LLMs Know the Constraint But Do Not Use It: Activation Bottlenecks in Pragmatic Constraint Reasoning

Anonymous ACL submission

Abstract

When a salient surface cue competes with an implicit feasibility constraint, LLMs often fail—but aggregate accuracy conflates genuine constraint inference with conservative defaulting. We formalize the distinction as *conditional constraint activation*: the constraint is internally encoded (**K**nowledge) symmetrically across constraint-present and -absent prompts (**S**ymmetry), yet only sometimes routed into the decision (**R**outing) and repairable by a donor activation (Re**p**air). A quartet diagnostic over 14 models reveals two failure modes; probes on two open weights decode the constraint above 88%, yet activation patching repairs one ($+6.4$ nats) and not the other (-0.07). On a mitigation frontier, *no* prompted intervention reaches the repair corner: all inflate conservative bias through a single mediation pathway—prerequisite mention. Hidden-constraint failure is a routing problem, not a knowledge problem.

1 Introduction

Large language models give plausible-sounding answers to questions whose implicit prerequisites they ignore. A widely-circulated example asks whether to walk or drive to a nearby car wash: most frontier models answer "walk," overlooking that a car wash needs the car to be present (Kévin (@knowmadd), 2026; Wunderlich, 2026; Jo, 2026). Behavioral evaluation across heuristic and constraint families confirms the failure is widespread: under strict aggregation, no model exceeds 75% on a 4 heuristic $\times$ 5 constraint taxonomy (Li et al., 2026). Yet aggregate accuracy on a constraint-active item rewards two distinct behaviors equally—a model that infers the hidden constraint and a model that defaults to a conservative option for unrelated reasons. The two are dissociable: when the constraint is *removed*, 12 of 14 evaluated models perform *worse* than on the original items, with drops up to 39 points (Li et al., 2026).

This paper. We argue hidden-constraint failures are best understood as a *routing* problem rather than a *knowledge* problem, and provide the diagnostic and mechanistic tools to test the claim. We formalize *conditional activation* as four falsifiable conditions on a model M and constraint C:

- **K. Knowledge:** a linear probe decodes "constraint applies" from M's hidden state above θ_K.
- **S. Symmetry:** probe accuracy is indistinguishable across constraint-active and -removed prompts.
- **R. Routing:** probe-projected magnitude predicts the gold–shortcut decision logit gap above θ_R.
- **P. Repair:** patching hidden states from an explicit-constraint prompt restores correctness without flipping constraint-removed pairs.

A model satisfying K, S, P but failing R exhibits a *conditional activation bottleneck*: the constraint is internally represented but not routed into the decision.

Two failure modes, mechanistically dissociated. A quartet-style evaluation—an ACTIVE item, a paired REMOVED control, an EXPLICIT variant, and a length- and frame-matched SALIENCE CONTROL—reveals two populations across 14 models. Over-activation models (Llama-4, Claude Opus 4.6, Kimi K2.5, Qwen3.5-27B, GPT-5.2, Claude Sonnet 4.5, Gemini 3 Pro) give the constraint-heavy answer even when the constraint is absent (CBI $+0.13$ to $+0.29$). Under-activation models (GPT-OSS-{20B, 120B}) under-apply the constraint when required (CBI -0.10 to -0.19). On the two open weights, the modes split mechanistically: in Qwen3-14B, patching constraint information from an Explicit prompt into a failing Active prompt shifts the gold–shortcut gap by $+6.4$ nats while leaving Removed pairs essentially unchanged (-0.84); in GPT-OSS-20B, an 88%-accurate probe at L_{20} confirms the constraint is internally available, yet patching moves the answer by only -0.07 nats. The under-activation

mode is the textbook routing failure; the over-activation mode is a prior-bias problem with routing intact.

Salience-controlled hint activation. A common reading of the explicitness-gradient observation (Li et al., 2026) is that hints "remind" the model of the constraint. We test this by inserting a length- and frame-matched neutral filler in place of the constraint cue. The Salience-Adjusted Hint Gain ranges from $+0.021$ (Gemini 3 Pro—no specificity above matched salience) to $+0.294$ (Llama-4—strong specificity), with matched controls flat across all five ladder levels.

The mitigation frontier. We re-evaluate four prompted interventions—CoT (Wei et al., 2022; Kojima et al., 2022), precondition listing, goal decomposition, and counterfactual checking (Wang et al., 2023c)—on a two-dimensional frontier (active-item gain vs. removed-pair harm). Across all 40 (model $\times$ strategy) cells, strategies cluster in the high-harm, near-zero-gain region: mean removed-pair harm $+0.44$ to $+0.47$ in CBI units, mean active gain only $+0.01$ to $+0.04$; *none* reaches the repair corner. They converge on a shared mediation pathway—mediated-correctness share ≥ 0.91 in every cell—each working by inducing the same surface behavior (prerequisite mention) that boosts Active-correctness while over-triggering on Removed prompts. A reasoning-budget sweep on four thinking-mode models reproduces the same harm-without-repair signature (Snell et al., 2025; Muennighoff et al., 2025).

Contributions. (i) A formal definition of *conditional constraint activation* as a falsifiable K/S/R/P claim, instantiated as a quartet diagnostic with salience controls. (ii) Empirical evidence for two distinct failure modes and a mechanistic dissociation via probes and patching on two open weights. (iii) The *mitigation frontier* protocol, showing that all four established prompted methods move along the harm axis. (iv) A measured mediation chain explaining *why* prompted compute helps when it does. Together: hidden-constraint failure is a circuit-level conditional-activation problem; prompted mitigations cannot repair it; the remedy must operate on the routing direction itself.

2 Related Work

Heuristic shortcuts and the knowing-vs-using gap. LLMs follow surface-level statistical regularities rather than perform the intended computation (Geirhos et al., 2020; McCoy et al., 2019; Nikankin et al., 2024), showing content effects (Lampinen et al., 2024) and easy distractibility (Shi et al., 2023; Mirzadeh et al., 2024). A parallel line shows models internally represent information they do not output: latent multi-hop chains (Yang et al., 2024), the reversal curse (Berglund et al., 2024), hallucinations whose truth status is encoded in hidden state (Orgad et al., 2025), and truth-direction inference-time interventions (Li et al., 2023; Zou et al., 2023). The Heuristic Override Benchmark (Li et al., 2026) extended the behavioral failure to pragmatic reasoning across four heuristic families and five constraint families, but leaves open whether a correct answer reflects genuine constraint inference or conservative defaulting. We push past that ambiguity with a quartet-based diagnostic and with mechanistic experiments that localize the failure to a specific routing pathway.

Linear probes and activation patching. Linear classifiers on intermediate representations (Alain and Bengio, 2017; Conneau et al., 2018; Hewitt and Manning, 2019; Belinkov, 2022) are the standard test for whether a feature is present in hidden state, though high probe accuracy does not by itself establish causal use (Hewitt and Liang, 2019). Many factual associations and truth-related axes are linearly decodable (Petroni et al., 2019; Geva et al., 2021; Dai et al., 2022; Gurnee and Tegmark, 2024; Burns et al., 2023; Zou et al., 2023; Orgad et al., 2025). Causal mediation (Vig et al., 2020; Geiger et al., 2021) and activation patching (Meng et al., 2022; Wang et al., 2023a; Hanna et al., 2023; Lieberum et al., 2023; Wu et al., 2023; Marks et al., 2024; Zhang and Nanda, 2024; Heimersheim and Nanda, 2024) have localized circuits for factual recall, indirect-object identification, and arithmetic (Stolfo et al., 2023). Most prior patching work targets *factual recall* or *single-step symbolic* computations. We apply patching to *conditional pragmatic reasoning* on two open weights with opposite behavioral failure modes, yielding a mechanistic dissociation aggregate behavior obscures.

