.4	9.8	56.6	10	11

Implicit conflict detection. Llama's first-token probability is 4.2pp lower on conflict items (0.751 vs. 0.793), and entropy is 16% higher (0.948 vs. 0.817), an implicit "hesitation" signature that parallels De Neys' conflict-detection findings [De Neys, 2012] and elevated Anterior Cingulate Cortex (ACC) activation in humans facing dual-process conflicts [Botvinick et al., 2001]. This is a pre-generation hesitation signature consistent with conflict sensitivity, even though Llama often fails to resolve the conflict in its final answer.

5.2 Linear Probes and Cross-Prediction

Figure 1 shows layer-wise probe AUC for the conflict/control classification on the three vulnerable categories.

Llama peaks at AUC = 0.974 (95% CI [0.952, 0.992]) at L16; R1-Distill peaks lower at AUC = 0.930 [0.894, 0.960] at L31 (CIs overlap slightly but point estimates differ by 0.044). The 15-layer shift (L16 $\rightarrow$ L31 of 32) is architecturally suggestive but not a localization proof: it shows that the conflict/control distinction is most linearly accessible much later in R1-Distill, near the hidden state that seeds the chain-of-thought rollout [Zhang et al., 2025]. Qwen 3-8B achieves high separability in both modes (no-think: 0.954; think: 0.970; both at L34), but cross-mode probe transfer is at chance (AUC = 0.496, cosine $= -0.005$): probes trained in one mode do not generalize to the other. Because the no-think template injects four tokens before `P0` (Section 4), the two modes' `P0` positions differ in both absolute index and local context, inducing a geometric subspace change through RoPE and attention to the injected tokens. The non-transferability is therefore an expected consequence of the template difference, not evidence by itself that the semantic conflict representation is mode-specific. The useful observation is narrower: each positional context independently supports high within-mode separability ($\sim$0.97), while the hard mode switch prevents a single `P0` linear probe from transferring across templates.

Selectivity and text-only baseline. Legacy Hewitt–Liang control probes achieve AUC ≤ 0.66 at all layers, indicating that the true-probe signal is not explained by probe capacity alone under that control protocol. Because the repository now uses stricter shuffled-label control semantics, these selectivity values should be treated as a diagnostic rather than a fresh confirmatory statistic.[1]

[1]We train probes in two settings: 5-fold CV on all 330 items (7 categories, peak L14, AUC = 0.999), used for cross-prediction and steering; and bootstrap CI analysis on 160 vulnerable-category items (3 categories, peak L16, AUC = 0.974 [0.952, 0.992]). The steering direction uses the legacy plain-label-stratified L14 CV probe; the L16 value is the post-selected bootstrap peak used for reporting decodability.

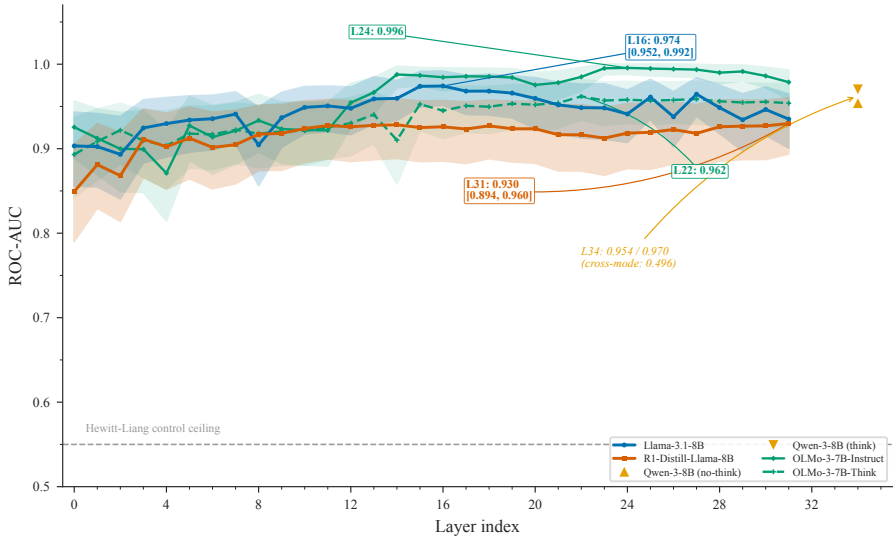

Figure 1: Layer-wise probe ROC-AUC (conflict vs. no-conflict, three vulnerable categories). Llama peaks at 0.974 [0.952, 0.992] (L16); R1-Distill at 0.930 [0.894, 0.960] (L31). Qwen 3-8B peaks at 0.954 (no-think) and 0.970 (think) at L34, but hard-template/position-shifted cross-mode transfer is at chance (0.496). Gray dashed: Hewitt–Liang control ceiling.

