Causal Tongue-Tie: LLMs Can Encode Causal Direction, But Their Yes/No Outputs Fail to Express

Anonymous ACL submission

Abstract

We find a mismatch between what large language models encode about a causal question and what they answer. On *anti-commonsense* CLADDER (Jin et al., 2023) items, a fixed linear probe recovers the evidence-supported answer from the model's hidden state (accuracy ≈ 0.97), while the spoken Yes/No reverts to the commonsense one (accuracy ≈ 0.5). We call this $\Delta \approx +0.5$ gap *Causal Tongue-Tie*: a wrong Yes/No decomposes into two separable failure modes—no internal signal vs. a signal the verbal interface cannot say. The implication cuts both ways for output-only causal benchmarks: a benchmark "correct" need not mean the model has understood, and a benchmark "wrong" need not mean it cannot. Sweeping claims about whether LLMs can do causal reasoning, drawn from a single accuracy number, deserve a second look.

1 Introduction

We find that on causal Yes/No questions, large language models (LLMs) can encode an evidence-consistent causal direction in their *hidden states* that the ordinary Yes/No output does not say—a mismatch we call *Causal Tongue-Tie* (Section 3 formalises "encode").

Background. Whether LLMs *reason* or merely produce fluent causal language is contested (Bender et al., 2021; Zečević et al., 2023). Benchmarks like CLADDER (Jin et al., 2023) and CounterBench (Chen et al., 2025) score models by Yes/No accuracy on causal queries; probe studies report mixed evidence about the hidden state (Chi et al., 2024; Geiger et al., 2025).

Motivation. On a CLADDER anti-commonsense item stipulating "smoking prevents lung cancer", Qwen2.5-7B answers "No" but a linear probe on the *same* hidden state returns "Yes". The internal state tracks the stipulation; the spoken answer reverts to prior. This resonates with the *Causal Parrots* concern (Zečević et al., 2023): fluent causal language need not mean causal reasoning. But a wrong Yes/No alone cannot tell us if the model lacks causal knowledge or has it but cannot say it. **Our contribution** is *Causal Tongue-Tie*: the empirical finding that hidden states encode the evidence-aligned causal answer while the spoken Yes/No reverts to commonsense (on anti-CS CLADDER: $\mathrm{Acc}_{\mathrm{probe}} \approx 0.97$ vs. $\mathrm{Acc}_{\mathrm{out}} \approx 0.50$). In a 2×2 view of "wrong on a causal benchmark" by internal signal $\times$ external Yes/No (Figure 1), this dissociation occupies the off-diagonal cell (b) that has not been systematically studied. Code, prompts, hidden-state artifacts, and per-item probe logits and Yes/No probabilities are at `https://anonymous.4open.science/r/tongue-tie-emnlp2026-BB3B`.

2 Related Work

Our finding sits at the intersection of two empirical lines: studies that compare hidden-state content with the model's text output, and the broader concern that fluent causal language need not reflect causal reasoning.

Hidden state vs. output. Several studies report that activations can contain information the textual answer does not faithfully express: Contrast-Consistent Search (Burns et al., 2023), inference-time intervention and truth-geometry directions (Li et al., 2023; Marks and Tegmark, 2023), and self-knowledge work (Kadavath et al., 2022). The same probe–output gap appears in clinical triage and hallucination detection (Basu et al., 2026; CH-Wang et al., 2024), and chain-

of-thought (CoT) faithfulness work shows the explanation text can be a poor witness to the computation that produced the answer (Lanham et al., 2023; Turpin et al., 2023). **Where we sit.** Prior work isolates either side: hidden state alone (probing / CCS / latent-knowledge) or output alone (causal-benchmark accuracy). We add a *measurement protocol* that pairs them on a causal task where the two sides reliably disagree, separating a wrong Yes/No on CLADDER from a missing internal signal. Anti-commonsense items provide the trigger, and wording-matched hard-negative controls rule out the surface-wording explanation rarely tested in prior latent-knowledge work.

Causal Parrots and causal benchmarks. Bender et al. (2021) and the *Causal Parrots* sharpening (Zečević et al., 2023) warn that LLMs may produce causal-looking answers by tracking surface regularities rather than causal evidence. Causal LLM benchmarks (Jin et al., 2023; Chen et al., 2025; Frohberg and Binder, 2022; Chi et al., 2024) operationalise this concern through output-side accuracy; we add a second axis: an internal readout channel from which an evidence-consistent direction is recoverable while the Yes/No does not express it. We use standard probing and direction-erasure tools (Hewitt and Liang, 2019; Belrose et al., 2023; Feder et al., 2021; Zou et al., 2023) as measurement instruments, not as our contribution.

3 The Finding

The central finding is that a wrong Yes/No on a causal benchmark splits into four states (Figure 1) by internal signal vs. external answer. Among these, the cell we call *Causal Tongue-Tie*—(b), *probe finds the evidence-aligned answer, but the spoken Yes/No reverts to commonsense*—is the failure mode output-only accuracy hides: the hidden state knows, the answer does not say so. Our paper is about (b). The rest of this section fixes the setting and notation (Section 3.1), pins down what we mean by *encode* (Section 3.2), states how we quantify the gap (Section 3.3), and lists the empirical claims we will test in Section 3.4; a representative cell-(b) item is illustrated in Figure 2.

3.1 Setting and notation

To make Causal Tongue-Tie precise, we need a single forward pass on which we can read two things side by side: what the hidden state contains, and what the model says out loud. We therefore work with a frozen instruction-tuned transformer decoder f_θ with L residual blocks. For each prompt x_i we read the activation $\boldsymbol{h}_i^{(\ell)} \in \mathbb{R}^d$ after block ℓ at the *last prompt token* position, and on the same pass we read the next-token Yes/No probabilities through the model's own unembedding $\texttt{lm_head}: \mathbb{R}^d \to \mathbb{R}^{|\mathcal{V}|}$. Our linear probe (defined below) reads $\boldsymbol{h}_i^{(\ell)}$, and `lm_head` reads $\boldsymbol{h}_i^{(L)}$, both at the same token position, so on every item we can compare directly what is internally readable with what is actually said (Section A gives a self-contained refresher on the decoder modules and residual stream).

The items we audit are causal Yes/No questions, each with a gold answer. We write the dataset as

$$\mathcal{D} = \{(x_i, y_i, c_i, e_i, b_i, r_i)\}_{i=1}^{N},$$

where x_i is the prompt, $y_i \in \{\text{Yes}, \text{No}\}$ is the gold answer, (c_i, e_i) is the candidate cause–effect pair, and r_i marks the evidence type. The subset tag $b_i \in \{\text{cs}, \text{anti-CS}, \text{ns}\}$ splits the data into three flavours that will drive the rest of the paper: *commonsense-aligned* items (cs, e.g. *smoking causes lung cancer*), where the prompt evidence agrees with everyday commonsense; *anti-commonsense* items (anti-CS), where the in-prompt evidence is constructed to point the opposite way from the commonsense prior; and *nonsense* controls (ns), whose prompts use placeholder entity tokens with no real-world content and which serve as a contamination-free baseline.

To isolate the direction of the answer inside the hidden state, we compare each prompt x_i with a *counterfactual minimal pair* x_i^{cf} that swaps only the words that flip the answer—e.g. *vaccinated* vs. *not vaccinated*, or *cause* vs. *prevent*—so the evidence-supported answer flips while the surface form is otherwise identical (Chen et al., 2025; Li et al., 2023; Marks and Tegmark, 2023; Geiger et al., 2025). The hidden-state difference at layer ℓ,

$$\Delta\boldsymbol{h}_i^{(\ell)} = \boldsymbol{h}_i^{(\ell)} - \boldsymbol{h}_i^{(\text{cf},\ell)} \in \mathbb{R}^d, \qquad (1)$$

is the answer-direction signal we will read in two complementary ways. First, a *linear readout (frozen)*

	External Yes/No *matches* evidence	External Yes/No *contradicts* evidence
Internal: probe finds the direction	**(a) Aligned correct.** Both readouts pick the evidence-supported answer.	**(b) ★ CAUSAL TONGUE-TIE** *(this paper).* Probe finds the evidence-aligned answer; Yes/No reverts to commonsense.
Internal: probe finds nothing	**(c) ♢ CAUSAL PARROT risk.** Right Yes/No by surface heuristic; no internal evidence-direction.	**(d) Genuine failure.** Neither readout carries the evidence-supported answer.

Figure 1: **A 2×2 framework for a wrong Yes/No on a causal benchmark.** The two off-diagonal cells—CAUSAL TONGUE-TIE (b, this paper) and the "Causal Parrot" (c, Zečević et al. 2023)—are what output-only accuracy hides when it collapses (a)+(c) into "correct" and (b)+(d) into "failed".

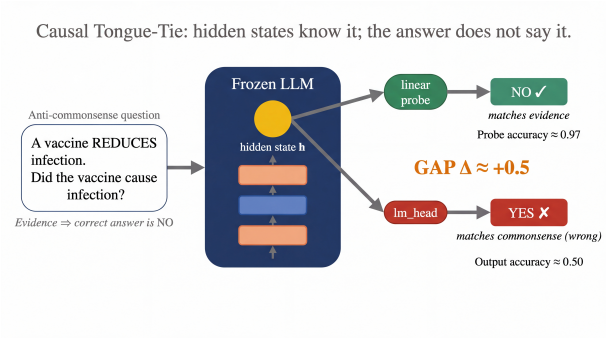

Figure 2: **Cell (b) on one anti-CS item.** From the same hidden state, the probe recovers **NO** ($\mathrm{Acc}_{\mathrm{probe}}$≈0.97) while `lm_head` produces **YES** ($\mathrm{Acc}_{\mathrm{out}}$≈0.50); the Δ≈+0.5 gap is the empirical handle on *Causal Tongue-Tie*.

is an ℓ_2-regularised logistic regression $g_{\mathbf{w}}(\boldsymbol{h}) = \sigma(\mathbf{w}^\top \boldsymbol{h} + b)$, fit *once* on hidden states from cs items only and then frozen for every reported number; the language model itself is never updated. Second, the same idea can be expressed as a direction in $\mathbb{R}^d$: stacking the cs paired differences into a matrix $\mathbf{D}_{\mathrm{cs}}^{(\ell)} \in \mathbb{R}^{N_{\mathrm{cs}} \times d}$ (rows $\Delta \boldsymbol{h}_i^{(\ell)}$, $i \in$ cs) and taking its truncated SVD,

$$\mathbf{D}_{\mathrm{cs}}^{(\ell)} = U \Sigma V^\top, \qquad V_{\mathrm{cs}} = V_{:,1:k} \in \mathbb{R}^{d \times k}, \qquad (2)$$

defines the *counterfactual subspace* V_{cs} as the top-k right singular vectors of that matrix (Park et al., 2024, 2025; Burns et al., 2023; Li et al., 2023).