Prompted mitigations and inference-time compute. A steady stream of prompting interventions has been introduced— chain-of-thought (Wei et al., 2022; Kojima et al., 2022), self-consistency (Wang et al., 2023c), tree-of-thoughts (Yao et al., 2023), plan-and-solve (Wang et al., 2023b), self-

refine (Madaan et al., 2023), verifier reranking (Cobbe et al., 2021; Lightman et al., 2024), reflexion (Shinn et al., 2023), step-back (Zheng et al., 2024)—alongside long-CoT thinking-mode models (OpenAI, 2024; DeepSeek-AI, 2025; Anthropic, 2025) and inference-scaling work (Snell et al., 2025; Brown et al., 2024; Muennighoff et al., 2025). Counter-evidence shows LLMs cannot reliably self-correct reasoning (Huang et al., 2024), CoT benefits are narrow (Sprague et al., 2025), and overthinking can degrade performance (Chen et al., 2024). Prior reports rarely audit whether constraint-removed minimal pairs are damaged by the same intervention. Our mitigation frontier re-evaluates four established strategies on a two-dimensional axis (active gain vs. pair harm), showing that gains in the literature are routinely along the conservative-bias axis rather than the repair axis.

Conservative bias and answer asymmetries. Sycophancy (Perez et al., 2023; Sharma et al., 2024; Wei et al., 2023), hedging (Zhou et al., 2024; Xiong et al., 2024), and calibration asymmetries (Kadavath et al., 2022) are documented RLHF-induced behaviors that shift outputs toward safe options. Our over-activation failure mode is their mechanistic counterpart—a constraint-heavy default that is correct when the constraint applies and wrong when it does not. Diagnosing it behaviorally requires the minimal-pair contrast; diagnosing it mechanistically requires checking whether the constraint signal in hidden state actually shifts the decision under patching.

3 Methods

We instantiate the four conditional-activation conditions (**K/S/R/P**, §1) through three measurement layers: (i) a behavioral quartet diagnostic with matched salience controls, (ii) a salience-controlled hint ladder, and (iii) hidden-state probes plus activation patching on open-weight models. We additionally re-evaluate four prompted mitigations on a two-axis frontier and probe inference-time compute through a budget sweep on four thinking-mode models.

3.1 Datasets and Conditions

Item universe. We start from the HOB benchmark (Li et al., 2026): a 4-heuristic $\times$ 5-constraint taxonomy ($4 \times 5 = 20$ cells) over $\approx$140 scenarios, each a goal + salient surface cue + hidden constraint C that, when active, makes the cue-following answer infeasible. We use CORE100: 100 base scenarios stratified across all 20 cells.

Condition	Definition
ACTIVE	Original scenario; C satisfied implicitly.
REMOVED	Constraint-removed counterfactual; cue preserved.
EXPLICIT	ACTIVE + C as one declarative sentence.
SAL. CTRL	ACTIVE + length-/frame-matched neutral filler at the same position.

Table 1: The four quartet conditions per scenario.

Quartet construction. For each scenario we construct four conditions paired by scenario id (Table 1). SALIENCE CONTROL controls for the alternative hypothesis that explicit-constraint gains are a non-specific salience effect: fillers are length-bucketed, neutral declarative sentences, assigned hash-deterministically by instance id.

Hint ladder. A five-level ladder $L_0 \prec \cdots \prec L_4$, where L_0 is ACTIVE and L_4 approaches EXPLICIT. At every level L_i ($i \geq 1$) we inject three matched negative controls: CTRL_LEX, CTRL_LEN, CTRL_HEUR.

Sample sizes. T trials per (model, condition): $T = 10$ for API models, $T = 8$ for local-GPU models, $T = 4$ for the budget sweep.

3.2 Models

We evaluate 14 LLMs spanning four providers (Claude Opus/Sonnet, GPT 5.2/5.4/OSS-20B/OSS-120B, Gemini 3 Pro, Grok 4.2, DeepSeek-R1, Kimi K2.5, Llama-4, Qwen3-14B/32B, Qwen3.5-27B). For probes and patching we use **Qwen3-14B** (balanced) and **GPT-OSS-20B** (under-activation), the two open weights spanning the failure modes.

3.3 Behavioral Metrics

Let Acc_c denote accuracy in condition c.

$$\mathrm{CBI} = \mathrm{Acc}_{\mathrm{A}} - \mathrm{Acc}_{\mathrm{R}}$$
$$\mathrm{SalAdjGain} = \mathrm{Acc}_{\mathrm{E}} - \mathrm{Acc}_{\mathrm{SC}}$$
$$\mathrm{PCA} = \frac{1}{|S|} \sum_{s} \widehat{\mathrm{Acc}}_{\mathrm{A}}(s) \cdot \widehat{\mathrm{Acc}}_{\mathrm{R}}(s)$$
$$\mathrm{CAS} = \mathrm{PCA} \cdot (1 - |\mathrm{CBI}|)$$

$\mathrm{CBI} > \theta_{\mathrm{CBI}} = 0.10$ flags over-activation; $\mathrm{CBI} < -\theta_{\mathrm{CBI}}$ flags under-activation. Over- and under-activation rates are $\mathrm{OAR} = 1 - \mathrm{Acc}_{\mathrm{R}}$ and $\mathrm{UAR} = 1 - \mathrm{Acc}_{\mathrm{A}}$. The ladder *specificity gap* at level i is

$\mathrm{Acc}_{L_i} - \max_\tau \mathrm{Acc}_{L_i^{(\tau)}}$; the minimum-effective-hint-level (MEHL) is the smallest i at which it exceeds a preregistered 0.10 margin.

3.4 Mitigation Frontier

We evaluate four prompted strategies grouped by hypothesized mechanism: Class A raises the conservative threshold (vanilla CoT (Wei et al., 2022; Kojima et al., 2022)); Class B induces prerequisite search (*precondition listing*, *goal decomposition*); Counterfactual-check (Wang et al., 2023c) is an oracle upper bound. Each is a single-prompt prefix; every other condition is held fixed. For (model, strategy) we report the two-axis frontier ACTIVEGAIN $= \mathrm{Acc}_{\mathrm{A}}^{\mathrm{strat}} - \mathrm{Acc}_{\mathrm{A}}^{\mathrm{zero}}$ vs. PAIRHARM $= \mathrm{CBI}^{\mathrm{strat}} - \mathrm{CBI}^{\mathrm{zero}}$; the repair corner (positive ACTIVEGAIN, non-positive PAIRHARM) sits at the bottom-right.

3.5 Reasoning-Budget Sweep

We sweep budget $\in \{256, 1024, 4096, 16384\}$ thinking tokens on the four thinking-mode models. Provider-specific budget routing (`thinking.budget_tokens`, `reasoning_effort`, `ThinkingConfig`) is detailed in Appendix G.

3.6 Mediation Analysis

We estimate the chain *intervention* $\rightarrow$ *prerequisite mention* $\rightarrow$ *correctness*. Mention is detected by content-word overlap with the canonical hidden constraint (Appendix G). We report $\Pr(\mathrm{c} \mid \mathrm{m})$, $\Pr(\mathrm{c} \mid \neg\mathrm{m})$, $\Pr(\mathrm{m})$, and the *mediated correctness share*, the fraction of correct answers attributable to mentioning traces.