Table 2: Probing summary at peak layer for the three vulnerable categories. 95% CIs: bootstrap (1000 resamples) for Llama/R1/OLMo-7B rows; CV-fold $\pm 1.96\sigma$ for OLMo-32B rows (†; per-sample predictions not cached for bootstrap). Qwen CIs omitted (only layer-level aggregates stored). Layer-selection uncertainty is not propagated in any row. Selectivity = true AUC − control AUC. Cross-mode column: AUC when a probe trained on one Qwen mode is tested on the other. Text baselines: three methods on input text only (no activations)—see range in last row.

Model	**Peak L**	**Peak AUC**	**95% CI**	**Ctrl AUC**	**Select.**	**Cross**
Llama-3.1-8B-Instruct	L16	0.974	[0.952, 0.992]	0.63	0.343	—
R1-Distill-Llama-8B	L31	0.930	[0.894, 0.960]	0.55	0.377	—
Qwen 3-8B (no think)	L34	0.954	—	—	—	0.496
Qwen 3-8B (think)	L34	0.970	—	—	—	0.496
OLMo-3-7B-Instruct	L24	0.996	[0.988, 1.000]	—	—	—
OLMo-3-7B-Think	L22	0.962	[0.934, 0.982]	—	—	—
OLMo-3.1-32B-Instruct	L20	0.9999	$[0.9997, 1.000]^\dagger$	—	—	—
OLMo-3-32B-Think	L20	0.9978	$[0.9960, 0.9996]^\dagger$	—	—	—
Text baselines	—	0.820–0.863	—	—	—	—

Three text-only baselines cluster in a narrow band: a length-only probe (word count, 5-fold CV) achieves AUC = 0.863, a TF-IDF unigram+bigram logistic regression achieves 0.820, and a global sentence-transformer baseline achieves 0.840, all substantially below the combined vulnerable-category activation probe (0.999). All three text-only syllogism baselines collapse to below chance ($\leq$0.086), because conflict and control syllogisms are surface-matched in length ($d = +0.02$) and vocabulary. By contrast, a syllogism-only activation probe still reaches 0.936, while the combined vulnerable-category probe reaches 0.999. The high 0.82–0.86 text-baseline band is itself an important caveat: many conflict/control labels are recoverable from prompt surface statistics. Our claim is only that the tested residual-stream representation makes the distinction more linearly accessible than these text encoders. Since activations are deterministic functions of the input text and model weights, this is not evidence of information absent from the prompt, nor does it rule out richer surface-template confounds.

Probe summary. Table 2 summarizes the key probing metrics.

Cross-prediction and immune-category specificity. Training on *vulnerable* categories and testing on *immune* categories yields non-transfer (L14 AUC = 0.378). The below-chance AUC reflects a small polarity reversal inside a compressed immune cluster: immune conflict/control separation is only $0.03\times$ the vulnerable separation, while the immune score spread is about $0.25\times$ as large. Base rate and conjunction probes transfer near-perfectly to each other (AUC > 0.99), while transfer to syllogism is weaker (0.59–0.63).

Cross-model probe transfer. A Llama-trained probe achieves AUC = 0.920 on R1-Distill (layer 23, the peak transfer layer from a full 32-layer sweep); the reverse achieves 0.954 at layer 15 (also the peak transfer layer from a full sweep). The processing-mode direction is substantially shared across this same-architecture pair despite different post-training pipelines, contrasting with Huang et al. [2026]'s near-orthogonal finding across architectures.

5.3 Held-in Steering Evidence

We steer with the normalized coefficient vector from the vulnerable-category L14 logistic-regression probe (CV AUC = 0.960), adding $\alpha\hat{\mathbf{w}}$ to the residual stream for $\alpha \in \{-5, -3, -1, 0, +1, +3, +5\}$ on the 80 vulnerable conflict items. This is a held-in behavioral test: the probe is fit on the same 160-item family, so the causal claim rests on direction-specific controls rather than held-out generalization.

Figure 2 shows a monotonic Llama response: lure rate moves from 68.8% at $\alpha = -5$ to 31.2% at $\alpha = +5$ (37.5pp). The entire lure reduction converts to correct answers (other rate stays 0), and 50 random unit directions show no systematic endpoint effect (mean lure 52–57%, $\sigma \leq 8.7$pp). Two controls sharpen the interpretation. Probe steering produces zero lure responses on matched non-conflict items, although large steering can still degrade control accuracy (Appendix H). The CogBias-style mean-difference vector is a stronger one-sided debiaser (52.5%→11.25% lure at $\alpha = +5$) but is not bidirectional; negative α leaves lure at 53.75%. The probe direction is weaker for debiasing but better evidence of a behaviorally effective boundary under this protocol.