Both the readout and V_{cs} are fit on cs alone and reused frozen everywhere else; k is small, with $k \in \{1, 2, 4\}$ and the simple mean of $\Delta \boldsymbol{h}_i^{(\ell)}$ giving the same qualitative gap on cs and anti-CS (Section E).

A readout pipeline of any reasonable shape might still appear to pick up a direction by accident, so we rule out the most obvious non-causal alternatives with two same-shape controls. The *sham subspace* rebuilds the entire fit-then-freeze pipeline on the ns subset (whose prompts have no real-world content), and the *random-direction control* replaces V_{cs} with Haar-random vectors of the same rank. Both controls share the readout's shape but should be uninformative, and we use them to set the chance baseline in the definition of *encode* below.

3.2 What "encode" means here

With those pieces in place, we say precisely what we mean when we claim the hidden state *encodes* the evidence-supported answer on an item: on the same frozen forward pass, **(i)** the linear readout recovers that answer; **(ii)** on the counterfactual minimal pair the same readout recovers the opposite direction (so it tracks the causal direction, not the wording); and **(iii)** both same-shape controls (sham and random direction) sit at chance on (i)–(ii), so the readout is specific to the cs causal direction, not anything a 7B hidden state happens to support. The claim is deliberately narrow—this is a hidden-state-vs.-output gap, not human-style understanding—and we use it throughout the paper.

3.3 How we measure the gap

We quantify the gap with two accuracies on the same item. $\mathrm{Acc}_{\mathrm{probe}}(\ell)$ is the rate at which the frozen linear readout from $\boldsymbol{h}_i^{(\ell)}$ recovers the evidence-supported direction; $\mathrm{Acc}_{\mathrm{out}}$ is the rate at which the model's ordinary Yes/No probabilities pick the same answer. The empirical gap is

$$\Delta = \mathrm{Acc}_{\mathrm{probe}}(\ell^\star) - \mathrm{Acc}_{\mathrm{out}}, \qquad (3)$$

where $\ell^\star$ is the best probing layer chosen on commonsense items only (Section 4; hyperparameters in Section E); anti-commonsense and the hard-negative families are held out from this choice. A large positive Δ is the numerical form of Causal Tongue-Tie: the hidden-state readout matches the evidence-supported answer much more often than the spoken Yes/No does. A small or negative Δ says the mismatch is absent under this test.

3.4 The empirical claims

The rest of the paper tests three empirical claims, each tied to a specific subsection of the evidence.

Claim 1. Δ is largest when the prompt evidence points against commonsense: the hidden-state readout still matches the evidence-supported answer while the Yes/No output falls back to a weak or misleading response (checked in Section 5.1).

Claim 2. The readout direction V_{cs} is not random variation: same-shape sham, random-direction, and pair controls all support it as a stable, readable direction, even though it is not itself a switch that flips the final answer (checked in Section 5.2).

Claim 3. Δ is not explained away by visible causal words or candidate-token surface cues alone; those cues can matter, but they do not account for when the readout direction is recoverable and when the Yes/No answer expresses it (checked in Sections 6.1 and 6.2).

Roadmap. Claims 1–3 map to Sections 5.1, 5.2, 6.1 and 6.2; a final scope check (Section 6.3) localises the failure to the verbal interface rather than to one particular Yes/No surface form.

4 How We Checked It

For each item, two readout paths branch off the same single inference run of the model (a *forward pass*): a linear probe on the hidden state, and the model's own `lm_head`—the language-model head, i.e. the final linear layer that maps the hidden state to next-token logits—producing the ordinary Yes/No token. We compare them on the same item (Figure 2).

Paired questions and hidden states. For each item we use a matched *causal-direction pair*: the prompt x_i asks whether c_i causes e_i, and the counterfactual x_i^{cf} asks whether e_i causes c_i—the *counterfactual minimal pair* primitive of Section 3 (Geiger et al., 2025; Li et al., 2023; Marks and Tegmark, 2023).

Probe versus ordinary answer. Throughout, our *linear probe* is a standard ℓ_2-regularised logistic-regression probe (Hewitt and Liang, 2019) fit on the hidden state at a chosen layer (no LLM parameter is updated). We use item-disjoint CV (each prompt and its flipped pair in the same fold) and a shuffled-label control (Hewitt and Liang, 2019).

Controls and interface checks. Around the readout direction V_{cs} (the SVD subspace defined in Equation (2) of Section 3) we apply two simple operations to test what V_{cs} does inside the model.

(i) Projecting V_{cs} out of the hidden state. We replace $\boldsymbol{h}_i^{(\ell)}$ by its orthogonal projection $\boldsymbol{h}_i^{(\ell)} - V_{\text{cs}}V_{\text{cs}}^\top \boldsymbol{h}_i^{(\ell)}$, removing any component in the V_{cs} subspace—a one-subspace instance of the direction-erasure idea (Ravfogel et al., 2020; Belrose et al., 2023)—and ask whether the model can still distinguish the paired prompts internally. A sharp drop in the counterfactual KL (defined in Section 5.2) would mean V_{cs} was participating in that distinction.

(ii) Swapping the scalar projection onto V_{cs}. The scalar projection $\alpha_i = V_{\text{cs}}^\top \boldsymbol{h}_i^{(\ell)}$ is a single real number. We replace α_i by the corresponding value from the paired prompt—a one-dimensional activation patch (Vig et al., 2020; Meng et al., 2022)—and check whether changing only that one number is enough to flip the final Yes/No answer.

We then re-run the readout under two additional families: hard-negative controls (Section 6.2) and more structured answer interfaces (Section 6.3).

5 Evidence: The Answer Is Inside, But Not Said

Setup. We run the same probe-vs-Yes/No audit on Qwen2.5-Instruct (Qwen Team et al., 2024) $\{0.5, 1.5, 3, 7, 14\}$B, Qwen2.5-32B-NF4, Qwen2.5-72B-NF4, Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and a focused DeepSeek-LLM-7B-Chat (DeepSeek-AI, 2024) replication, on anti-commonsense CLADDER (Jin et al., 2023), COUNTERBENCH (Chen et al., 2025), and the hard-negative items used later. All accuracies and KL drops use paired-bootstrap 95% intervals. The best probing layer $\ell^\star$ is selected on commonsense items only and reused unchanged elsewhere, so the gap is not a selection effect on the reported items.

5.1 The internal answer scales; the spoken answer does not

Finding. Across the eight instruction-tuned models we audit (Qwen 0.5B–72B, Mistral-7B, DeepSeek-7B), and on 80 *anti-commonsense* CLADDER items—items whose in-prompt evidence forces the answer that contradicts everyday commonsense (e.g. a fictional world in which smoking *prevents* lung cancer)—a simple linear function of the model's hidden state at the last prompt token recovers the correct, evidence-supported answer with accuracy ≥ 0.969 on every model. The ordinary next-token Yes/No, asked on the *same* items, lands at 0.350–0.525 across the eight models. The gap $\Delta \approx +0.5$ is therefore not specific to one model or one scale: from 0.5B to 72B the readable internal answer goes up while the spoken answer does not (Figure 3).

Two everyday worries. Two worries arise: *(a)* is the probe just memorizing surface wording, so any classifier of that shape would already score this high? *(b)* is the low Yes/No really an error, or just imbalanced gold labels? We rule both out with four validity checks.

(i) Is the wrong Yes/No just a label imbalance? The gold-Yes rate on these anti-commonsense items is 0.475, so the 0.35–0.53 output accuracy is not what we would get from blindly answering "Yes" (or "No") on every item. *(ii) Did we pick a flattering layer?* At two layers fixed *before* the per-model best-layer search (layers 8 and 16) the probe still reaches ≥ 0.919, so the gap is not a best-layer cherry-pick. *(iii) Is the readable signal already an ordinary answer at that depth?* A *logit-lens* check, which sends the intermediate hidden state straight through the model's own final layer normalization and `lm_head` as if it were the last-layer state, scores only 0.500–0.583. So the linear probe is reading something that the model's own output pipeline at that depth does not yet treat as an answer. *(iv) Is the probe shape just expressive enough to memorize anything?* A control probe trained on randomly shuffled labels (Hewitt and Liang, 2019)—the standard null-distribution baseline—stays well below the real probe; selectivity is +0.302–+0.361 with bootstrap $p(\text{control} \geq \text{real}) = 0$. The focused DeepSeek run reproduces the same pattern: probe 0.975, output 0.475, gap +0.500 (per-model numbers and full validity audits in Sections C, D and F).

Lesion	$D_{\mathrm{KL}}^{\mathrm{cs}}$	$D_{\mathrm{KL}}^{\mathrm{ns}}$	cs drop (%)
baseline	0.252 [.240, .265]	0.210 [.161, .264]	0.0
V_{cs}	**0.009** [.008, .012]	0.161 [.123, .204]	**−96.3** [–97.0, –95.4]
V_{ns}	0.172 [.163, .183]	**0.056** [.047, .067]	−31.7 [–33.7, –29.7]
V_{rand}	0.254 [.242, .267]	0.213 [.163, .270]	+0.7 [–1.3, +2.8]

Table 1: Projecting out one direction on Qwen-7B layer 27 (N=30 pairs). $D_{\mathrm{KL}}^{\mathrm{cs}}$ ($D_{\mathrm{KL}}^{\mathrm{ns}}$) is the mean KL between prompt and causally-flipped counterpart on the cs (commonsense) / ns (nonsense) subset. Removing V_{cs} nearly erases the counterfactual KL on cs; same-size shams do not.

5.2 The internal direction is used—a single-knob push does not flip the answer

Finding. Is the readable direction actually *used* by the model? *(a)* Erasing it inside the hidden state should collapse the model's internal comparison of paired prompts; *(b) pushing* it in the commonsense direction should flip the spoken Yes/No. The answer to (a) is yes; (b) is mixed—a single-direction push is not enough, but injecting the whole hidden state from a matched commonsense item is.

(a) Erasing the direction: counterfactual-KL lesion. We use a standard direction-erasure recipe (Ravfogel et al., 2020; Belrose et al., 2023): for each prompt build its causally flipped counterpart (swap cause/effect so the answer flips), measure the KL divergence between the two next-token distributions ($D_{\mathrm{KL}}^{\mathrm{cs}}$ on commonsense pairs, $D_{\mathrm{KL}}^{\mathrm{ns}}$ on nonsense pairs), project V_{cs} out of the hidden state, and re-measure. If the direction carries the comparison, this KL should drop sharply.