3.7 Hidden-State Probes (K, S, R)

For Qwen3-14B and GPT-OSS-20B we cache the post-block residual hidden state at the final input token across all transformer layers on every CORE100 quartet item. We train one per-layer logistic regression (`liblinear`, $C = 1$, with per-feature standardization) to predict ACTIVE vs. REMOVED. Cross-validation is *scenario-grouped* 5-fold: the four conditions for a scenario co-occur in the same fold. **K** is satisfied at the best layer when CV accuracy exceeds $\theta_K = 0.80$. **S** (symmetry) is the test-accuracy gap between balanced Active-only and Removed-only training subsets, threshold $\theta_S = 0.05$. **R** (routing) is the Spearman correlation between probe projection and gold–shortcut decision logit gap at the best probe layer, threshold $\theta_R = 0.30$.

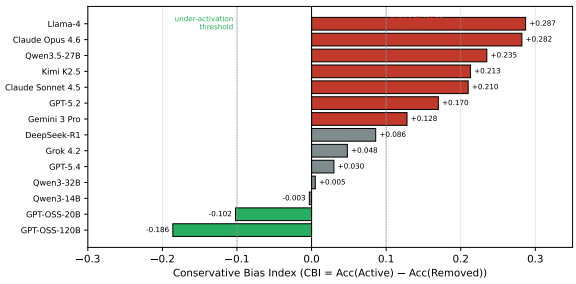

Figure 1: Per-model Conservative Bias Index, rank-ordered. Two populations are separated by $|\mathrm{CBI}| = 0.10$ (grey dashed): over-activation (red, 7 models), balanced (grey, 5 models), and under-activation (green, 2 models).

3.8 Activation Patching (P)

For each scenario we forward the EXPLICIT (donor) prompt and cache the final-token hidden state $h_{\ell^\star}^{\mathrm{donor}}$ at the probe's best layer $\ell^\star$. We then forward the recipient (ACTIVE or REMOVED) prompt with a forward hook on $\ell^\star$ that overwrites the final-token output with $h_{\ell^\star}^{\mathrm{donor}}$, and recompute the gold–shortcut log- probability gap. The patched-minus-baseline difference Δ measures how much the constraint-encoding activation changes the decision. **P** requires specificity: $\Delta^{\mathrm{Active}} > \theta_P = 1.0$ nat *and* $|\Delta^{\mathrm{Removed}}| \leq \theta_P$, preregistered against the typical logit-gap magnitude ($\sigma_\Delta \approx 3$ nats). Implementation, donor-shuffling control, and layer/position sweeps are in Appendix G.

4 Results

We present results in the order of the four conditions: **K** (knowledge present), **S** (symmetric across paired conditions), **R** (routed into the decision)—all in §4.5— and **P** (repair via patching, §4.6). §4.1 reports the behavioral signature on all 14 models; §4.2 reports the salience-controlled hint ladder; §4.3 reports the mitigation frontier; §4.4 reports the reasoning-budget sweep with mediation analysis.

4.1 Behavioral signature: two failure modes

Table 2 reports per-model accuracy in all four quartet conditions, plus the derived metrics. Across 14 models we observe *two* populations separated by the sign of CBI (Fig. 1).

Over-activation (CBI $> +0.10$): 7 models. Llama-4, Claude Opus 4.6, Qwen3.5-27B, Kimi

K2.5, Claude Sonnet 4.5, GPT-5.2, and Gemini 3 Pro (CBI $+0.13$ to $+0.29$; see Table 2). These models pick the constraint-heavy answer even when the constraint is absent: Removed accuracy falls 11–33 points below Active. Pair-Consistent Accuracy correspondingly collapses (e.g. 0.342 for Llama-4 vs. 0.792 marginal Active accuracy).

Balanced (|CBI| ≤ 0.10): 5 models. DeepSeek-R1 ($+0.086$), Grok 4.2 ($+0.048$), GPT-5.4 ($+0.030$), Qwen3-32B ($+0.005$), Qwen3-14B (-0.003). Active and Removed accuracies are within 9 points across all five.

Under-activation (CBI < -0.10): 2 models. GPT-OSS-20B (-0.103) and GPT-OSS-120B (-0.186). These models are *more* accurate on Removed than on Active—they fail to apply the constraint when it matters.

Salience-Adjusted Hint Gain. SalAdjGain ranges from $+0.021$ (Gemini 3 Pro: no specificity beyond matched salience) to $+0.294$ (Llama-4) and $+0.265$ (Qwen3-32B). The under-activation models retain large positive SalAdjGain (GPT-OSS-20B: $+0.248$; GPT-OSS-120B: $+0.230$), confirming the constraint *can* be activated when made explicit. Gemini 3 Pro's near-zero SalAdjGain demonstrates that a model can show large *nominal* hint gains entirely explained by added textual salience.

4.2 Hint ladder with matched controls

On the six API models with full ladder collection, explicit-condition accuracy rises monotonically with hint level for every model; the three matched controls remain within ± 0.04 of L_0 across all four ladder rungs. The specificity gap exceeds the preregistered 0.10 margin by L_2 for over-activation models and by L_3 for balanced models (DeepSeek-R1, GPT-5.4). Per-model curves and specificity gaps are in Figs. 5 and 6 (Appendix).

4.3 The mitigation frontier

Table 3 reports per-strategy means over 10 API models; full per-(model, strategy) detail is in Table 7 (Appendix).

Existing strategies inflate CBI without repairing routing. Across all 40 (model, strategy) cells, PAIRHARM is large ($+0.28$ to $+0.65$) and ACTIVEGAIN is small (-0.08 to $+0.11$); none reaches the repair corner. Counter to our preregistered hypothesis, Class-B interventions (*precondition_listing*, *goal_decomposition*) do not move closer to repair than Class-A: mean PAIRHARM $+0.44$ vs. $+0.47$, mean ACTIVEGAIN $+0.038$ vs. $+0.012$. The *counterfactual_check* oracle shows the same pair-harm signature ($+0.44$ mean). Llama-4 is the worst case: *every* strategy reduces its Active accuracy (-0.018 to -0.082).

Mediation through prerequisite mention. Fig. 2 visualizes the per-strategy frontier; every point sits in the high-harm, low-gain region.

Across all 40 cells, the conditional gap $\Pr(\mathrm{c} \mid \mathrm{m}) - \Pr(\mathrm{c} \mid \neg\mathrm{m})$ is large ($+0.34$ to $+0.64$), $\Pr(\mathrm{m})$ is high (0.84 to 0.96), and the mediated-correctness share is 0.91 to 0.99. Every strategy "works" by inducing the same surface behavior (prerequisite mention); they differ only in how much they over-trigger it on Removed items.

4.4 Reasoning-budget sweep

We sweep budget $\in \{256, 1024, 4096, 16{,}384\}$ thinking tokens on the four thinking-mode models (Claude Opus 4.6, GPT-5.4, Gemini 3 Pro, DeepSeek-R1) over the full CORE100 quartet ($T = 4$ trials per cell). Per-budget Active/Removed accuracy, CBI, and mediation metrics are summarized in Fig. 3.

Thinking-mode *itself* moves all four models well into the over-activation regime: at budget=256 tokens, CBI is $+0.85$ for Claude Opus (vs. $+0.28$ zero-shot), $+0.57$ for GPT-5.4 (vs. $+0.03$), $+0.65$ for Gemini 3 Pro (vs. $+0.13$), $+0.66$ for DeepSeek-R1 (vs. $+0.09$). Removed accuracy collapses (≤ 0.22) while Active accuracy holds within 0.05 of zero-shot. Budget within thinking-mode is essentially a no-op: CBI varies by ≤ 0.05 across the four-point grid. Mediation is uniformly strong: $\Pr(\mathrm{m}) \in [0.82, 0.92]$, conditional gap $\Pr(\mathrm{c} \mid \mathrm{m}) - \Pr(\mathrm{c} \mid \neg\mathrm{m}) \in [+0.43, +0.67]$, mediated-correctness share ≥ 0.93 throughout. Extra compute does not change *which* items get the mention, only the strength of the bias. Per-budget detail is in Table 8 (Appendix).