R1-Distill steering contrast: layer sweep. We sweep R1-Distill layers L14/L25/L28/L31 under continuous 2048-token steering in the original source-run scoring regime (Appendix A). No tested layer shows a coherent dose-response: lure fluctuates between 5–14%, and at the probe peak L31 it changes from 10.0% to 11.25% between $\alpha = -5$ and $+5$ (5.0pp full range), versus Llama's 37.5pp endpoint swing. The direction that controls Llama therefore does not yield a stable dose-response in R1 under this static continuous-addition protocol, and no matched R1 random-direction control band is included for this source-run sweep. However, the corrected final-answer-span audit changes the R1 baseline and is the stricter diagnostic.

As an exploratory post-hoc position-targeted check, we add the L31 vector only at `P0` and disable steering during generation. Across all 80 vulnerable conflict items, R1 gives identical verdicts at $\alpha = -5, 0, +5$ (56.2% lure; paired endpoint CI [0.0, 0.0]pp), and the pattern also holds at $\alpha = \pm 20$. This check argues against the narrow hypothesis that the continuous sweep failed only because later-token steering washed out the `P0` perturbation. It is not a clean held-out category-transfer claim because a probe-check showed below-chance held-out projection transfer for syllogism; moreover, $\alpha = \pm 5$ prompt-prefill KL is $< 10^{-6}$ and matched all-vulnerable R1 random-direction controls are not included. A narrower base-rate/conjunction P0-only random-control audit gives the same 0pp endpoint swing for the probe and 10 random directions, but it shares the near-zero-KL limitation (Appendix A). A completed high-alpha base-rate/conjunction diagnostic instead produces macroscopic prompt-logit KL and changes R1 behavior, reinforcing that the main result is a unit-scale transfer failure rather than evidence of general non-writability. The comparison is therefore not a matched-perturbation test: Llama shows a large behavioral swing, whereas the R1 `P0` diagnostic has near-zero prompt-logit impact. Thus the supported claim is a static unit-scale steering-transfer failure: linear readability persists, but this specific protocol does not provide matched behavioral control. It is not a claim against dynamic, calibrated multi-layer, SAE-feature, or learned nonlinear interventions.

5.4 Within-CoT Probing

At Qwen 3-8B L34, seven-position probes trace a non-monotonic path (Figure 3): P0 0.938 → T0 0.973 → T75 0.754 → Tend 0.971; P2 returns to chance (0.500). We interpret this conservatively: the boundary is non-stationary over the trace, but the AUC changes could reflect boundary rotation,

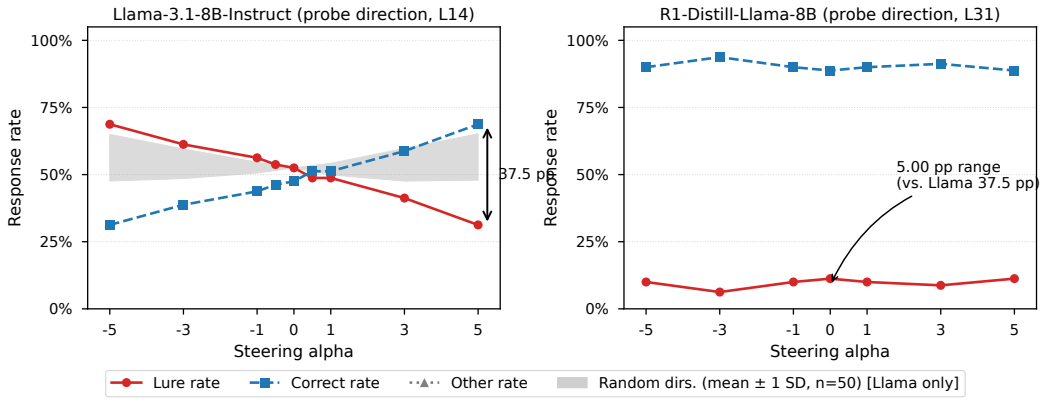

Figure 2: Probe-direction steering (80 vulnerable-category conflict items). *Left*: Llama-3.1-8B-Instruct (L14) shows a 37.5pp dose-response ($\alpha = -5$: 68.8% lure; $\alpha = +5$: 31.2%), with 100% of lure reduction converting to correct answers (other rate stays 0); 50 random directions (gray band, $n = 50$) show no systematic effect. *Right*: R1-Distill-Llama-8B (L31, its probe-peak layer) shows no coherent dose-response (5.0pp range, compared to Llama's 37.5pp) under the original static continuous-addition source-run scoring protocol.