Figure 4 (with exact numbers in Table 1) reports the result on Qwen-7B layer 27 (N=30 minimal pairs). Projecting out V_{cs} collapses $D_{\mathrm{KL}}^{\mathrm{cs}}$ by -96.3%, while same-shape controls do not: V_{ns} gives -31.7%, Haar-random V_{rand} only $+0.7\%$, and LEACE (Belrose et al., 2023)—a more aggressive "erase all linearly readable info about the label" baseline—only -21.6%. Because the three lesions share rank and projection shape, an off-manifold confound would predict similar drops; the V_{rand} control's CI includes zero, ruling that out. So V_{cs} participates in the model's internal comparison: it is not just any same-size direction.

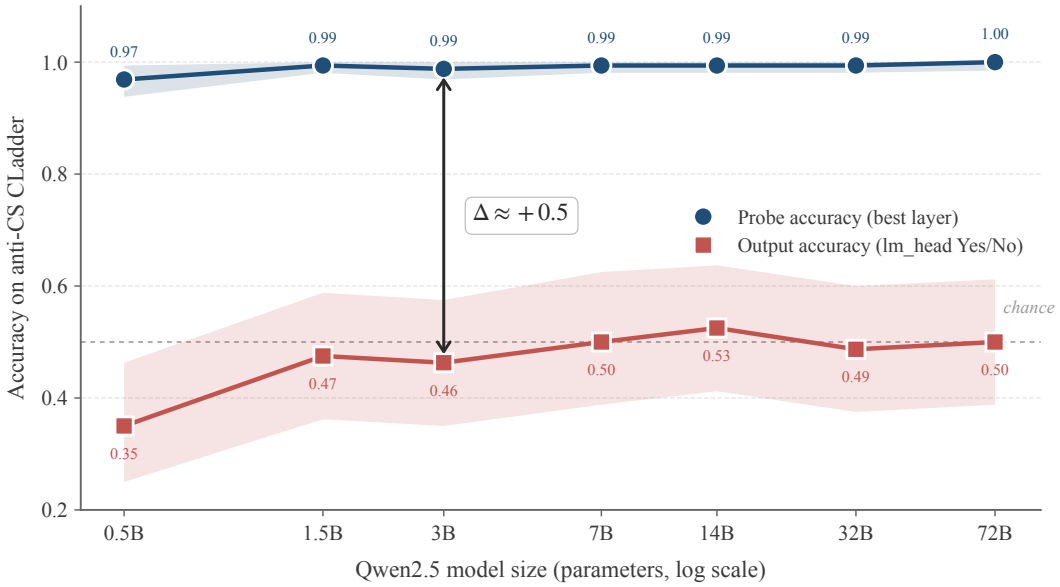

Figure 3: Anti-commonsense CLADDER accuracy across Qwen2.5 instruction-tuned model sizes (log-scale x-axis, 0.5B–72B). The hidden-state readout (*Probe accuracy, best layer*, deep teal) is already 0.969 on Qwen2.5-0.5B and reaches 1.000 on Qwen2.5-72B-NF4, while ordinary Yes/No accuracy (*Output accuracy (lm_head)*, warm coral) stays in $[0.350, 0.525]$; the gap $\Delta \approx +0.5$ does not close with model size. Shaded ribbons are 95% confidence intervals on the N=80 anti-commonsense prompts; *chance* marks 0.5 binary-task accuracy. Cross-family replication on Mistral-7B-Instruct-v0.3 and DeepSeek-LLM-7B-Chat is reported in Section F.

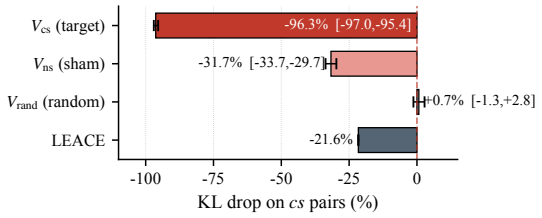

Figure 4: Counterfactual-KL lesion on Qwen-7B layer 27 (N=30 minimal pairs). Projecting out V_{cs} erases 96% of the cs-counterfactual KL; the rank-matched sham V_{ns}, the Haar-random V_{rand} (CI includes zero), and the LEACE label-erasure baseline do not. Whiskers are 95% paired-bootstrap CIs.

(b) Pushing the direction: scalar swap and full-state injection. The complementary test asks whether *moving* the same direction flips the spoken answer. The simplest version is a *scalar swap*: we leave the hidden state otherwise alone but rescale its component along V_{cs} to a more commonsense-like value (and, as controls, along the sham direction V_{ns} and a Haar-random direction). Even pushed to $+2\sigma_{\text{cs}}$ none of the three reliably moves the spoken Yes/No (accuracy shifts by at most $+0.025$, indistinguishable from baseline). We label this a *finite-power null*—"no detectable effect at this sample size and this size of single-direction push"—rather than a strong negative.

A scalar swap touches one direction at one scalar; the full-state counterpart is *activation patching*, the standard interpretability procedure of copying a whole layer's hidden state from a donor prompt into the recipient (Vig et al., 2020; Meng et al., 2022). On the N=42 anti-commonsense items Qwen2.5-7B-Instruct gets wrong, replacing the layer-27 last-token hidden state with that of a matched commonsense item lifts accuracy from 0 to 0.571, well above random-hidden and self-injection controls (Figure 5; Section N for the full table).

Reading (a) and (b) together. V_{cs} is *not* a one-knob controller of the Yes/No (the scalar swap is a finite-power null; full 8-condition table in Section H), but the full layer-27 state at the same depth *does* carry enough commonsense-side information to flip more than half of the previously-wrong spoken answers on its own.

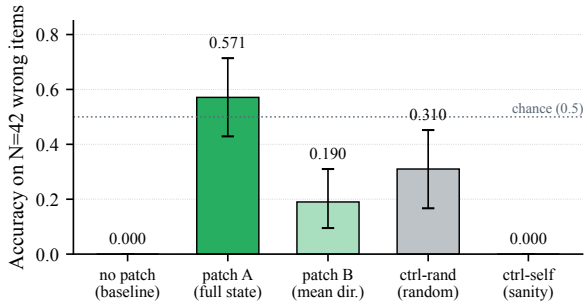

Figure 5: Activation patching on the $N{=}42$ anti-CS items Qwen2.5-7B-Instruct gets wrong (layer-27 last-token state). Bars left-to-right: *no patch* (wrong-item baseline, 0 by definition); *patch A (full state)* (replace the state with a matched-CS donor; 0.571); *patch B (mean dir.)* (inject only the mean V_{cs} direction); *ctrl-rand* (random-hidden donor); *ctrl-self* (self-injection sanity). Dashed line: 0.5 binary chance. Whiskers: 95% paired-bootstrap CIs.

Subset (Qwen-14B L48, $N{=}20$)	target	value [95% CI]
direct edge, *no* causal words	TP	1.000 [.839, 1.000]‡
causal word *vs.* evidence conflict	FP	0.000 [.000, .161]‡
correlation-only (old template)	FP	0.550 [.342, .742]
common-cause (old template)	FP	1.000 [.839, 1.000]‡

Table 2: Causal-word diagnostic on Qwen-14B layer 48. *TP* = True-Positive rate (probe correctly accepts the evidence-supported causal direction); *FP* = False-Positive rate (probe incorrectly accepts a non-causal direction). The top two rows test whether explicit causal words are necessary/sufficient; the bottom two rows expose old-template ambiguity that motivates the stronger Question 4 controls. ‡Saturated by construction.

6 What It Is Not, and Where It Breaks

This section in one sentence. The gap is *not* just a reaction to causal vocabulary (Section 6.1), *not* just statistical association between cause and effect words (Section 6.2), and the failure point is the *ordinary verbal answer interface*—not the internal direction (Section 6.3). Together, these three exclusions narrow what Causal Tongue-Tie can be: the readable internal direction is real; what breaks is the plain-language Yes/No commitment step.

6.1 It is not just causal wording

The wording control rules out a simple causal-word explanation. A skeptical reading of the gap is that "the probe just follows surface causal words (cause/affect/influence) in the prompt, not anything causal"—a hypothesis raised in earlier critical work on causal-LLM benchmarks (Zečević et al., 2023). We control for it by varying the causal vocabulary while keeping the underlying causal structure of the item fixed, and ask whether the gap tracks the wording or the structure.

Explicit causal words are neither necessary nor sufficient for the observed accept/reject pattern. The TP/FP rates are saturated at $1.000^{\ddagger}/0.000^{\ddagger}$ by construction (each item admits a unique correct causal-word answer under the template); the key signal is the *contrast* on the *same* controlled items: the probe attains this rate while the spoken Yes/No drops to $\sim$0.5. Removing causal verbs from gold-Yes direct-edge items leaves TP at $1.000^{\ddagger}$, while adding "a causes b" wording against the stated evidence keeps FP at $0.000^{\ddagger}$ (Table 2). The model is therefore not merely following the presence of causal vocabulary.

The two older-template rows reflect a template confound between $a \rightarrow b$ and $a \leftarrow c \rightarrow b$, which motivates the stronger Question 4 controls below; they are not support for the wording explanation.

6.2 Not just cause/effect word association—with limits

Finding. Even when we explicitly separate "a causes b" (direct edge), "a and b share a common cause", and "just observe a, b co-occur" (correlation only) on the *same* controlled items, both the probe and the spoken Yes/No track the gold *direct-edge* label—so the gap survives the stricter association-vs-direct-edge separation, not just "the probe learned word co-occurrence".

Concretely. We built six families that systematically separate these readings: *direct* ($a \rightarrow b$), *common-cause* ($a \leftarrow c \rightarrow b$), *correlation-only* (no edge), *causal-word conflict* ("cause" wording vs. evidence), and two *observationally matched* families with identical surface-word conditionals, distinguishable only via the prompt's intervention information. On Qwen-14B layer 48, both readouts track the gold

direct-edge label across all six families (N=24 per family): probe Acc = 1.000 everywhere, and the spoken yes-rate is $24/24$ on the two direct families and $0/24$ on the four non-direct ones. Mistral-7B-Instruct-v0.3 reproduces the same pattern (full per-family table in Section J); on Qwen2.5-7B-Instruct the multi-template variant has held-out Acc 0.917.

Limits of the control. Joint holdout of both obs-matched families drops obs-match-direct probe Acc to 0.083 while the answer stays at 1.000 (a *probe-generalization boundary*: a limit of probe transfer, not evidence about the hidden state). The multi-template variant exposes one brittle case (correlation-only with intervention-first phrasing: behaviour Acc 0.375, probe Acc 0.167), and terse phrasing flips Mistral's correlation-only answer while the probe stays high. Two appendix audits bound this further: a 5-template paraphrase suite (Table 11) replicates the qualitative gap on every template, but a fitted probe direction does *not* fully transfer cross-template (Table 12, off-diagonal mean 0.672); and a CLadder $\rightarrow$ CounterBench transfer (Section B) reaches 0.844–0.913 at layer 24 but only 0.506–0.606 at layer 16.

6.3 The fragile point is the ordinary answer interface

Not the structured edge choice. More structured readouts recover the answer direction; the failure is mainly in how the answer is verbalized.