4.5 Conditions K, S, R: open-weight probes

(K) Knowledge. A linear probe on per-layer hidden states decodes the binary Active-vs-Removed label well above chance for both models (Table 4). Qwen3-14B peaks at layer 27 with $94.5\% \pm 1.9\%$ cross-validated accuracy; GPT-OSS-20B peaks at layer 20 with $88.0\% \pm 4.3\%$. The constraint is internally encoded in both models, well above the preregistered $\theta_K = 0.80$ threshold. We additionally

Model	Per-condition Acc				CBI	PCA	CAS	SalAdjGain
	Active	Removed	Explicit	Sal Ctrl				
Over-activation								
Llama-4	0.792	0.505	0.926	0.632	+0.287	0.342	0.400	+0.294
Claude Opus 4.6	0.886	0.604	0.923	0.833	+0.282	0.523	0.535	+0.090
Qwen3.5-27B	0.925	0.690	0.969	0.851	+0.235	0.645	0.638	+0.118
Kimi K2.5	0.928	0.715	0.956	0.853	+0.213	0.650	0.664	+0.103
Claude Sonnet 4.5	0.837	0.627	0.936	0.766	+0.210	0.496	0.525	+0.170
GPT-5.2	0.824	0.654	0.929	0.802	+0.170	0.513	0.539	+0.127
Gemini 3 Pro	0.907	0.779	0.910	0.889	+0.128	0.693	0.707	+0.021
Balanced								
DeepSeek-R1	0.850	0.764	0.957	0.795	+0.086	0.627	0.649	+0.162
Grok 4.2	0.883	0.835	0.954	0.790	+0.048	0.725	0.737	+0.164
GPT-5.4	0.822	0.792	0.938	0.723	+0.030	0.634	0.651	+0.215
Qwen3-32B	0.771	0.766	0.932	0.667	+0.005	0.580	0.591	+0.265
Qwen3-14B	0.762	0.765	0.944	0.694	−0.003	0.567	0.583	+0.250
Under-activation								
GPT-OSS-20B	0.765	0.868	0.941	0.694	−0.103	0.657	0.664	+0.248
GPT-OSS-120B	0.686	0.872	0.857	0.627	−0.186	0.595	0.598	+0.230

Table 2: Quartet behavioral metrics on CORE100, 14 models, sorted by CBI. PCA is the soft (per-scenario expectation) variant; CAS combines PCA with the conservative-bias penalty. SalAdjGain is the explicit-minus-salience-control accuracy gap.

Strategy	ActGain	PairHarm	MedShare
cot (A)	+0.012	+0.475	0.962
precond. listing (B)	+0.043	+0.423	0.974
goal decomp. (B)	+0.034	+0.458	0.975
counterfactual (oracle)	+0.020	+0.436	0.954

Table 3: Per-strategy means over 10 API models. Every strategy moves overwhelmingly along the harm axis; ActiveGain is small and indistinguishable across classes. Per-(model, strategy) detail is in Table 7 (Appendix).

	Qwen3-14B	GPT-OSS-20B
Probe best layer	L_{27}	L_{20}
CV accuracy (K)	0.945 ± 0.019	0.880 ± 0.043
Sym. gap (S)	0.011	0.024
Routing ρ @ best L (R)	0.125	0.224
Max ρ over layers (R)	0.242	0.302

Table 4: Conditions K, S, R on the two open-weight models. Both satisfy K and S; both fall below $\theta_R = 0.30$ at the probe's best layer (Qwen3-14B clearly; GPT-OSS-20B marginally).

observe broad-layer accessibility: ≥ 0.92 accuracy across layers 25–36 in Qwen3-14B and ≥ 0.85 across layers 4–24 in GPT-OSS-20B (full per-layer curves in Table 9, Appendix).

(S) Symmetry. Re-training the probe on a balanced constraint-active/constraint-removed split shows the test-accuracy gap between the two halves is within $\theta_S = 0.05$ for both models at their best layer ($\Delta = 0.011$ for Qwen3-14B; $\Delta = 0.024$ for GPT-OSS-20B). The knowledge is present symmetrically; an asymmetry-based explanation of the behavioral CBI is ruled out.

(R) Routing. The probe-projected magnitude at the best layer correlates only weakly with the per-trial gold–shortcut logit gap (Spearman $\rho = 0.125$ for Qwen3-14B, 0.224 for GPT-OSS-20B); the per-layer maxima (0.242, 0.302) also fall below the preregistered $\theta_R = 0.30$. The probe direction is present but not strongly read by the decision head—the signature of a routing failure.

4.6 Condition P: activation patching

Patching the constraint-encoding hidden state from the Explicit prompt into the Active prompt at the probe's best layer produces opposite signatures in the two models (Table 5, $n = 100$ scenarios each).

Qwen3-14B. The Active gold–shortcut gap rises from +1.40 to +7.78 nats ($\Delta = +6.38$), an order-of-magnitude shift, while Removed pairs are only mildly affected ($\Delta = -0.84$). This is a clean repair signature: the constraint-encoding direction *exists* in the donor and *causally* shifts the recipient's decision when patched. **P is satisfied**. **GPT-OSS-20B.** The same intervention yields no effect: $\Delta = -0.069$ on Active patches, +0.029 on Removed (both within the $\sim$ 3-nat noise floor). Despite the 88%-accurate probe at L_{20}, the constraint direction does not causally control the decision. **P fails**.