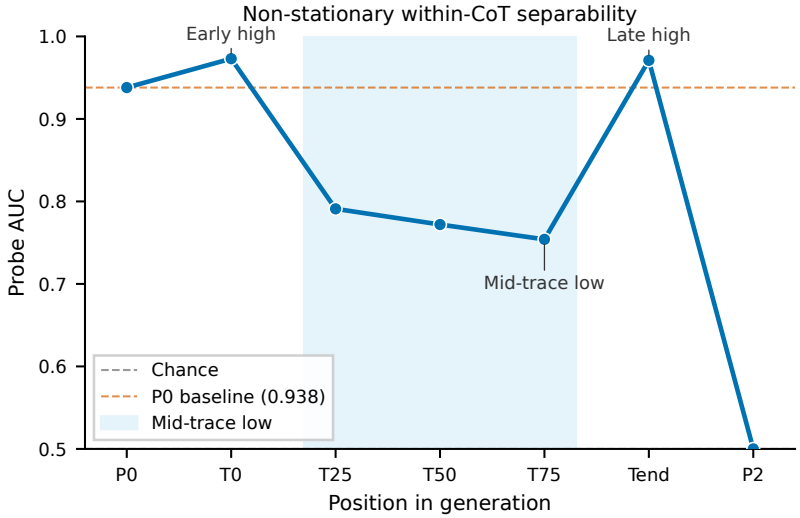

Figure 3: Within-CoT probe AUC for Qwen 3-8B (think mode) at L34. The trajectory shows an early rise (P0→T0, +3.5pp), a mid-trace dip (T0→T75, −22pp), and a late rebound (T75→Tend, 0.971). P2 (control) = 0.500.

dispersion, position effects, or other context-dependent shifts. They do not establish stepwise reasoning or causal use of specific CoT tokens [Boppana et al., 2026, Zhao et al., 2025]. Full trajectories in Appendix B.

References

Anum Afzal, Florian Matthes, Gal Chechik, and Yftah Ziser. Knowing before saying: LLM representations encode information about chain-of-thought success before completion. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphael Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. Reasoning theater: Disentangling model beliefs from chain-of-thought. *arXiv preprint arXiv:2603.05488*, 2026.

Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. Conflict monitoring and cognitive control. *Psychological Review*, 108(3):624–652, 2001.

Oliver Brady, Paul Nulty, Lili Zhang, Tomas Ward, and David McGovern. Dual-process theory and decision-making in large language models. *Nature Reviews Psychology*, 4:777–792, 2025. doi: 10.1038/s44159-025-00506-1.

Julian Coda-Forno, Zhuokai Zhao, Qiang Zhang, Dipesh Tamboli, Weiwei Li, Xiangjun Fan, Lizhu Zhang, Eric Schulz, and Hsiao-Ping Tseng. Exploring system 1 and 2 communication for latent reasoning in LLMs. *arXiv preprint arXiv:2510.00494*, 2025.

Kyle Cox, Darius Kianersi, and Adrià Garriga-Alonso. Decoding answers before chain-of-thought: Evidence from pre-CoT probes and activation steering. *arXiv preprint arXiv:2603.01437*, 2026.

Wim De Neys. Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, 7(1):28–38, 2012.

Wim De Neys, editor. *Dual Process Theory 2.0*. Routledge, London, 2017.

Wim De Neys. Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 46:e111, 2023.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241, 2013.

Gerd Gigerenzer and Ulrich Hoffrage. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684–704, 1995.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10):833–838, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2733–2743, 2019.

Fan Huang, Songheng Zhang, Haewoon Kwak, and Jisun An. CogBias: Measuring and mitigating cognitive bias in large language models. *arXiv preprint arXiv:2604.01366*, 2026.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

Su Hwan Kim, Sebastian Ziegelmayer, Felix Busch, Christian J Mertens, Matthias Keicher, Lisa C Adams, Keno K Bressem, Rickmer Braren, Marcus R Makowski, Jan S Kirschke, Dennis M Hedderich, and Benedikt Wiestler. LLM reasoning does not protect against clinical cognitive biases—an evaluation using BiasMedQA. *medRxiv*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Deep Ganguli, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

George Ma, Zhongyuan Liang, Irene Y. Chen, and Somayeh Sojoudi. Do sparse autoencoders identify reasoning features in language models? *arXiv preprint arXiv:2601.05679*, 2026.

Simon Malberg, Roman Poletukhin, Carolin M Schuster, and Georg Groh. A comprehensive evaluation of cognitive biases in LLMs. *arXiv preprint arXiv:2410.15413*, 2024.

David E Melnikoff and John A Bargh. The mythical number two. *Trends in Cognitive Sciences*, 22 (4):280–293, 2018.

Maxime Méloux, Giada Dirupo, François Portet, and Maxime Peyrard. The dead salmons of AI interpretability. *arXiv preprint arXiv:2512.18792*, 2025.

Meta AI. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

Team OLMo. OLMo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025.

Jiachen Zhao, Yiyou Sun, Weiyan Shi, and Dawn Song. Can aha moments be fake? identifying true and decorative thinking steps in chain-of-thought. *arXiv preprint arXiv:2510.24941*, 2025.