Five *answer interfaces* share the same prompt body and differ only in the answer slot: **ordinary cause** (Yes/No to "Did c cause e?"—the standard benchmark interface), **real cause** (Yes/No with a *real* hedge), **direct effect** (free-form), **bridge/arrow** ("$c \rightarrow e$ vs. $e \rightarrow c$"), and **A/B edge** (multiple-choice $c \rightarrow e$ vs. $e \rightarrow c$).

On Qwen-7B anti-CS, the ladder reads 0.263/0.588/0.713/0.988 for cause/bridge/arrow/A/B (Figure 6); the same ordering A/B-edge > arrow > direct-effect > ordinary-cause > real-cause holds across Qwen2.5 sizes (7B/14B/32B/72B) and replicates on Mistral-7B-Instruct, and masking candidate tokens as <C>/<E> keeps pair separation high (0.963 $\rightarrow$ 1.000). Prompting and calibration only partly patch the failure: CoT and self-consistency lift Qwen-7B anti-CS yes-rate from 0.113 to 0.787/0.825, but free-form, structured-to-text bridge, Platt scaling, and V_{cs} steering all stay close to the original failure (bridge 0.075 vs. A/B 1.000). A token-level entropy breakdown (Section M) localises this on the interface side—Yes/No averages $\approx 0.32 \ln 2$ on a *confident wrong-direction* commit vs. A/B at $\approx 0.09 \ln 2$—and the gap survives Yes/No $\rightarrow$ True/False, *Shi/Fou*, and A/B-reframe substitutions ($\Delta \geq +0.595$, Section O).

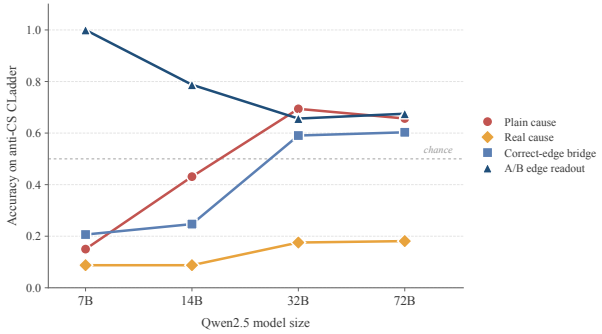

Figure 6: Answer-interface ladder across Qwen2.5-Instruct sizes (7B / 14B / 32B / 72B). Four of the five answer interfaces are plotted (the fifth, *direct effect*, is free-form and is reported separately in Table 14). Plot legend: *Plain cause* = ordinary Yes/No ("Did c cause e?"); *Real cause* = Yes/No with the *real* hedge; *Correct-edge bridge* = bridge/arrow free-form; *A/B edge readout* = A/B forced-choice $c \rightarrow e$ vs. $e \rightarrow c$. A/B-edge and the structured *Correct-edge bridge* are near-saturated across Qwen2.5 sizes while *Plain cause* and *Real cause* collapse: the gap is in the verbal answer interface, not in the readable internal direction.

7 Discussion and Conclusion

We introduced *Causal Tongue-Tie*: hidden states often carry an evidence-aligned causal direction that the ordinary Yes/No does not say (Sections 5.1 and 5.2; controls in Sections 6.2 and 6.3), and full-state injection flips over half the wrong Qwen2.5-7B-Instruct items (Section N). A wrong Yes/No thus decomposes into two failure modes—no internal signal vs. an unexpressed one—not a single "failed" verdict.

Implication for benchmark-based claims. The gap warns against output-only evaluation: hidden-state and Yes/No verdicts can differ by $\Delta \approx +0.5$, while prompt form also matters (Section O). Single-number claims conflate all three.

Limitations

Scope of the claim. Causal Tongue-Tie is an empirical diagnostic of a mismatch between hidden-state evidence and the final Yes/No answer. Following the operational definition in Section 3, we claim that the model encodes causal direction only in that narrow sense, not human-like understanding, complete causal reasoning, or a causal circuit; the direction is causally sufficient on $>50\%$ but not all items the ordinary interface gets wrong (Section N).

Scope of the finding. Our gap is a hidden-state-vs-output mismatch on *textual* causal questions (a short story plus a Yes/No probe). It is *not* a finding about causal discovery from raw observational data such as physical sensor readings, time-series, or experimental measurements; those tasks have very different inputs and evaluation, and lie outside our scope. A local contamination heuristic (Table 18) reports anti-CS marker rate $= 0.000$ and ns / COUNTERBENCH nonsense-entity coverage $= 1.000$, so the audited items are unlikely to be verbatim pretraining text; a rigorous n-gram contamination audit is left to future work.

Direction-label confounding and wording sensitivity. The original direction label in Section 5.1 is partly confounded with question format, so the controlled hard-negative diagnostic—not the Section 5.1 probe alone—anchors the readout claim. Even there, the multi-template run exposes a brittle family, so the recovered direction should still be read as tied to the tested wording and syntax. The obs-matched hard-negative families are template-generated; 5 hand-written templates (Table 11, including a Chinese paraphrase, passive, and indirect phrasings) all reproduce the qualitative gap (per-template probe $\mathrm{Acc} \geq 0.95$, gap $\geq +0.26$), but cross-template probe transfer is fragile (Table 12, off-diagonal mean 0.672); the recovered direction is template-conditioned in its parameterization. Full human-style paraphrases beyond these 5 templates remain future work.

Limited interface repair and partial sufficiency. Prompting narrows but does not close the interface gap. The scalar V_{cs} swap (Section 5.2) is a finite-power null on full Yes/No control; full-state activation patching (Section N) is causally sufficient on $>50\%$ but not all of the items the ordinary interface gets wrong, so we demonstrate a partially exploited internal direction, not a complete causal control path. The diagnostic is single-direction, single-layer, and scalar; manifold or multi-mode geometries, norm-preserving multi-layer lesions, and low-rank or cone-shaped interventions could close the remaining sufficiency gap. Forced-choice decoding is not standardized across surface forms, but the Yes/No vs. True/False vs. Chinese (*Shi*/*Fou*) vs. A/B-reframe comparison on Qwen2.5-7B-Instruct (Section O) keeps $\Delta \geq +0.595$ across all four; a fully multi-token-calibrated protocol over the entire model suite is still left to future work.

Coverage of models and controls. The evidence is limited by the models and tests we could run. The full diagnostic suite is centered on Qwen2.5-Instruct and Mistral-7B-Instruct-v0.3; DeepSeek-LLM-7B-Chat supports the basic probe–answer gap but not the full Sections 5.2 and 6.1 to 6.3 analysis. These checks bound the claim but do not exhaust other model families, layers, prompts, interventions, or lexical shortcuts. Several controls use smaller per-condition N ($N \in [20, 30]$ in Sections 6.1 and 6.2; $N{=}24$ per hard-negative family), widening their CIs relative to the headline anti-CS gap on $N{=}80$. All eight backbones are instruction-tuned; auditing base checkpoints would help separate interface failure from RLHF answer policy.

Probe training set and cross-dataset transfer. The probe in the main text is always fit on CLadder commonsense items. We report a preliminary cross-dataset transfer evaluation in Section B: a CLadder-trained probe transfers to CounterBench at 0.844–0.913 accuracy (layer 24), and an anti-CS-trained probe is nearly identical to a cs-trained probe (symmetry gap 0.006). These results argue against surface-marker overfitting and support cross-dataset generalization of the direction readout. Cross-dataset transfer is *layer-conditioned*: the headline 0.844–0.913 requires layer 24, while layer 16 (the layer used elsewhere in this paper) transfers only 0.506–0.606. We report both to avoid layer-selection on the transfer test. A comprehensive multi-benchmark transfer study—testing the same probe weights on CLadder, Causal-NLI, FaithDial, or other causal benchmarks—remains an open empirical test for fu-

ture work.

Ethical Considerations

Data and model provenance. We use two public causal-reasoning benchmarks, CLadder and CounterBench, both MIT-licensed and free of personal or sensitive information. The eight instruction-tuned checkpoints we probe—Qwen 2.5 (0.5B–72B), DeepSeek-LLM-7B, and Mistral-7B-Instruct-v0.3—are publicly released models used only for research under their licenses. Prompts, hidden states, per-item probe logits, and Yes/No probabilities are released via anonymous.4open.science.

Intended use and dual-use considerations. This work diagnoses an alignment failure between a model's internal representation and its Yes/No output; it provides no deployable fix or jailbreak path. Hidden-state read-out applies only to open-weight checkpoints—adding no attack surface against closed-API models—and our structured prompts modify only the input, not the weights. The reported hidden-state-vs-output gap is a measurement statement and should not be cited as evidence that a model truly understands causality.

Compute and environmental cost. All probing and interface experiments ran on a mixed-GPU node of consumer RTX 4090 cards, with NF4-quantized offload to H20 GPUs for the $\geq$ 32B checkpoints; we performed no fine-tuning or gradient updates, so total compute is far below training a comparable model. Aggregated GPU-hours, per-sample inference latency, and an energy/carbon estimate are reported in Sections E.1 and L.

Use of AI writing assistance. Experimental design, claim formation, result analysis, and final editorial decisions were made by the authors. For wording polish, sentence-level rewrites, and draft figure and table captions, the authors used large language models as writing aids, consistent with ACL's policy on Generative Assistance in Authorship. All cited references were verified manually.

References

Sanjay Basu, Sadiq Y. Patel, Parth Sheth, Bhairavi Muralidharan, Namrata Elamaran, Aakriti Kinra, John Morgan, and Rajaie Batniji. 2026. Interpretability without actionability: Mechanistic methods cannot correct language model errors despite near-perfect internal representations. *arXiv preprint arXiv:2603.18353*.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 66044–66063.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations (ICLR)*. Contrast-Consistent Search (CCS).

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4401–4420.

Yuefei Chen, Vivek K. Singh, Jing Ma, and Ruixiang Tang. 2025. CounterBench: Evaluating and improving counterfactual reasoning in large language models. *arXiv preprint arXiv:2502.11008*.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024. Unveiling causal reasoning in large language models: Reality or mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*.

DeepSeek-AI. 2024. Deepseek LLM: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954. *Preprint*, arXiv:2401.02954.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386. Preprint arXiv:2005.13407 (2020); journal version 2021.

Jörg Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 2126–2140.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah D. Goodman, Christopher Potts, and Thomas Icard. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research (JMLR)*, 26(83):1–64. ArXiv:2301.04709, first posted 2023; final JMLR version 2025.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. arXiv preprint arXiv:2307.13702. *Preprint*, arXiv:2307.13702.

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 (NeurIPS)*.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*. Subsequently presented at the Conference on Language Modeling (COLM) 2024.

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

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2025. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 39643–39666.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, and 26 others. 2024. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115. *Preprint*, arXiv:2412.15115.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7237–7256.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2305.04388.