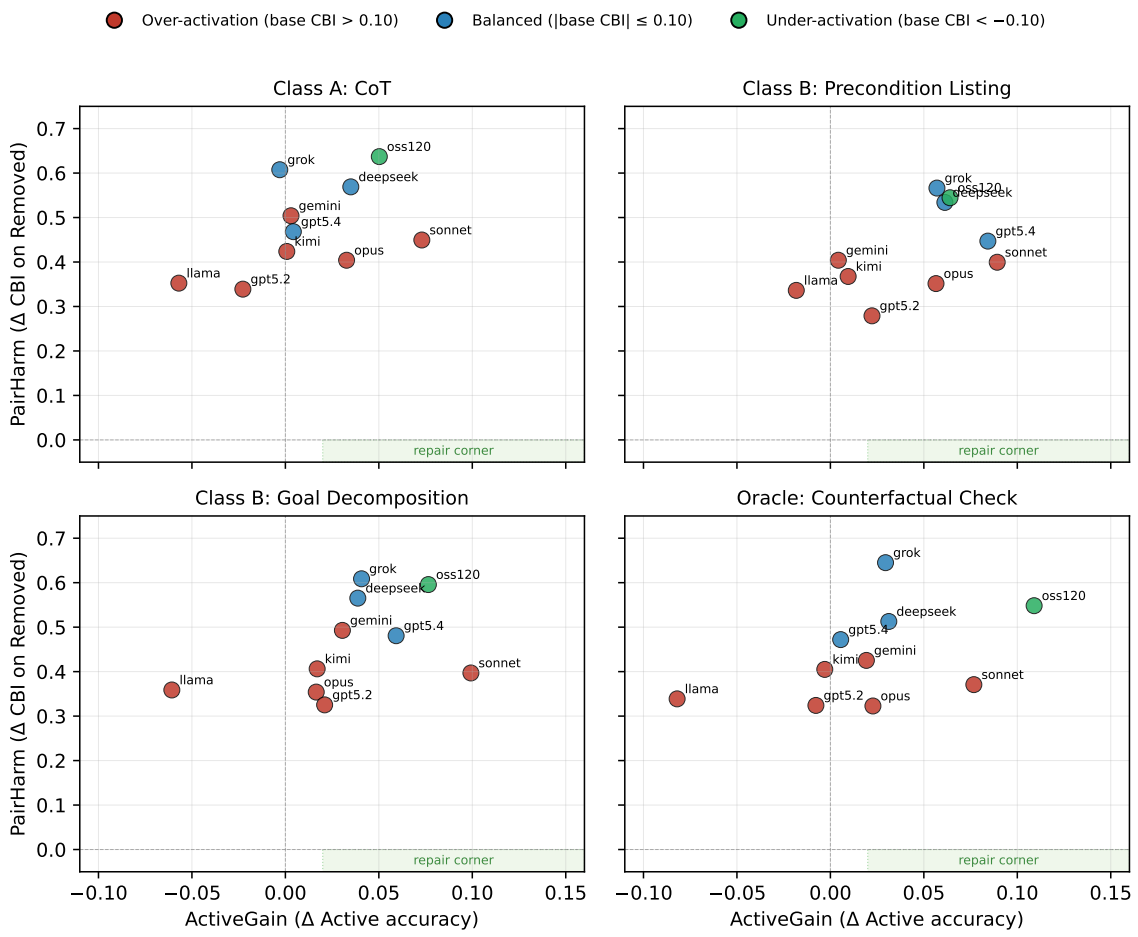

Figure 2: Mitigation frontier across ten API models and four prompted strategies. Each point is one (model, strategy) cell at the model's own zero-shot baseline. The green "repair corner" (bottom-right of each panel: positive ActiveGain, non-positive PairHarm) is empty for every strategy. Points colored by failure mode of the base model.

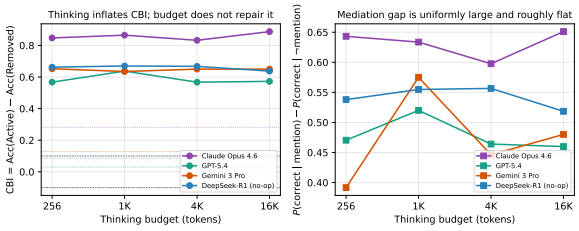

Figure 3: Reasoning-budget sweep on four thinking-mode models. *Left:* CBI as a function of budget; the four dotted horizontals show each model's zero-shot (non-thinking) baseline. Thinking-mode moves all four models well above the zero-shot CBI; budget within thinking-mode does not. *Right:* the mediation conditional gap $\Pr(c \mid m) - \Pr(c \mid \neg m)$ stays uniformly large and budget-insensitive.

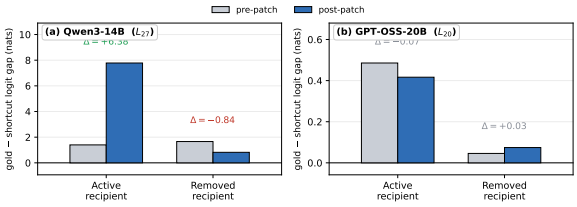

Figure 4: Activation-patching deltas at the probe's best layer. Pre-patch (grey) vs. post-patch (blue) mean gold–shortcut logit gap. Qwen3-14B shows a clean repair signature on Active recipients ($\Delta = +6.38$) with controlled Removed deviation; GPT-OSS-20B shows no effect, despite the strong probe at the same layer.

K/S/R/P verdict. Qwen3-14B is K✓ S✓ R∂P✓ (causally available, weakly routed); GPT-OSS-20B is K✓ S✓ R∂P× (textbook routing failure). The 7 over-activation models share this K/S/R structure but add a strong constraint-respecting prior.

5 Discussion

5.1 Routing, not knowledge

The K/S/R/P measurements disconfirm the simplest account of hidden-constraint failure—that the model does not know the constraint. Probes decode

	pre-gap	patched	Δ (nats)
Qwen3-14B, L_{27} Explicit→Recip.			
Active recip.	$+1.398$	$+7.781$	$\mathbf{+6.382}$
Removed recip.	$+1.660$	$+0.820$	-0.840
GPT-OSS-20B, L_{20} Explicit→Recip.			
Active recip.	$+0.486$	$+0.417$	$\mathbf{-0.069}$
Removed recip.	$+0.046$	$+0.074$	$+0.029$

Table 5: Activation-patching repair test. Qwen3-14B exhibits a clean repair signature on Active patches with controlled Removed harm. GPT-OSS-20B exhibits no effect, despite a strong probe at the same layer (K satisfied).

constraint applicability at 94.5% (Qwen3-14B, L_{27}) and 88.0% (GPT-OSS-20B, L_{20}), broadly across 10+ layers, and the symmetry test rules out a constraint-active-only artifact. Knowledge *is* there.

What separates the two open weights is the *causal* test: patching the donor Explicit activation repairs Qwen3-14B ($+6.4$ nats, Removed unchanged) but does nothing in GPT-OSS-20B (-0.07)—the direction is decodable yet unread by the decision head. The implication: Active-item accuracy alone cannot tell competent inference from a constraint-heavy prior; the Removed pair makes the distinction visible.

5.2 Why every prompted mitigation fails the same way

The mitigation-frontier evidence is the most surprising part of this paper. CoT, precondition listing, goal decomposition, and even counterfactual checking—an instruction to explicitly imagine the salient resource absent—all produce the same harm profile: large PairHarm ($+0.28$ to $+0.65$), small ActiveGain (-0.08 to $+0.11$), $\geq 91\%$ mediated-correctness share. The strategies are not differentiated by how they reason; they are differentiated only by how often they trigger the *same downstream behavior*—namely, mentioning the hidden constraint.

A single common cause fits. Each strategy lifts $\Pr(\text{mention})$ from a low baseline (0.3–0.5) to a uniformly high value (0.84–0.96); whether or not the constraint applies, the model talks about it, and the conservative answer then dominates. The conditional gap is large ($+0.34$ to $+0.64$) but *symmetric*: on Removed items, mention flips the answer the wrong way, netting to near-zero on Pair-Consistent Accuracy—yet the literature reports only the constraint-active half. Prompted reasoning cannot *decide* when to mention the constraint, only globally bias the rate; selective activation is a routing problem, not a prompt-level property.

5.3 Two failure modes, two remedies

The mechanistic measurements suggest different remedies per population. **Over-activation** (7 models) combines intact routing with a prior bias (large SalAdjGain); prompted mitigations push exactly the coefficient that already over-fires. **Balanced** (5 models) is where Qwen3-14B's patching success shows a *selective* intervention—inject the constraint direction only when it applies— is mechanistically feasible. **Under-activation** (2 models) is the hard case: GPT-OSS-20B's direction exists at L_{20} but is unread by the decision head, and the single-layer patch is insufficient. A *learned activation gate* that decides per item when to patch would test whether circuit-level intervention can reach the repair corner; the released probe weights, cross-architecture replication, and a multi-label extension to multi-constraint scenarios are the natural next steps.

6 Conclusion

We recast hidden-constraint failure in LLMs as *conditional activation*: the constraint is decodable from hidden state (K) symmetrically (S) but not reliably routed into the decision (R). Patching dissociates the two failure modes, and a mitigation frontier shows every prompted intervention inflates conservative bias rather than repairing routing—so the fix must target the routing direction itself.

Limitations

Scenario count. CORE100 stratifies across the HOB cell taxonomy at the cost of statistical power per cell. Our between-condition contrasts use the matched paired design to recover power, but per-cell error bars remain wide for finer-grained slicing.