Alireza S Ziabari, Nona Ghazizadeh, Zhivar Sourati, Farzan Karimi-Malekabadi, Payam Piray, and Morteza Dehghani. Reasoning on a spectrum: Aligning LLMs to system 1 and system 2 thinking. *arXiv preprint arXiv:2502.12470*, 2025.

A R1 Continuous Layer-Sweep Steering Details

Table 3: R1-Distill layer-sweep steering (continuous, 80 vulnerable conflict items, original source-run scoring). Lure rate (%) at each α; none of the four tested layers shows a monotonic dose-response (full-sweep ranges: L14 = 8.8pp, L25 = 7.6pp, L28 = 5.0pp, L31 = 5.0pp; endpoint swing at L31: $10.0\% \rightarrow 11.25\% = 1.25$pp). Compare Llama L14: 68.8%→31.2% (37.5pp endpoint swing).

Layer	-5	-3	-1	0	$+1$	$+3$	$+5$
L14	6.2	10.0	7.5	6.2	5.0	7.5	13.8
L25	6.2	6.2	8.8	11.2	10.0	13.8	8.8
L28	8.8	6.2	10.0	11.2	7.5	11.2	10.0
L31	10.0	6.2	10.0	11.2	10.0	8.8	11.2

Position-targeted diagnostics. Table 4 reports exploratory post-hoc R1 diagnostics that perturb fewer positions than the continuous layer sweep. These runs are useful checks against the simplest washout objection, but they are not confirmatory: matched R1 random-direction controls for the continuous and position-targeted runs are not included, the `P0` KL shift is tiny at $|\alpha| = 5$, the all-vulnerable rows include syllogism despite below-chance held-out projection transfer in a probe-check, and answer-after-think covers only the 55 base-rate/conjunction conflict items.

A narrower unit-scale random-control audit on the 55 base-rate/conjunction conflict items gives 69.1% lure at $\alpha = -5, 0, +5$ for the probe direction, and 10 matched random directions also have 0pp endpoint swing (2.5–97.5% range [0.0, 0.0]pp). Because prompt-prefill KL remains $\leq 2\times10^{-6}$ and syllogism is excluded, this is a random-direction sanity check for the unit-scale diagnostic, not evidence of calibrated R1 resistance.

Table 4: R1-Distill position-targeted diagnostics. Entries are conflict-item lure rates (%); all rows use the L31 probe direction except the explicit L25+L28+L31 multilayer row. The final column reports mean prompt-prefill KL at $\alpha = +5$ when that position is perturbed; n/a means the protocol does not perturb prompt-prefill logits.

Protocol	**Items**	$\alpha = -5$	$\alpha = 0$	$\alpha = +5$	**KL@+5**
P0 only, L31	80	56.2	56.2	56.2	4.0×10^{-7}
T0 only, L31	80	56.2	56.2	56.2	n/a
P0 only, L25+L28+L31	80	58.8	56.2	50.0	5.3×10^{-6}
Answer-after-think, L31	55	70.9	69.1	65.5	n/a

High-alpha P0 diagnostic. Table 5 reports a completed, post-hoc base-rate/conjunction P0-only diagnostic that deliberately leaves the unit-scale regime. The result is not a calibrated confirmation of a mechanism: it has no random-direction controls, covers 55 conflict/control pairs rather than all three vulnerable categories, and uses very large perturbations whose prompt-prefill KL is around 10–13 with a 100% top-token flip rate. It is nevertheless important for scope. At $\alpha = -500$ or -1000, R1 lure falls from 69.1% to 18.2%; at $\alpha = +500$ or $+1000$, lure falls to 52.7%. Thus R1 is not claimed to be unwritable in general. The paper's main claim is that the unit-scale static direction that controls Llama does not transfer as a matched behavioral-control protocol to R1-Distill.

Table 5: R1-Distill high-alpha P0 diagnostic on base-rate/conjunction conflict items ($n = 55$), final-answer-span scoring. These very large perturbations produce macroscopic prompt-prefill KL and change behavior, delimiting the unit-scale claim rather than supporting a general non-writability conclusion.

α	lure	correct	other	mean KL
-1000	18.2	81.8	0.0	11.49
-500	18.2	81.8	0.0	10.09
0	69.1	30.9	0.0	0.00
$+500$	52.7	45.5	1.8	11.88
$+1000$	52.7	45.5	1.8	13.30

B Within-CoT Probing Details

This appendix details the within-chain-of-thought probing protocol for Qwen 3-8B (think mode). For each model–item pair, we extract residual stream activations at five canonical positions within the <think>...</think> trace: T0 (first token), T25, T50, T75, and Tend (final token before </think>). We train independent ℓ_2-regularized logistic regression probes (5-fold stratified CV) at each (layer, position) combination on the conflict vs. no-conflict classification ($n = 160$).