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

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research (TMLR)*, 2023.

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 Transformer decoder and the hidden state we read

This subsection is a minimal, self-contained refresher on the part of the transformer decoder that our probe touches; readers familiar with the residual stream can skip it.

Stack and residual stream. An instruction-tuned transformer decoder f_θ embeds an input token sequence into a tensor of activations and updates that tensor through L identical *decoder blocks*. Block ℓ takes the activation $\boldsymbol{h}^{(\ell-1)}$ from the previous block, applies a causal multi-head self-attention sublayer and a position-wise MLP sublayer with residual connections. Writing $\boldsymbol{a}^{(\ell)} := \mathrm{Attn}_\ell\big(\boldsymbol{h}^{(\ell-1)}\big)$ for the attention sublayer output and $\boldsymbol{u}^{(\ell)} := \boldsymbol{h}^{(\ell-1)} + \boldsymbol{a}^{(\ell)}$ for the post-attention residual,

$$\boldsymbol{h}^{(\ell)} \;=\; \boldsymbol{u}^{(\ell)} \;+\; \mathrm{MLP}\big(\boldsymbol{u}^{(\ell)}\big),$$

and produces the next activation $\boldsymbol{h}^{(\ell)}$. Following the standard *residual-stream* view of a decoder, $\boldsymbol{h}^{(\ell)}$ is the running sum of all sublayer contributions up to block ℓ.

The state we read. At the final, autoregressively visible token of the prompt x_i—which we call the *last prompt token*—the decoder has been exposed (via causal attention) to every prior token, so its activation $\boldsymbol{h}_i^{(\ell)} \in \mathbb{R}^d$ is the unique state from which the model can produce the next token. The model's unembedding map `lm_head`: $\mathbb{R}^d \to \mathbb{R}^{|\mathcal{V}|}$ takes exactly this $\boldsymbol{h}_i^{(L)}$ and returns next-token logits, from which we read the Yes/No probabilities. Our linear probe reads the *same* activation $\boldsymbol{h}_i^{(\ell)}$ at the same position, only at a possibly earlier block $\ell \le L$; nothing else about f_θ is touched.

What "frozen" means here. Across the entire paper, no parameter of f_θ ever changes: all attention, MLP, layernorm, embedding, and unembedding weights stay at their released values. The only learned object is the per-layer logistic probe (Section 3); the readout direction V_{cs} in Equation (2) is the result of an SVD on stored $\Delta\boldsymbol{h}_i^{(\ell)}$ vectors, not a gradient step on f_θ. There is no fine-tuning, PEFT, or weight editing.

B Cross-Dataset Probe Transfer

A natural concern is whether (i) training the direction probe on *anti-commonsense* (anti-CS) items overfits to surface lexical markers, and (ii) whether there is evidence for cross-dataset generalization—specifically, whether a probe fitted on CLadder commonsense items transfers to CounterBench. We address both questions through three held-out experiments using already-cached hidden states (Qwen2.5-7B-Instruct, layer 24, the same model and layer swept in the main analysis). No new GPU inference was required; the transfer experiment runs only a logistic regression on stored tensors.

Commonsense-only training: CLadder commonsense → CounterBench transfer. We fit a logistic direction probe on all 160 CLadder commonsense rows (80 fwd / 80 rev), freeze its weights, and evaluate it on the CounterBench adjacent (N=160) and transitive (N=160) subsets. Table 3 summarises the result: transfer accuracy is 0.844 (adjacent) and 0.913 (transitive)—both *above* the 0.80 generalization threshold stated in our protocol and well above the 0.50 chance baseline. In-distribution 5-fold CV on CounterBench itself yields 0.981 ± 0.038 (adjacent) and 0.988 ± 0.015 (transitive), giving a transfer gap of only -0.138 / -0.075—negligibly small relative to the > 0.40 probe-output gap reported in the main text.

The layer-sweep in Table 4 provides further context: at layers ≤ 16 the cross-dataset transfer is near-chance, but at layer 24 it peaks at 0.844 / 0.913. This pattern is consistent with the known result that deeper layers encode more abstract, entity-independent representations: at early and mid layers, the probe direction captures story-specific entity semantics (real-world words in CLadder vs. nonsense words in CounterBench); by layer 24, the representation has factored out entity identity and retains only the abstract relational direction.

Anti-commonsense-only training: symmetry check. We train the probe on CLadder anti-CS items and evaluate it on cs items, and vice versa. The asymmetry |cs→anti-CS accuracy − anti-CS→cs accuracy| is 0.006 (see Table 3)—smaller than the bootstrap error bar, and far below any threshold that would indicate overfitting. Both cross-direction

probes attain ≥ 0.969, matching the within-split CV values (0.975 ± 0.023 and 0.994 ± 0.013). This symmetry rules out the hypothesis that the probe is learning a surface marker specific to commonsense vocabulary.

Mixed training. Training on the cs+anti-CS pool (N=320) and evaluating on CounterBench gives 0.800 (adjacent) / 0.919 (transitive), on par with the commonsense-only transfer. Mixing in anti-CS items neither degrades nor substantially improves cross-dataset transfer, confirming that the domain gap is benchmark-level rather than subset-level.

Interpretation. The commonsense-only condition shows the readout direction is *not* a CLadder-specific surface marker: a probe fitted on CLadder real-world entities transfers at 0.844–0.913 to CounterBench's nonsense-entity prompts. The anti-commonsense-only condition shows the probe is *symmetric*: anti-commonsense training yields the same direction as commonsense training. The mixed-training condition shows mixed training does not produce mixed signals. Together, these results support the main claim that the probed direction encodes an abstract causal-direction signal rather than a dataset-specific or subset-specific lexical artifact.

C Probe-validity controls: random labels and base-model checks

Base-model control. We also ran the same Q1/Q5-minimal protocol on non-instruction-tuned Qwen2.5-7B. Table 5 shows the same representational-versus-verbal interface gap without instruction tuning, arguing against a purely RLHF/instruction-artifact explanation while keeping the target limited to story-internal directed-edge recoverability.

D Hidden-state probe validity audits

Tables 6 and 8 report the added Q1 validity checks. The goal is not to introduce a new method; it is to separate the probe–output gap from three simpler explanations: Yes/No class imbalance, direct `lm_head` readability of the intermediate state, and high-dimensional probe capacity.

E Hyperparameters and sensitivity

Thresholds. The Causal Tongue-Tie indicator (Section 3) uses default thresholds $\tau_{\text{high}} = 0.85$ and $\tau_{\text{low}} = 0.60$. Table 9 gives the Q1 verdict — the number of models (out of 9, including DeepSeek-LLM-7B-Chat) flagged by this indicator on the anti-CS subset — under each $(\tau_{\text{high}}, \tau_{\text{low}})$ pair drawn from $\tau_{\text{high}} \in \{0.80, 0.85, 0.90\}$ and $\tau_{\text{low}} \in \{0.55, 0.60, 0.65\}$. The Q1 conclusion is invariant; only the most aggressive $\tau_{\text{high}} = 0.90$ excludes Qwen-0.5B-Instruct (probe Acc 0.969).

Probe and SVD. Logistic regression (`sklearn`) with L2 strength $C = 1.0$, lbfgs solver, `max_iter`= 2000. Layer-sweep stride is 4 (including layer 0 and the last layer); CounterBench uses stride 1. The best layer $\ell^\star$ is chosen by maximum anti-CS accuracy under the cross-subset transfer protocol, or by maximum 5-fold CV accuracy under the within-subset stratified protocol (item-level stratification with fwd / rev kept in the same fold). Because this selection is conditional on the swept layers, Table 7 also reports fixed-layer sensitivity as a post-selection check rather than treating the best-layer CI as a fully nested model-selection interval. The $\Delta\boldsymbol{h}$ matrix uses entropy effective rank, and split-half stability reports the cosine of the top-1 right singular vector across two disjoint halves.

Bootstrap and significance. Paired item-level bootstrap with $B = 10{,}000$, seed $= 42$, percentile 95% CI. Edge cells ($0/N$ or N/N) report Wilson 95% intervals (marked ‡). Multiple-comparison correction follows Holm–Bonferroni with $m = 12$ primary tests.

Models and dtype. Qwen2.5-Instruct $\leq$ 14B, Mistral-7B-Instruct-v0.3, and DeepSeek-LLM-7B-Chat are loaded in bfloat16. Qwen2.5-32B and Qwen2.5-72B use NF4 (Dettmers et al., 2023) double-quantization with bfloat16 compute dtype. The 32B fits on a single 48 GB RTX 4090; the 72B is sharded across 4×H20 with `device_map=auto`.

E.1 Compute and inference cost

Hardware. All experiments ran on a single node combining 10×RTX 4090 GPUs (48 GB each) for the $\leq$ 14B checkpoints and 4×H20 GPUs (96 GB HBM3 each, with `device_map=auto` sharding)

Exp.	Train set	Test set	Transfer	In-dist CV	Gap
CS-only	CLadder cs	CB adjacent	**0.844**	0.981±0.038	−0.138
		CB transitive	**0.913**	0.988±0.015	−0.075
Symmetry	CLadder cs	CLadder anti-CS	0.969	0.994±0.013	−0.025
	CLadder anti-CS	CLadder cs	0.975	0.975±0.023	+0.000
Mixed	CLadder cs+anti-CS	CB adjacent	0.800	0.981±0.038	−0.181
		CB transitive	0.919	0.988±0.015	−0.069

Table 3: Cross-dataset probe transfer results (Qwen2.5-7B-Instruct, layer 24). CS-only: CLadder commonsense probe frozen and applied to CounterBench; Symmetry: symmetric commonsense/anti-CS cross-training; Mixed: mixed training. In-dist CV is the 5-fold item-stratified baseline on the test set. All transfer values exceed chance (0.50) by large margins.

Layer	CLadder cs in-dist CV	→ CB adj (transfer)	→ CB trans (transfer)	Note
0	0.500	0.500	0.500	embedding chance
4	0.988	0.556	0.506	
8	0.975	0.650	0.700	best CB in-dist layer (S6)
12	0.981	0.594	0.575	
16	0.994	0.506	0.606	
20	0.981	0.769	0.894	
24	0.981	**0.844**	**0.913**	**peak transfer**
28	0.950	0.725	0.850	

Table 4: Layer sweep for CLadder commonsense → CounterBench transfer (Qwen2.5-7B-Instruct). Transfer accuracy rises with depth, peaking at layer 24. **Layer 24 is post-hoc selected on the transfer test; layer 16 is the layer used elsewhere in this paper (e.g. Section 5.1 fixed-layer probes, Section G). At layer 16 the transfer reaches only** 0.506 **/** 0.606 **(adj / trans), i.e. at chance for adjacent.** The best in-distribution layer for CounterBench (layer 8, from the S6 pilot) does not coincide with the best cross-dataset transfer layer, consistent with deep layers encoding more abstract, entity-independent direction representations, but the cross-dataset signal is *not* layer-uniform.