Two open weights for the causal claim. The K/S/R/P mechanism is established on Qwen3-14B and GPT-OSS-20B only. While the behavioral signature extends to all 14 models, the causal patching evidence is backed by two open-weight models, chosen for release-license reasons and to span the two failure modes.

Single-layer patching. Our Repair condition is satisfied by patching at the probe's single best layer. A multi-layer or block-wise intervention may reveal repair in models where the single-layer patch

fails (e.g. GPT-OSS-20B), and our analysis therefore underestimates the rate at which Repair holds.

English, single-domain. HOB scenarios are English everyday-domain. Cross-lingual replication and domain-shifted versions (e.g. medical, legal) are open questions; we expect the K/S/R/P formalism to transfer but the failure-mode partition may not.

Prompted-mitigation coverage. We evaluate four prompted strategies that span the published taxonomy of pragmatic-reasoning interventions. Reinforcement-learned mitigations (Lambert, 2024; Gulcehre et al., 2023) and fine-tuned debiasing remain untested. Our prediction, based on the mediation chain, is that any intervention that operates on the trace surface will fall in the same harm-without-repair region; we leave that to empirical verification.

Judge model. All 54,400 quartet trials and 128,000 mitigation trials are judged by a single Qwen3-32B local model with deterministic decoding. This is a standard choice in the heuristic-shortcut literature, but a single-judge evaluation may systematically miss certain failure modes (e.g. correct answers given for the wrong reason). We release the judging prompt and all raw generations so reanalysis with alternative judges is straightforward.

API model versioning. Frontier models on closed APIs (Claude, GPT, Gemini, Grok, Kimi, DeepSeek) may be updated between data collection (May 2026) and publication. We pin exact model identifiers in Table 6 (Appendix A) and release the full raw generations, but pin-violation risk on closed APIs is non-zero.

Heuristic prerequisite-mention detection. Our mediation analysis uses content-word overlap with the canonical hidden constraint to detect "the model mentions the prerequisite." This catches cases where the model paraphrases the constraint but may miss implicit references; we report mediation conservatively.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (Workshop)*.

Anthropic. 2025. Claude's extended thinking. https://www.anthropic.com/news/visible-extended-thinking.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations*.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2024. Do NOT think that much for 2+3=? on the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *ACL*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *ACL*.

DeepSeek-AI. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *EMNLP*.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Randal Douc, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *International Conference on Learning Representations*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *EMNLP-IJCNLP*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *NAACL-HLT*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*.

Heejin Jo. 2026. Prompt architecture determines reasoning quality: A variable isolation study on the car wash problem. *arXiv preprint arXiv:2602.21814*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Samuel R. Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kévin (@knowmadd). 2026. Car wash reasoning test. Mastodon post.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Nathan Lambert. 2024. Reinforcement learning from human feedback: a tutorial. *arXiv preprint arXiv:2404.12358*.

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C.Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36.

Yubo Li, Lu Zhang, Tianchong Jiang, Ramayya Krishnan, and Rema Padman. 2026. The model says walk: How surface heuristics override implicit constraints in LLM reasoning. *arXiv preprint arXiv:2603.29025*.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in Chinchilla. *arXiv preprint arXiv:2307.09458*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time

scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20275–20321.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2024. Arithmetic without algorithms: Language models solve math with a bag of heuristics. *arXiv preprint arXiv:2410.21272*.

OpenAI. 2024. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations*.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of ACL*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *EMNLP-IJCNLP*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *EMNLP*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023a. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *ACL*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. 2023. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems*.

Felix Wunderlich. 2026. Car wash test on 53 leading AI models: "i want to wash my car. the car wash is 50 meters away. should i walk or drive?". Opper AI blog.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *ACL*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Take a step back: Evoking reasoning via abstraction in large language models. In *International Conference on Learning Representations*.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *ACL*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Model versions

Table 6 pins the exact model identifiers and access modality for all 14 evaluated models. All API data were collected in May 2026; closed-API models may be updated by their providers after this date. Open-weight models are served locally from the listed Hugging Face checkpoints and are version-stable.

Model	Provider	Identifier
GPT-5.2	OpenAI	`gpt-5.2`
GPT-5.4	OpenAI	`gpt-5.4`
Claude Opus 4.6	Anthropic	`claude-opus-4-6`
Claude Sonnet 4.5	Anthropic	`claude-sonnet-4-5-20250929`
DeepSeek-R1	DeepSeek	`deepseek-reasoner`
Gemini 3 Pro	Google	`gemini-3.1-pro-preview`
Grok 4.2	xAI	`grok-4.20-beta-0309-reasoning`
Kimi K2.5	Moonshot	`kimi-k2.5`
Llama-4 Scout	Groq	`meta-llama/llama-4-scout-17b-16e-instruct`
GPT-OSS-120B	Groq	`openai/gpt-oss-120b`
Qwen3-14B	local	`Qwen/Qwen3-14B`
Qwen3-32B	local	`Qwen/Qwen3-32B`
Qwen3.5-27B	local	`Qwen/Qwen3.5-27B`
GPT-OSS-20B	local	`openai/gpt-oss-20b`

Table 6: Exact model identifiers and access modality (API vs. local Hugging Face checkpoint). API data collected May 2026. The judge model is Qwen3-32B for all trials. The four local checkpoints (bottom block) are also the open weights used for hidden-state probing and activation patching (Qwen3-14B, GPT-OSS-20B).

B Per-(model, strategy) mitigation frontier

Table 7 reports the full $10 \times 4 = 40$ mitigation cells (10 API models, 4 prompted strategies) referenced in §4.3. All values are relative to each model's zero-shot baseline on the same 100 CORE100 scenarios.