Table 6: Within-CoT probe AUC trajectory for Qwen 3-8B (think mode) at L34, 7 positions.

Position	**Description**	**Probe AUC**
P0	Pre-generation	0.938
T0	Thinking onset	**0.973**
T25	25% of trace	0.791
T50	Midpoint	0.772
T75	75% of trace	0.754
Tend	Final thinking token	0.971
P2	Control (chance)	0.500

The U-shaped trajectory (T0 peak, T25–T75 dip, Tend rebound) is inconsistent with a flat trace and with monotonic sharpening, but it is not by itself a causal trace-faithfulness test. The orthogonal P0 directions between Qwen's think/no-think modes (cross-mode AUC = 0.496) should be interpreted with the hard-template confound in mind: the result shows that template-specific P0 geometries do

not transfer linearly, while the within-trace curve shows that the think-mode representation changes substantially before the answer.

C Scale Analysis: Full Results

OLMo-3.1-32B-Instruct and OLMo-3-32B-Think (64 layers, 5120 hidden dims) were evaluated on the full 470-item benchmark. Lure rate rises from 14.9% (7B) to 19.6% (32B Instruct), with base rate neglect doubling (46%→74.3%). The 32B Think checkpoint has 0.4% scored lure rate. Probes achieve AUC = 0.9999 (Instruct L20) and 0.9978 (Think L20): the behavioral gap widens (19.2pp at 32B vs. 14.0pp at 7B) while the representational gap shrinks ($0.034 \rightarrow 0.0021$).

D SAE Features and Attention Entropy

We apply a Goodfire SAE (65,536 features) to Llama-3.1-8B-Instruct L19 `P0` residual activations. Reconstruction fidelity is moderate (MSE 0.011, explained variance 74.0%, mean L0 47.9 active features), so we treat feature findings as exploratory supporting evidence. After BH-FDR correction across features, 41 features show significant conflict/control activation differences; all 41 pass the Ma et al. token-injection falsification protocol (0 spurious under our threshold).

For attention, we test all 1,024 heads per model using rank-biserial correlation between conflict and control items, BH-FDR corrected at $q < 0.05$. Llama has 473 significant heads and 30 S2-like-specialized heads; R1-Distill has 515 significant heads and 57 S2-like-specialized heads. At the more conservative KV-group granularity, the counts are 13/256 for Llama and 24/256 for R1-Distill. These analyses support the representation-level story but are not causal evidence.

E Multi-Seed Robustness

Multi-seed evaluation with sampled decoding (temperature 0.7) on the 11-category robustness set shows that Llama's sampled lure rate is stable across seeds (27.5% ± 1.3pp) while R1-Distill's is also stable (12.1% ± 1.0pp) in the current robustness artifacts, but category profiles shift and these sampled artifacts have not been regenerated as a fresh final-answer-rescored behavioral table. Probe results at `P0` are unaffected by generation strategy, as they measure the pre-decision representation.

E.1 Exact Generation Parameters

Table 7 reports the exact generation keyword arguments used for each model family.

Table 7: Generation keyword arguments for all models. "Greedy" denotes the main-text configuration; "Sampled" denotes the robustness configuration used in this appendix.

Parameter	Greedy (main text)	Sampled (this appendix)
`do_sample`	`False`	`True`
`temperature`	1.0	0.7
`top_p`	—	0.95
`max_new_tokens`	128 (standard) / 2048 (reasoning)	same
System prompt	none	none

Vendor-recommended settings. Our greedy configuration departs from several vendor recommendations: DeepSeek recommends `temperature=0.5-0.7` for R1 models; Qwen recommends `temperature=0.6` with `top_p=0.95` for Qwen 3. We use deterministic greedy decoding throughout the main text to (a) ensure exact reproducibility without seed dependence, (b) isolate representational differences from generation-strategy effects, and (c) simplify the probing analysis, since `P0` activations are deterministic given the prompt regardless of downstream sampling. This appendix demonstrates that the qualitative findings— in particular the S1/S2 lure-rate gap and the stability of probe AUC at `P0`—are preserved under vendor-recommended sampled decoding (`temperature=0.7`).

Robustness argument. The key robustness claim is twofold: (1) probe results at P0 are *invariant* to generation strategy because P0 activations depend only on the prompt, not on downstream sampling; and (2) behavioral lure-rate gaps between S1 and S2 models are qualitatively preserved under sampled decoding, though the magnitudes shift for reasoning models whose chain-of-thought is sensitive to stochastic token selection.

F Token-Length Confound Analysis

Near-perfect probe AUC (e.g., OLMo-32B at 0.9999) can in principle arise from trivial confounds like systematic prompt-length differences between conflict and control items. Table 8 reports subword token counts per category using the Llama-3.1-8B tokenizer.