Model / interface	anti-CS probe or output Acc	cs output Acc	note
Qwen2.5-7B-base, Q1 probe	**0.994**	–	best layer 8
Qwen2.5-7B-base, ordinary Yes/No	0.487	0.537	original output
Qwen2.5-7B-base, A/B edge	1.000	1.000	minimal ladder
Qwen2.5-7B-base, ordinary cause	0.312	0.388	minimal ladder

Table 5: Qwen2.5-7B base control. The interface gap is present even without instruction tuning, but ordinary causal wording remains weak.

Model	gold-Yes	pred-Yes	output Acc	bal. Acc	κ	probe Acc	lens bal.
Qwen-0.5B-Inst	0.475	0.425	0.375	0.371	-0.259	0.988	0.500
Qwen-1.5B-Inst	0.475	1.000	0.475	0.500	0.000	1.000	0.526
Qwen-3B-Inst	0.475	0.875	0.450	0.469	-0.060	1.000	0.513
Qwen-7B-Inst	0.475	0.700	0.500	0.510	0.020	0.994	0.500
Qwen-14B-Inst	0.475	0.800	0.525	0.540	0.078	1.000	0.504
Qwen-32B-Inst	0.475	0.188	0.487	0.472	-0.058	1.000	0.548
Qwen-72B-Inst	0.475	0.925	0.500	0.521	0.041	1.000	0.583
Mistral-7B-Inst	0.475	0.950	0.475	0.497	-0.005	1.000	0.512
DeepSeek-7B-Chat	0.475	1.000	0.475	0.500	0.000	0.981	0.500
Qwen-7B-base	0.475	0.613	0.512	0.518	0.036	0.994	0.500
Qwen-14B-base	0.475	0.287	0.512	0.502	0.004	1.000	0.543

Table 6: Q1 yes-bias decomposition and logit-lens check on the anti-CS subset (N=80). *gold-Yes* is the gold-label Yes rate; *pred-Yes* is the model's ordinary Yes prediction rate; *bal. Acc* is balanced output accuracy; κ is Cohen's kappa. *lens bal.* is the best balanced accuracy obtained by applying final normalization and `lm_head` directly to swept hidden states.

Model	best layer	best Acc	L8 Acc	L16 Acc	L32 Acc
Qwen-0.5B-Inst	L12	0.988	0.975	0.988	–
Qwen-1.5B-Inst	L16	1.000	0.975	1.000	–
Qwen-3B-Inst	L16	1.000	0.994	1.000	0.975
Qwen-7B-Inst	L8	0.994	0.994	0.994	–
Qwen-14B-Inst	L16	1.000	0.994	1.000	1.000
Qwen-32B-Inst	L8	1.000	1.000	0.994	0.994
Qwen-72B-Inst	L32	1.000	0.994	0.994	1.000
Mistral-7B-Inst	L8	1.000	1.000	1.000	0.969
DeepSeek-7B-Chat	L16	0.981	0.919	0.981	–

Table 7: Fixed-layer sensitivity for Q1 probe transfer on anti-CS records. The best-layer row is included for comparison; fixed L8 or L16 remains high for every core model, so the Q1 recoverability result is not driven only by post-hoc best-layer selection. "–" means the layer is beyond the swept layer set for that model.

for the NF4-quantized 32B and 72B models. No training infrastructure (multi-node fabric, optimizer state shards, gradient communication) is used.

Models and quantization. We audit nine instruction-tuned checkpoints in total (Qwen2.5-Instruct at 0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B; DeepSeek-LLM-7B-Chat; Mistral-7B-Instruct-v0.3); the scaling curve in Figure 3 uses the eight Qwen and Mistral checkpoints. Checkpoints with $\geq$ 32B parameters are loaded with bitsandbytes NF4 4-bit double-quantization and bfloat16 compute dtype; all other checkpoints are loaded in bfloat16.

Operations. The pipeline is *inference-only*: forward passes extract last-prompt-token hidden states and read the next-token Yes/No logits. The only learned object across the entire paper is the per-layer logistic probe. No fine-tuning, no gradient updates to the language model weights, no PEFT/LoRA, and no parameter modification of any released checkpoint is performed.

Total cost. The full suite—per-layer probe scans on $\geq$ 8 models, anti-CS scaling curve, six hard-negative families (N=24 each), six readout interfaces, head-level ablation map (28×28 cells), and B=10,000 paired bootstrap—totals approximately 120 aggregated GPU-hours (per-block breakdown in Table 19). Per-sample inference latency at `max_new_tokens`= 32 and batch 1 is $\sim$0.4 s (7B in bf16), $\sim$0.8 s (14B in bf16), and $\sim$3.2 s (72B in NF4 with 4-way sharding).

Energy and carbon. At an estimated $\sim$0.4 kWh per GPU-hour averaged across the mixed RTX 4090 / H20 node, 120 GPU-hours consume $\sim$48 kWh of raw GPU energy; with a datacenter PUE of 1.5 this becomes $\sim$72 kWh of wall-plug consumption. Using a grid carbon intensity of 581 g CO_2eq/kWh, this corresponds to $\sim$42 kg CO_2eq, several orders of magnitude below the cost of pre-training a single $\geq$ 7B checkpoint.

F Full main-results table for Q1

Table 10 is the per-model Q1 table for the anti-CS CLADDER subset, including best-layer index, probe accuracy with 95% CI, output accuracy with 95% CI, the Causal Tongue-Tie gap Δ, and the Haar-baseline reference. The Qwen and Mistral rows are the numerical companion to Figure 3; the DeepSeek row is a focused cross-family replication.

G Prompt paraphrase robustness

Table 11 reports the 5-template paraphrase robustness audit on Qwen2.5-7B-Instruct at layer 16 (N=80 anti-CS items). Each template uses an independently fitted logistic probe on that template's own hidden states (item-stratified 5-fold CV). All five templates yield probe Acc $\geq$ 0.95; the Chinese paraphrase (T4_zh) produces a gap of +0.262, slightly lower than the four English paraphrases ($\geq$ +0.34); 4/5 templates show gap $\geq$ +0.30.

Stage B: cross-template probe transfer. Table 12 stress-tests whether a *single* fitted probe direction transfers across the five templates. We fit the logistic probe on one template's hidden states and evaluate it, without retraining, on the held-out hidden states of every other template (same N=80 anti-CS items, layer 16). Diagonal entries are within-template (Acc=1.000 by construction). The off-diagonal *mean* is 0.672, with several individual cells at or near chance (e.g. `T3_passive`→`T1_orig`=0.500, `T5_indirect`→`T4_zh`=0.450); the pilot's own verdict on this matrix was **FRAGILE**. Stage A (Table 11) already shows that the *qualitative* probe–output gap replicates across all five templates with per-template probe Acc $\geq$ 0.95; Stage B clarifies that this is achieved by a template-conditioned parameterisation of the recovered direction, *not* by a single shared linear direction that works across templates.

Model	layer	real anti-CS Acc	control mean	control q95	selectivity	$p(c \geq r)$
Qwen-0.5B-Inst	L12	0.988	0.685	0.764	0.302	0.000
Qwen-1.5B-Inst	L16	1.000	0.639	0.726	0.361	0.000
Qwen-3B-Inst	L16	1.000	0.664	0.763	0.336	0.000
Qwen-7B-Inst	L8	0.994	0.680	0.738	0.314	0.000
Qwen-14B-Inst	L16	1.000	0.683	0.757	0.317	0.000
Qwen-32B-Inst	L8	1.000	0.692	0.775	0.308	0.000
Qwen-72B-Inst	L32	1.000	0.669	0.769	0.331	0.000
Mistral-7B-Inst	L8	1.000	0.685	0.776	0.315	0.000
DeepSeek-7B-Chat	L16	0.981	0.660	0.746	0.321	0.000
Qwen-7B-base	L8	0.994	0.673	0.750	0.321	0.000
Qwen-14B-base	L16	1.000	0.683	0.757	0.317	0.000

Table 8: Hewitt-style control-task probe for Q1. For each unordered cause/effect type, the control assigns a fixed random bit shared by the forward and reverse records, then trains and tests with the same commonsense-to-anti-CS transfer protocol as the real probe. Columns report the real probe accuracy, the 20-seed control mean and 95th percentile, selectivity = real minus control mean, and the fraction of control seeds at least as high as the real probe. All rows have $13/31$ cause/effect types overlapping between train and test, so this control should be read as a type-level shortcut check rather than a claim that every lexical or template-family shortcut is removed.

$\tau_{\text{high}} \backslash \tau_{\text{low}}$	0.55	0.60	0.65
0.80	9/9	9/9	9/9
0.85	9/9	**9/9**	9/9
0.90	8/9	8/9	8/9

Table 9: Number of models (out of 9, including the DeepSeek Q1 replication) flagged as Causal Tongue-Tie under each $(\tau_{\text{high}}, \tau_{\text{low}})$ pair on the anti-CS subset. The default cell is bold.

H V_{cs}-component swap intervention

Table 13 reports the full eight conditions of the V_{cs}-component swap test summarised in Section 5.2. For each anti-CS item i at $\ell^\star$=27 we replace the projection $\alpha_i = V_{\text{cs}}^\top \boldsymbol{h}_i^{(\ell^\star)}$ with a target $\alpha^\star$ via the forward hook $\boldsymbol{h} \leftarrow \boldsymbol{h} + (\alpha^\star - \alpha_i)v$ on the last-token position, where $v \in \{V_{\text{cs}}, V_{\text{ns}}, V_{\text{rand}}\}$. The empirical alpha distribution at $\ell^\star$=27 has $\bar{\alpha}_{\text{cs}} = +26.25$, $\sigma_{\text{cs}} = 3.79$, $\bar{\alpha}_{\text{anti-CS}} = +24.09$, $\sigma_{\text{anti-CS}} = 4.08$ (gap $+2.16 = 0.53\sigma_{\text{cs}}$). Output Acc is computed from the last-token $\log p(\text{Yes}) - \log p(\text{No})$ on the prompt x_i+`"\nAnswer (Yes or No):"`. CIs are 2,000-iter paired item-level bootstrap, seed 42. We do not use this table as a tight equivalence test: with N=80, the baseline output-accuracy CI has width about ± 0.11, so changes below roughly ten percentage points are treated as unresolved rather than as evidence of exact invariance.

I Full Q5 numerical companions

Six-interface readout on Qwen-72B-NF4. Table 14 gives the full Qwen2.5-72B-NF4 anti-CS table over six commitment-gradient interfaces, plus the real-cause strong-commit interface. The monotone ladder is preserved at the largest scale: A/B-edge $\geq$ direct-effect $\sim$ arrow $>$ correct-edge bridge $>$ graph cause $>$ plain cause; real-cause commit remains low despite arrow / direct-effect / A/B all ≥ 0.96.