Strategy	Model	ActGain	PairHarm	ΔCBI	Pr(c | m)	Pr(c | ¬m)	Pr(m)
cot							
cot	claude-opus-4.6	+0.033	+0.404	+0.437	0.759	0.182	0.928
cot	claude-sonnet-4.5	+0.073	+0.450	+0.523	0.764	0.124	0.935
cot	deepseek-r1	+0.035	+0.569	+0.604	0.786	0.270	0.848
cot	gemini-3-pro	+0.003	+0.504	+0.507	0.766	0.312	0.957
cot	gpt-5.2	−0.023	+0.339	+0.316	0.751	0.409	0.852
cot	gpt-5.4	+0.004	+0.468	+0.473	0.767	0.348	0.849
cot	gpt-oss-120b	+0.050	+0.637	+0.687	0.683	0.079	0.924
cot	grok-4.2	−0.003	+0.607	+0.604	0.792	0.365	0.839
cot	kimi-k2.5	+0.001	+0.424	+0.424	0.792	0.157	0.942
cot	llama-4	−0.057	+0.353	+0.295	0.642	0.124	0.927
precondition_listing							
prec	claude-opus-4.6	+0.056	+0.351	+0.408	0.788	0.241	0.946
prec	claude-sonnet-4.5	+0.089	+0.399	+0.489	0.775	0.292	0.944
prec	deepseek-r1	+0.061	+0.534	+0.595	0.775	0.259	0.934
prec	gemini-3-pro	+0.004	+0.404	+0.408	0.799	0.447	0.959
prec	gpt-5.2	+0.022	+0.279	+0.301	0.771	0.424	0.917
prec	gpt-5.4	+0.084	+0.447	+0.531	0.798	0.336	0.911
prec	gpt-oss-120b	+0.064	+0.544	+0.608	0.706	0.191	0.944
prec	grok-4.2	+0.057	+0.566	+0.623	0.799	0.306	0.927
prec	kimi-k2.5	+0.009	+0.367	+0.377	0.814	0.333	0.954
prec	llama-4	−0.018	+0.336	+0.318	0.691	0.178	0.900
goal_decomposition							
goal	claude-opus-4.6	+0.016	+0.354	+0.370	0.761	0.261	0.941
goal	claude-sonnet-4.5	+0.099	+0.397	+0.496	0.784	0.294	0.942
goal	deepseek-r1	+0.039	+0.565	+0.604	0.757	0.264	0.912
goal	gemini-3-pro	+0.030	+0.493	+0.523	0.784	0.353	0.964
goal	gpt-5.2	+0.021	+0.325	+0.346	0.760	0.386	0.903
goal	gpt-5.4	+0.059	+0.481	+0.540	0.779	0.294	0.913
goal	gpt-oss-120b	+0.076	+0.596	+0.672	0.688	0.120	0.951
goal	grok-4.2	+0.041	+0.609	+0.649	0.794	0.260	0.913
goal	kimi-k2.5	+0.017	+0.406	+0.423	0.809	0.248	0.948
goal	llama-4	−0.061	+0.359	+0.298	0.644	0.123	0.927
counterfactual_check (oracle)							
cf	claude-opus-4.6	+0.023	+0.323	+0.345	0.787	0.300	0.916
cf	claude-sonnet-4.5	+0.077	+0.371	+0.448	0.797	0.205	0.908
cf	deepseek-r1	+0.031	+0.513	+0.544	0.801	0.296	0.854
cf	gemini-3-pro	+0.019	+0.425	+0.445	0.810	0.345	0.936
cf	gpt-5.2	−0.008	+0.324	+0.316	0.777	0.376	0.865
cf	gpt-5.4	+0.006	+0.472	+0.478	0.772	0.380	0.845
cf	gpt-oss-120b	+0.109	+0.548	+0.657	0.730	0.177	0.940
cf	grok-4.2	+0.029	+0.645	+0.674	0.796	0.286	0.863
cf	kimi-k2.5	−0.003	+0.405	+0.402	0.815	0.302	0.912
cf	llama-4	−0.082	+0.339	+0.257	0.664	0.212	0.845

Table 7: Per-(strategy, model) mitigation frontier with mediation detail. ActGain is the change in Active accuracy relative to zero-shot; PairHarm is the change in Removed accuracy expressed as CBI inflation; ΔCBI is ActGain + PairHarm; Pr(c | m) and Pr(c | ¬m) are correctness conditional on prerequisite mention; Pr(m) is the marginal mention rate.

C Per-budget breakdown

Table 8 reports per-(model, budget) Active accuracy, Removed accuracy, CBI, mention rate, and mediated correctness share for the four thinking-mode models referenced in §4.4.

Model	B	A	R	CBI	Pr(m)	MedSh
claude-o	256	0.89	0.04	+0.85	0.87	0.99
claude-o	1K	0.87	0.01	+0.87	0.87	0.99
claude-o	4K	0.87	0.03	+0.83	0.87	0.97
claude-o	16K	0.87	-0.02	+0.89	0.88	0.99
deepseek	256	0.83	0.17	+0.66	0.85	0.94
deepseek	1K	0.86	0.19	+0.67	0.85	0.94
deepseek	4K	0.84	0.17	+0.67	0.87	0.95
deepseek	16K	0.83	0.20	+0.64	0.87	0.93
gemini-3	256	0.87	0.22	+0.65	0.90	0.95
gemini-3	1K	0.83	0.20	+0.64	0.92	0.97
gemini-3	4K	0.84	0.19	+0.65	0.91	0.95
gemini-3	16K	0.87	0.21	+0.65	0.92	0.96
gpt-5.4	256	0.78	0.21	+0.57	0.84	0.93
gpt-5.4	1K	0.81	0.17	+0.64	0.83	0.95
gpt-5.4	4K	0.79	0.22	+0.57	0.83	0.93
gpt-5.4	16K	0.79	0.22	+0.57	0.82	0.93

Table 8: Per-budget breakdown of the reasoning-budget sweep. B is in tokens; CBI is signed ($> +0.10$ = over-activation).

D Per-layer probe accuracy

Table 9 reports cross-validated probe accuracy at every transformer layer for both open-weight models. Bold marks the best layer used for downstream routing correlation (§4.5) and activation patching (§4.6).

L	Qwen 14B	GPT-OSS 20B	L	Qwen 14B	GPT-OSS 20B
0	0.595	0.560	21	0.930	**0.875**
1	0.715	0.730	22	0.930	0.870
2	0.770	0.820	23	0.935	0.870
3	0.800	0.835	24	0.940	0.860
4	0.815	0.875	25	0.940	–
5	0.820	0.860	26	0.940	–
6	0.830	0.860	27	**0.945**	–
7	0.835	0.880	28	0.940	–
8	0.835	0.860	29	0.935	–
9	0.840	0.860	30	0.935	–
10	0.840	0.860	31	0.930	–
11	0.850	0.870	32	0.930	–
12	0.855	0.870	33	0.935	–
13	0.870	0.875	34	0.930	–
14	0.870	0.875	35	0.930	–
15	0.880	0.870	36	0.930	–
16	0.900	0.870	37	0.920	–
17	0.910	0.860	38	0.920	–
18	0.920	0.860	39	0.905	–
19	0.925	0.870	40	0.905	–
20	0.930	**0.880**			

Table 9: Per-layer cross-validated probe accuracy on the Active vs. Removed binary label, scenario-grouped 5-fold. Standard deviations range from 0.018 to 0.052 (omitted for compactness). Bold = layer chosen for downstream patching. Qwen3-14B has 41 layers, GPT-OSS-20B has 25; the right column for layers ≥ 25 is empty for GPT-OSS-20B.

E Hint-ladder details

Fig. 5 reports the per-model hint-ladder activation curve with matched negative controls; Fig. 6 shows the per-level specificity gap. The activation-specific component of the hint gain grows monotonically with hint level on every ladder-complete model.

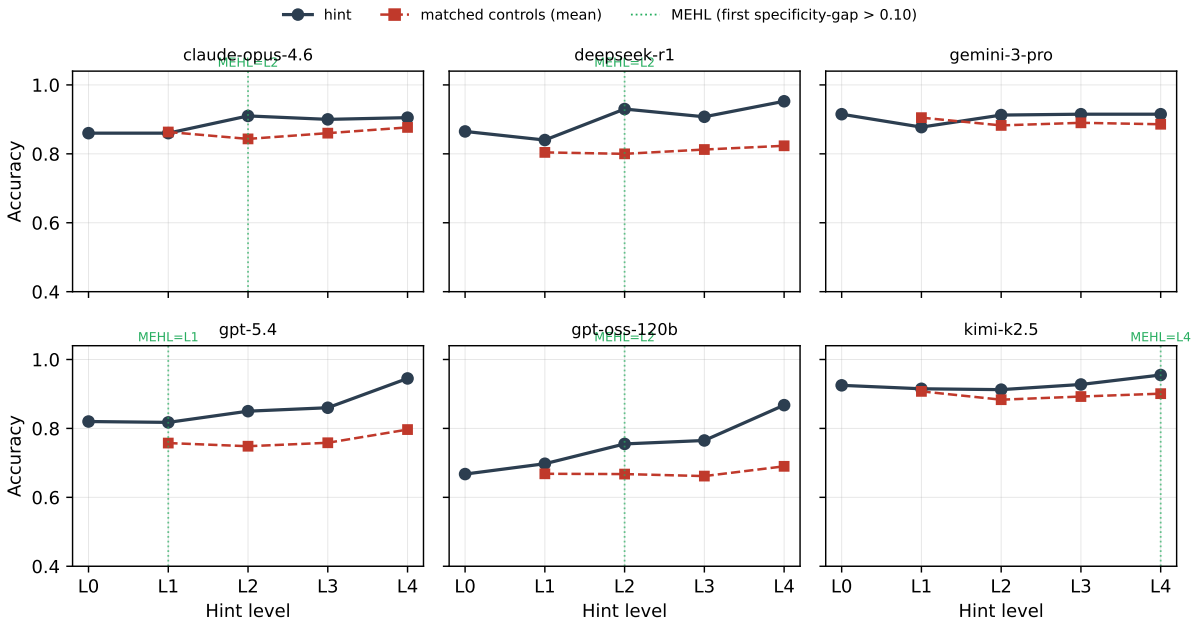

Figure 5: Hint-ladder activation curves with matched negative controls on six API models. Hint accuracy (solid black) climbs monotonically; the mean of the three matched controls (dashed red) stays flat, demonstrating the gain is hint-specific rather than salience-driven.