Table 8: Prompt subword length (Llama-3.1-8B tokenizer) for conflict vs. control items by category. "d" is Cohen's d; positive values indicate conflict items are longer.

Category	**Conflict**	**Control**	**Cohen's d**	p
Base rate	98.0	66.7	$+3.69$	$< 10^{-4}$
Conjunction	87.7	63.4	$+6.36$	$< 10^{-4}$
Syllogism	93.3	93.2	$+0.02$	0.953
CRT	47.0	51.8	-0.44	0.097
Arithmetic	70.9	51.0	$+6.21$	$< 10^{-4}$
Framing	97.2	106.2	-4.86	$< 10^{-4}$
Anchoring	60.4	45.5	$+4.62$	$< 10^{-4}$
Overall	**81.1**	**68.4**	$+0.45$	$< 10^{-4}$

Why length is unlikely to be the primary signal. Length differences are substantial on base rate and conjunction but balanced on syllogism and reversed on framing. Three observations argue against length as the dominant confound. First, a probe trained on syllogism alone (where length is balanced, $d = +0.02$) achieves within-category AUC = 0.936 and transfers to base rate (0.950) and conjunction (0.990)—the shared direction cannot be length. Second, length direction is inconsistent across categories, so a pure-length probe would produce anti-correlated cross-category predictions rather than the immune-cluster compression we observe. Third, the text baseline, which has direct access to length as a feature, caps at 0.840; the 16 pp gap to the activation probe (0.999) is not accounted for by these text baselines alone, though richer surface-template features remain a possible confound. For OLMo-32B specifically, probe AUC = 0.9999 on the three vulnerable categories (including length-balanced syllogism) implies the probe separates syllogism conflict vs. control despite near-identical prompt lengths (93.3 vs. 93.2 tokens).

G Final-Answer Behavioral Rescoring Details

Table 9 gives the category-level final-answer rescoring used for the Llama/R1/Qwen rows in Table 1. The rescoring recomputes verdicts from preserved 470-item raw responses after removing generated thinking text; close-only R1-style continuations containing "</think>" without generated "<think>" are scored only after the close tag. This table replaces the older legacy source-run behavioral table for these models. OLMo rows in Table 1 still use their source-run verdict artifacts because the corresponding raw-response artifacts were not cached for final-answer rescoring.

Table 9: Final-answer rescoring of preserved 470-item raw-response artifacts. For Llama/R1/Qwen, verdicts are recomputed from the final-answer span after removing generated thinking text. Acc, Lure, and Other are percentages on conflict items; Ctrl is matched-control accuracy.

Model	Category	n	Acc	Lure	Other	Ctrl
Llama	base_rate	35	11.4	88.6	0.0	100.0
Llama	conjunction	20	45.0	55.0	0.0	100.0
Llama	syllogism	25	68.0	32.0	0.0	52.0
Llama	CRT	30	56.7	10.0	33.3	83.3
Llama	arithmetic	25	100.0	0.0	0.0	100.0
Llama	framing	20	55.0	45.0	0.0	100.0