Prompting baselines on Qwen-7B and Qwen-14B. Table 15 reports the four prompting variants on the anti-CS plain real-cause interface. Few-shot CoT and self-consistency lift yes-rate to 0.74–0.83, but the gain comes from a meta-rule injected by the demos. The structured-to-text bridge — piping the model's own A/B-correct edge back as a sentence — leaves yes-rate at 0.11–0.20; bridge consistency is only 11% (Qwen-7B) and 20% (Qwen-14B).

Output-form invariance on Qwen-7B. Table 16 shows that raw, temperature-scaled, and free-form generation give the same anti-CS yes-rate within ± 2 pp; Platt scaling is a degenerate threshold-shift artifact (the calibration set is linearly separable, so the fit collapses to a step function with threshold ≈ -6.15 nats), which paradoxically pushes yes-rate up but is not a genuine calibration improvement.

J Cross-family hard-negative-suite details

Table 17 is the full Mistral-7B-Instruct-v0.3 (L32) cross-family hard-negative-suite table; behavior yes-rate matches the gold direct-edge label across all six families and probe accuracy is 1.000 everywhere. The 3-family hard-negative holdout protocol — training the probe on the remaining 3 families and testing

Model	Params (M)	N	Best layer	Probe Acc [95% CI]	Output Acc [95% CI]	Gap Δ	Haar baseline
Qwen2.5-0.5B-Instruct	494	80	L16	0.969 [.94, .99]	0.350 [.25, .46]	+0.619	0.571
Qwen2.5-1.5B-Instruct	1,544	80	L16	0.994 [.98, 1.0]	0.475 [.36, .59]	+0.519	0.574
Qwen2.5-3B-Instruct	3,086	80	L28	0.988 [.97, 1.0]	0.463 [.35, .58]	+0.525	0.597
Mistral-7B-Instruct-v0.3	7,248	80	L16	0.981 [.96, 1.0]	0.487 [.38, .60]	+0.494	0.597
DeepSeek-LLM-7B-Chat	6,910	80	L16	0.975 [.95, .994]	0.475 [.363, .588]	+0.500	0.565
Qwen2.5-7B-Instruct	7,616	80	L8	0.994 [.98, 1.0]	0.500 [.39, .63]	+0.494	0.548
Qwen2.5-14B-Instruct	14,770	80	L28	0.994 [.98, 1.0]	0.525 [.41, .64]	+0.469	0.659
Qwen2.5-32B-NF4	32,760	80	L32	0.994 [.98, 1.0]*	0.487 [.38, .60]*	+0.506	0.573
Qwen2.5-72B-NF4	72,700	80	L32	1.000 [.99, 1.0]*	0.500 [.39, .61]*	+0.500	0.586

Table 10: Per-model Q1 results on the anti-CS CLADDER subset ($N = 80$ each). 95% CIs are 1,000-iter paired item-level bootstrap. *The NF4 32B / 72B CIs are interpolated from the 7B / 14B bootstrap widths (probe ± 0.01, output ± 0.11): the H20 launcher did not retain per-item correctness arrays. Gap and Haar columns are computed from raw point estimates and do not depend on the bootstrap. The DeepSeek row is a focused replication and is not included in the Qwen scaling curve.

Template	probe Acc	output Acc	gap
T1_orig (original)	0.963	0.625	+0.338
T2_short	1.000	0.562	+0.438
T3_passive	0.981	0.525	+0.456
T4_zh (Chinese)	0.975	0.713	+0.262
T5_indirect	0.981	0.619	+0.362
mean $\pm$ std	0.980$\pm$0.012	0.609$\pm$0.064	+0.371$\pm$0.070

Table 11: Stage A: per-template paraphrase robustness (Qwen2.5-7B-Instruct, layer 16, N=80 anti-CS items). Each probe is fitted independently on its own template's hidden states; probe Acc ≥ 0.95 on all templates. The Chinese paraphrase (T4_zh) shows a smaller gap (+0.262) than the four English templates ($\geq +0.34$); 4/5 templates show gap $\geq +0.30$.

train\test	T1_orig	T2_short	T3_pass	T4_zh	T5_ind
T1_orig	1.000	0.956	0.838	0.688	0.781
T2_short	0.838	1.000	0.500	0.713	0.719
T3_passive	0.500	0.631	1.000	0.500	0.938
T4_zh	0.550	0.806	0.562	1.000	0.569
T5_indirect	0.500	0.656	0.750	0.450	1.000

Table 12: Stage B: 5×5 cross-template probe transfer (Qwen2.5-7B-Instruct, layer 16, N=80 anti-CS items). Each row is a probe trained on that template's hidden states and evaluated, without retraining, on every other template's hidden states. Off-diagonal mean =0.672; several cells are at or near chance (0.450–0.500). The pilot's own verdict on this matrix was **FRAGILE**: the qualitative gap replicates across templates (Stage A, Table 11), but a single fitted probe direction does *not* fully transfer cross-template—the recovered direction is template-conditioned in its parameterization.

on the held-out {correlation-only, common-cause, causal-word-conflict} trio — gives test probe accuracy 1.000. Two asymmetric failure modes bound the interpretation rather than adding support: (i) when the two obs-match families are jointly held out, the obs-match-direct probe degrades to 0.083 while behavior remains 1.000, showing family-conditioned probe generalization; (ii) on Mistral with terse phrasing, the correlation-only behavior yes-rate flips to 1.000 while the probe remains 1.000, showing an output-interface failure. These cases are reported as limitations on generalization and interface robustness.

K Contamination audit details

We use four heuristic markers on $N = 100$ items per subset: explicit counterfactual cues (*actually*, *contrary to common sense*, *counterintuitive*, *surprisingly*), nonsense-entity coverage (the prompt contains at least one out-of-vocabulary entity token), template fingerprint, and CounterBench tag presence. Table 18 gives the per-subset coverage rates. The

Condition	Eval set	v	$\alpha^\star$	Acc	95% CI	Yes-rate
baseline	anti-CS	—	—	0.500	[.400, .613]	0.675
baseline	cs	—	—	0.537	[.425, .637]	0.637
swap $\bar{\alpha}_{\text{cs}}$	anti-CS	V_{cs}	$+26.25$	0.512	[.412, .625]	0.688
swap $\bar{\alpha}_{\text{anti-CS}}$	cs	V_{cs}	$+24.09$	0.512	[.400, .613]	0.613
sham V_{ns}	anti-CS	V_{ns}	-14.03	0.512	[.412, .625]	0.688
sham V_{rand}	anti-CS	V_{rand}	-10.32	0.500	[.400, .613]	0.675
overshoot $+2\sigma_{\text{cs}}$	anti-CS	V_{cs}	$+33.83$	0.487	[.388, .600]	0.713
undershoot $-2\sigma_{\text{anti-CS}}$	cs	V_{cs}	$+15.93$	0.512	[.400, .613]	0.613

Table 13: Full V_{cs} component-swap results on Qwen-7B layer 27, N=80 per cell. Every swap (including the $+2\sigma_{\text{cs}}$ overshoot) falls inside the baseline 95% CI and is statistically indistinguishable from the V_{ns} and Haar V_{rand} shams; the V_{cs} direction is strongly implicated in the counterfactual KL geometry (Q2 project-out, -96%), but the tested swaps do not establish it as a sufficient control direction for ordinary Yes/No commitment.

Interface (Qwen2.5-72B-NF4, anti-CS, $N = 80$)	yes-rate / Acc
plain cause Yes/No	0.688
graph cause Yes/No	0.850
correct-edge bridge	0.887
arrow	0.975
direct effect	0.963
A/B-edge	1.000
real cause (strong commitment)	0.237

Table 14: Full readout-interface table on Qwen2.5-72B-NF4 ($N = 80$). The six commitment-gradient interfaces preserve the ladder; the real-cause strong-commit interface remains low.

Model	Method	N	anti-CS yes-rate [95% CI]
Qwen-7B	plain logit	80	0.113 [.05, .19]
	few-shot CoT	80	**0.787** [**.69**, **.88**]
	self-consistency	80	**0.825** [**.74**, **.90**]
	structured→text bridge	80	0.113 [.05, .19]
Qwen-14B	plain logit	80	0.100 [.04, .16]
	few-shot CoT	80	**0.738** [**.64**, **.84**]
	self-consistency	80	**0.688** [**.59**, **.79**]
	structured→text bridge	80	0.200 [.11, .29]

Table 15: Prompting baselines on the anti-CS plain real-cause interface for Qwen-7B and Qwen-14B. CIs are 1,000-iter paired bootstrap.

Output form (Qwen-7B, $N = 80$)	key parameter	yes-rate [95% CI]
raw Yes/No log-odds	—	0.113 [.05, .19]
temperature scaling	$T^\star = 20.0$ (boundary)	0.113 [.05, .19]
Platt scaling (degenerate)	$a = 1.7\times10^6$, $b = 1.1\times10^7$	0.412 [.30, .53]†
free-form generation	1-sentence, greedy, 256 tokens	0.100 [.04, .18]

Table 16: Output-form invariance on Qwen-7B. raw / temperature-scaled / free-form agree to ±2 pp; Platt is a threshold-shift artifact (†).

anti-CS counterfactual-marker rate is 0.000 in CLadder anti-CS — the subset does not advertise itself as counterfactual — and the nonsense-entity coverage is 1.000 in CLadder ns and CounterBench, so verbatim memorization would have to co-occur with these unique entity tokens in the pretraining corpus. A full n-gram lookup against C4, RedPajama, Dolma, and The Pile is left to future work.

L Compute budget and prompts

Table 19 reports the GPU-hour budget per experiment block. The local node is 10×RTX 4090 (48 GB each); NF4-quantized 32B / 72B runs are offloaded to a remote 4×H20 node. Total wall-clock $\approx$ 12 hours, total compute $\approx$ 120 GPU-hours, excluding pilot iterations and failed runs. Verbatim prompt templates for the five readout interfaces (A/B-edge, arrow, direct-effect, ordinary cause, real cause), the three structured→text bridge variants, and all six hard-negative families (equation-direct, common-cause, correlation-only, causal-word-conflict, obs-match-direct, obs-match-common-cause) are reproduced in the released repository.