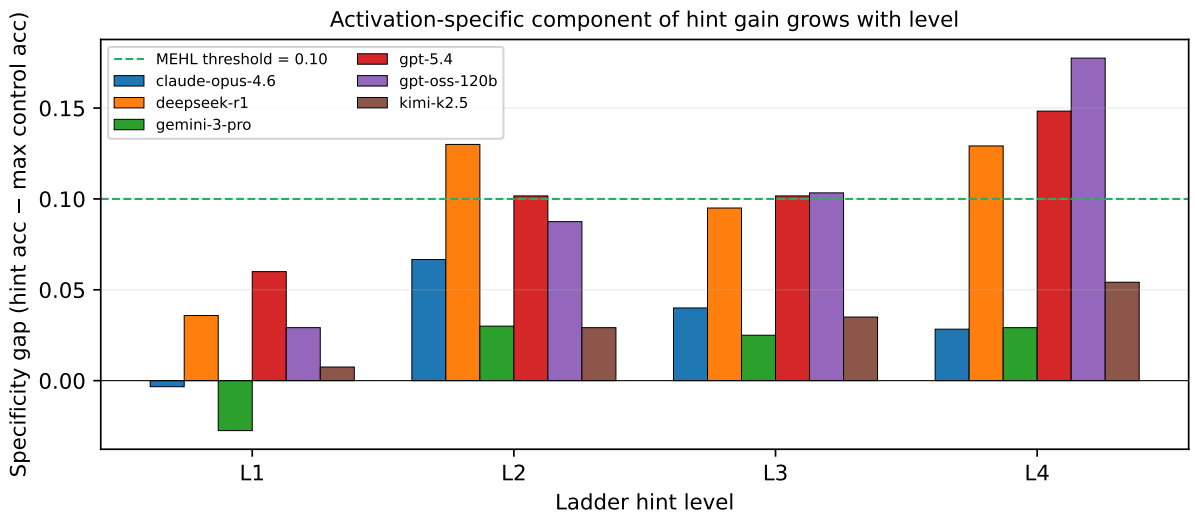

Figure 6: Hint-ladder specificity gap per level (hint accuracy minus best matched control). Reaches MEHL threshold (green dashed) at L_2 or L_3 for most models.

F Per-layer probe accuracy curve

Fig. 7 visualizes the per-layer probe accuracy reported numerically in Table 9.

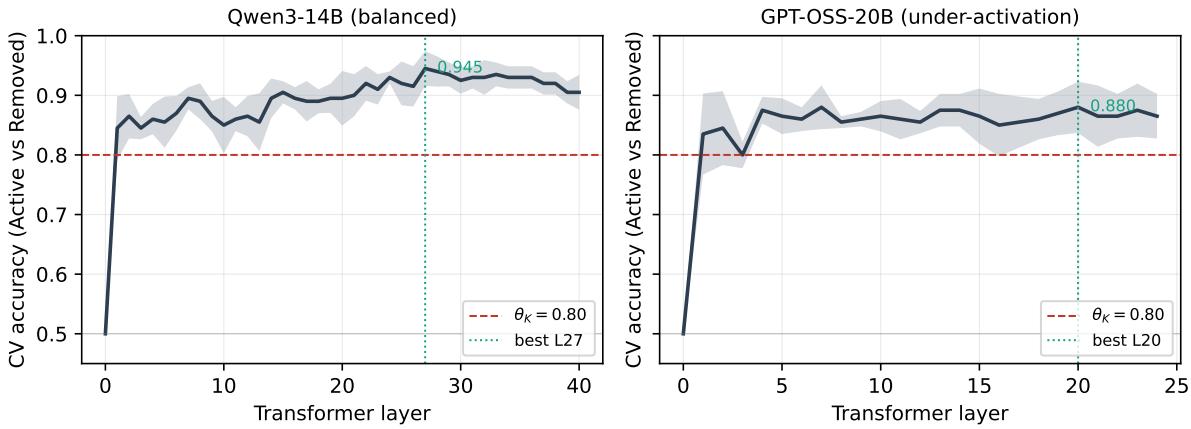

Figure 7: Per-layer linear-probe CV accuracy on the Active vs. Removed label. Shaded band is ± 1 SD over scenario-grouped folds; red dashed at $\theta_K = 0.80$; green dotted marks best layer.

G Methods detail: reproducibility

Code and data release. All prompts, the salience-control bank, mitigation prefixes, ladder templates, probe and patching code, and metric implementations will be released. We seed ladder construction (deterministic hash by `instance_id`) and probe training (`numpy` seed 42); all stochasticity is at the model-API level. The full quartet, ladder, mitigation, and budget collection is resume-safe by (instance_id, trial, strategy, budget) keys, with retry passes that preserve the original prompt prefix and thinking budget so re-runs are exchangeable with first-time runs.

Provider-specific budget routing. Anthropic models accept `thinking={enabled, budget_tokens=`b`}` with b clamped to a minimum of 1,024 tokens. OpenAI o-series and GPT-5 series accept `reasoning_effort`$\in$`{low, medium, high}`, mapped from b via thresholds 1024 and 4096. Gemini accepts `ThinkingConfig(thinking_budget=`b`)`. DeepSeek-R1 emits chain-of-thought traces but does not expose a budget knob; the requested b is recorded for parity but is a no-op.

Activation patching mechanics. For each scenario, we forward the donor (EXPLICIT) prompt and cache the final-token hidden state at the probe's best layer $\ell^\star$ via a PyTorch forward hook on `model.model.layers[`$\ell^\star$`]`. We then forward the recipient prompt, with a second hook that overwrites the layer-output tensor at the recipient's last position with the cached donor activation; logits are extracted at the same position and the gold–shortcut log-probability gap is recorded. The gap difference $\Delta = \Delta^{\text{patched}} - \Delta^{\text{pre}}$ is reported as the patching effect.

Prerequisite-mention detection. A model response is classified as "mentioning the prerequisite" iff it shares ≥ 2 distinct content-word matches with the canonical hidden constraint string, after lowercasing, stripping punctuation, and filtering a small stoplist of high-frequency English function words. This is a conservative, lexical-overlap definition; it may undercount paraphrastic mentions, which would only strengthen the mediated-correctness-share lower bounds reported in §4.3.

Judge. All $54{,}400 + 128{,}000 + 25{,}600 = 208{,}000$ trials are judged by Qwen3-32B with deterministic decoding, using the multiple-choice extraction prompt from Li et al. (2026). Trials with API errors at collection are retried with exponential backoff; verdicts for retried trials are recomputed by the same judge to maintain consistency.