Model	Category	n	Acc	Lure	Other	Ctrl
Llama	anchoring	20	55.0	10.0	35.0	75.0
Llama	availability	15	86.7	13.3	0.0	40.0
Llama	certainty_effect	15	26.7	73.3	0.0	100.0
Llama	loss_aversion	15	100.0	0.0	0.0	100.0
Llama	sunk_cost	15	93.3	6.7	0.0	0.0
Llama	**Overall**	235	**59.6**	**33.2**	**7.2**	**80.4**
R1-Distill	base_rate	35	37.1	62.9	0.0	100.0
R1-Distill	conjunction	20	20.0	80.0	0.0	100.0
R1-Distill	syllogism	25	80.0	20.0	0.0	80.0
R1-Distill	CRT	30	66.7	16.7	16.7	100.0
R1-Distill	arithmetic	25	96.0	0.0	4.0	100.0
R1-Distill	framing	20	65.0	35.0	0.0	95.0
R1-Distill	anchoring	20	60.0	5.0	35.0	55.0
R1-Distill	availability	15	86.7	13.3	0.0	60.0
R1-Distill	certainty_effect	15	100.0	0.0	0.0	100.0
R1-Distill	loss_aversion	15	73.3	13.3	13.3	93.3
R1-Distill	sunk_cost	15	100.0	0.0	0.0	0.0
R1-Distill	**Overall**	235	**68.1**	**25.5**	**6.4**	**84.3**
Qwen think	base_rate	35	37.1	2.9	60.0	82.9
Qwen think	conjunction	20	5.0	30.0	65.0	60.0
Qwen think	syllogism	25	64.0	0.0	36.0	68.0
Qwen think	CRT	30	3.3	0.0	96.7	10.0
Qwen think	arithmetic	25	12.0	0.0	88.0	92.0
Qwen think	framing	20	0.0	5.0	95.0	55.0
Qwen think	anchoring	20	60.0	0.0	40.0	55.0
Qwen think	availability	15	53.3	26.7	20.0	33.3
Qwen think	certainty_effect	15	40.0	40.0	20.0	66.7
Qwen think	loss_aversion	15	100.0	0.0	0.0	53.3
Qwen think	sunk_cost	15	60.0	0.0	40.0	0.0
Qwen think	**Overall**	235	**35.7**	**7.7**	**56.6**	**54.9**
Qwen no-think	base_rate	35	28.6	71.4	0.0	100.0
Qwen no-think	conjunction	20	5.0	95.0	0.0	100.0
Qwen no-think	syllogism	25	48.0	0.0	52.0	52.0
Qwen no-think	CRT	30	60.0	13.3	26.7	100.0
Qwen no-think	arithmetic	25	100.0	0.0	0.0	100.0
Qwen no-think	framing	20	55.0	45.0	0.0	100.0
Qwen no-think	anchoring	20	75.0	15.0	10.0	65.0
Qwen no-think	availability	15	100.0	0.0	0.0	60.0
Qwen no-think	certainty_effect	15	13.3	86.7	0.0	100.0
Qwen no-think	loss_aversion	15	100.0	0.0	0.0	100.0
Qwen no-think	sunk_cost	15	100.0	0.0	0.0	0.0
Qwen no-think	**Overall**	235	**59.1**	**31.1**	**9.8**	**83.0**

H Steering Specificity Controls

Table 10 reports specificity controls for Llama-3.1-8B-Instruct at L14. Four sweeps: (1) probe-direction on conflict items (main result); (2) probe-direction on matched non-conflict control items (null-effect specificity); (3) mean-difference direction on conflict items (CogBias baseline); (4) mean-difference on control. cos(probe, mean-diff) $= 0.314$: the two directions differ substantially.

Table 10: Steering specificity controls on Llama-3.1-8B-Instruct, L14, continuous steering. Conflict items carry an S1 lure answer; control items carry only the normative answer.

Direction	**Items**	$\alpha=-5$	-3	-1	0	$+1$	$+3$	$+5$
Lure rate (%)								
Probe	Conflict	68.75	61.25	56.25	52.50	48.75	41.25	31.25
Probe	Control	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean-diff	Conflict	53.75	55.00	55.00	52.50	36.25	21.25	11.25
Mean-diff	Control	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Correct rate (%)								
Probe	Conflict	31.25	38.75	43.75	47.50	51.25	58.75	68.75
Probe	Control	78.75	87.50	98.75	100.00	100.00	100.00	100.00
Mean-diff	Conflict	46.25	45.00	45.00	47.50	63.75	78.75	88.75
Mean-diff	Control	60.00	73.75	97.50	100.00	100.00	100.00	100.00

Three observations. (i) The probe direction produces *zero* lure responses on control items at any α, so it does not induce the specific heuristic-lure answer on matched no-conflict prompts. The small drop in correctness at $\alpha = -5$ on control items (78.75%) reflects residual-stream coherence loss (the 21.25% of items become "other" responses, not lures), the same pattern seen with random directions in the main text. (ii) The probe-weight direction is *bidirectional*: moving toward S1 (negative α) increases lure by 16.25pp, moving toward S2 (positive α) decreases lure by 21.25pp. Bidirectional dose-response is stronger evidence that this direction controls the measured lure behavior than a unidirectional effect. (iii) The mean-difference baseline is *unidirectional*: negative α has essentially no effect (lure stays around 53–55%), while positive α strongly reduces lure (to 11.25% at $\alpha = +5$). Mean-difference steering efficiently improves behavior but cannot test causal sufficiency by worsening it—a key methodological motivation for probe-weight steering.

I Limitations and Safety Note

The main claim is restricted to the tested static continuous and limited position-targeted unit-scale probe-direction protocols. The completed high-alpha P0 diagnostic shows that much larger perturbations can change R1 behavior, so the result should not be read as non-writability in general. The R1 continuous sweep and position-targeted diagnostics still lack complete matched random-direction controls, strong cross-model perturbation calibration, sampled-decoding steering, and dynamic or SAE-feature interventions. Linear probes may miss nonlinear structure; Qwen cross-mode transfer is confounded by the hard template's positional shift; within-CoT phase labels are descriptive; and the benchmark covers English-language cognitive-bias items on models up to 32B parameters. The safety implication is cautionary: steering safeguards validated on instruction-tuned models should be revalidated before deployment on reasoning-distilled models. We release no new model weights or personal data.