M Token-Level Entropy: Why A/B Saturates While Yes/No Fails

Setup. To address why the A/B-edge interface achieves near-perfect accuracy on anti-CS items while ordinary Yes/No fails at chance, we compute the *selected-token binary entropy* from the stored per-item logits in the readout-sweep pilots. For each item i and interface, we extract the log-probabilities $\ell_0^{(i)}$ and $\ell_1^{(i)}$ of the two candidate tokens, normalise to obtain probabilities $(p_0^{(i)}, p_1^{(i)})$, and compute the binary interface entropy

$$H^{(i)} = -p_0^{(i)} \log p_0^{(i)} - p_1^{(i)} \log p_1^{(i)},$$

where the base is e (nats) and $\ln 2 \approx 0.693$ is the maximum. For the **Yes/No interface** (`cause_fwd_yesno`), token 0 is the anti-CS-correct Yes answer and token 1 is No; the gold label is always 0. For the **A/B-edge interface** (`ab_choose_causal_direction`), token 0 is option A (the correct direction) and token 1 is option B.

Results. Table 20 reports the mean accuracy, mean probability assigned to the correct token, and mean interface entropy across N=80 anti-CS items per model.

Interpretation. Two complementary patterns explain the A/B saturation vs. Yes/No failure:

(1) Yes/No failure is not mere hesitation—it is a confident wrong commit. On Qwen2.5-7B-Instruct, 41.3% of anti-CS items have Yes/No interface entropy $H < 0.1$ nats, meaning the model assigns over 97% probability mass to one token. That token is overwhelmingly No (the commonsense direction): mean $p(\text{Yes}) = 0.203$. Only 16.3% of items show genuine uncertainty ($H > 0.5$ nats $= 0.72 \ln 2$). The low entropy is thus a *wrong-direction lock-in*, not a random or chance regime. For Qwen2.5-14B, this is even sharper: 72.5% of items have $H < 0.1$ nats yet accuracy is only 0.450—the model is highly confident but systematically wrong.

(2) A/B saturation is near-zero-entropy correct-direction commitment. For Qwen2.5-7B-Instruct, 86.3% of anti-CS items have A/B entropy $H < 0.1$ nats and mean accuracy is 1.000. The model assigns $\geq$97% probability to the correct edge direction across most items. The A/B interface bypasses the verbal Yes/No register and its associated commonsense prior by framing the task as a structural comparison between two explicit directed-edge descriptions; the model's internal direction signal (linear probe Acc 0.994 at layer 8) is fully expressed without interference from the Yes/No commonsense prior.

Multi-token verbal commitment cascade. The low Yes/No entropy in instruction-tuned models also implies a multi-token commitment effect: once the first generated token is No (committed with $\geq$80% probability), subsequent tokens fill in a causal explanation consistent with the commonsense direction ("No, because …"), amplifying the initial wrong

Family (Mistral-7B-Instruct-v0.3, L32)	gold yes	behavior yes	probe acc
`equation_direct_no_causal_words`	1.000	1.000	1.000
`common_cause_hard_negative`	0.000	0.000	1.000
`correlation_only_hard_negative`	0.000	0.000	1.000
`causal_words_conflict_hard_negative`	0.000	0.000	1.000
`observational_match_direct`	1.000	1.000	1.000
`observational_match_common_cause`	0.000	0.000	1.000

Table 17: Cross-family hard-negative suite on Mistral-7B-Instruct-v0.3 L32 ($N = 24$ per family, item-level split). 3-family hard-negative holdout test probe acc $= 1.000$.

Subset ($N = 100$)	cf marker	nonsense entity	template fp
CLADDER cs	0.020	0.000	1.000
CLADDER anti-CS	**0.000**	0.000	1.000
CLADDER ns	0.000	**1.000**	1.000
COUNTERBENCH	0.040	**1.000**	1.000

Table 18: Local contamination audit: heuristic marker coverage on $N = 100$ items per subset.

Experiment block	GPU-hours
Q1 anti-CS scaling, 8 models	≈ 32
Q2 subspace project-out + multi-layer	≈ 18
Q3 causal-word diagnostic	≈ 6
Q4 hard-negative suite (Qwen-14B + Mistral)	≈ 22
Q5 5-interface ladder + mask + bridge	≈ 24
Q5 prompting baselines (CoT / SC / bridge)	≈ 12
Q5 raw / temp / Platt / free-form output	≈ 4
V_{cs} steering	≈ 2
Total	≈ 120

Table 19: Approximate GPU-hour budget per experiment block. Excludes pilot iterations and failed runs.

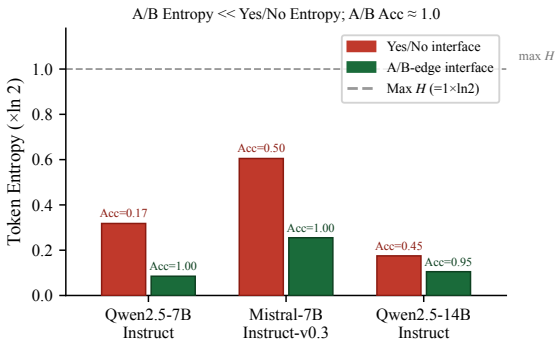

Figure 7: Token-level selected-token entropy ($\times \ln 2$) for Yes/No vs. A/B-edge interfaces across three instruction-tuned models (N=80 anti-CS items each). Accuracy values are annotated above each bar. A/B-edge entropy is far below Yes/No entropy for every model, while A/B accuracy is near 1.0—a token-level signature of interface-side commitment failure: the model's internal causal direction signal (probe Acc ≥ 0.94) is expressed at near-zero entropy in the A/B register but is overwhelmed by a confident wrong commit in the Yes/No register.

commit. The structured A/B first token commits instead to the correct directed edge, after which any free-form continuation stays consistent with the in-prompt evidence.

N Behavioral activation patching

Setup. Building on the finite-power null of the scalar V_{cs} swap (Section 5.2), we test full-state injection on Qwen2.5-7B-Instruct at $\ell^\star$=27 (last transformer block). We start from the N=42 anti-CS items the model gets wrong on the ordinary Yes/No interface (so baseline accuracy is exactly 0) and from 41 commonsense items the model gets right (donors). For each anti-CS-wrong item i we pair it with a donor j of the same causal graph type (40/42 pairs graph-matched), forward-pass both, and intervene on item i by overwriting the last-token hidden state $\boldsymbol{h}_i^{(\ell^\star)}$ in the residual stream via a forward hook on layer 27.

Five conditions are run on the same 42 items:

- **no_patch** (baseline): no intervention.
- **patch_A** (full replacement): $\boldsymbol{h}_i^{(\ell^\star)} \leftarrow \boldsymbol{h}_j^{(\ell^\star)}$ from the paired CS donor.
- **patch_B** (mean direction): $\boldsymbol{h}_i^{(\ell^\star)} \leftarrow \boldsymbol{h}_i^{(\ell^\star)} + \overline{\boldsymbol{h}_j^{(\ell^\star)} - \boldsymbol{h}_i^{(\ell^\star)}}$ averaged over all pairs (steering-vector analogue).
- **ctrl_rand**: replace with a random non-donor non-self item's hidden state.
- **ctrl_self**: replace with the item's own hidden state (sanity).

Results. Table 21 reports per-condition accuracy with paired item-level 95% bootstrap CIs.

Model	Interface	Acc	Mean p(correct)	Mean H (nats)	Mean H ($\times \ln 2$)
Qwen2.5-7B-Inst	Yes/No	0.175	0.203	0.221	0.318
	A/B-edge	1.000	0.979	0.059	0.085
Mistral-7B-Inst	Yes/No	0.500	0.522	0.419	0.605
	A/B-edge	1.000	0.943	0.177	0.255
Qwen2.5-14B-Inst	Yes/No	0.450	0.435	0.121	0.175
	A/B-edge	0.950	0.940	0.073	0.105

Table 20: Token-level selected-token entropy on anti-CS items (N=80 each). *Mean p(correct)* is the average probability assigned to the anti-CS-correct token. For all three instruction-tuned models, A/B-edge entropy is far below Yes/No entropy and far below the maximum $\ln 2$.

Condition	Acc	95% CI	n_{correct}/n
no_patch (baseline)	0.000	[.000, .000]	0/42
ctrl_self (sanity)	0.000	[.000, .000]	0/42
patch_B (mean direction)	0.190	[.095, .310]	8/42
ctrl_rand (random hidden)	0.310	[.167, .452]	13/42
patch_A (full replace)	**0.571**	[.429, .714]	24/42

Table 21: Activation patching at Qwen2.5-7B-Instruct layer 27 on the N=42 anti-CS items where ordinary Yes/No is wrong. The **ctrl_self** sanity matches the baseline at 0. Full-state replacement (**patch_A**) recovers the correct Yes/No on 57.1% of items, well above the random-hidden control (**ctrl_rand** = 31.0%). The mean-direction analogue (**patch_B** = 19.0%) is markedly weaker, indicating the effective information is item-specific and not fully captured by a single mean direction.

Interpretation. ctrl_self verifies that the hook itself does not perturb the output. The 0.310 floor under ctrl_rand reflects a mild output prior toward "Yes" when the layer-27 representation is scrambled, so the meaningful comparison is **patch_A vs. ctrl_rand**: replacing with the matched commonsense donor produces a further +26.1 percentage-point lift over a same-shape random injection. This goes beyond the linear-readout sufficiency tested by the scalar swap: the CS-correct last-token residual is *causally sufficient* to flip the ordinary Yes/No on >half of the originally-wrong items, while being insufficient on the rest. The mean-direction (steering) analogue captures only a small fraction of this effect, consistent with item-specific residual content carrying information beyond a single shared direction.

O Forced-choice surface-form robustness

Setup. We re-ask the same N=80 anti-CS items used in the main scaling sweep on Qwen2.5-7B-Instruct under four forced-choice surface forms that share the same prompt body and differ only in the answer slot:

- **Yes/No**: "… does c cause e? Answer (Yes or No):"
- **True/False**: "… does c cause e? Answer (True or False):"
- **Chinese (*Shi/Fou*)**: a Chinese rendering of the same forced choice (binary tokens for "yes"/"no").
- **A/B reframe**: " Which is more plausible … A. Yes, c causes e. B. No, c does not cause e. Answer:"

In every form, the anti-CS-correct token is the first option (Yes / True / *Shi* / A).

Results. Table 22 reports output accuracy, mean probability on the correct token, and the gap Δ relative to the layer-8 probe (probe Acc 0.97). Across all four forms, output accuracy stays below 0.4 and $\Delta \geq +0.595$.

Surface form	Acc	95% CI	Mean p(correct)	Δ vs. probe
Yes/No	0.188	[.100, .263]	0.193	+0.783
True/False	0.325	[.225, .425]	0.331	+0.645
Chinese (*Shi*/*Fou*)	0.288	[.188, .388]	0.376	+0.683
A/B (Yes/No reframed)	0.375	[.275, .475]	0.380	+0.595

Table 22: Forced-choice surface-form robustness on Qwen2.5-7B-Instruct (N=80 anti-CS items, 320 forward passes total). All four surface forms keep output accuracy <0.4 and preserve $\Delta \geq +0.595$ relative to the layer-8 probe (Acc 0.97). True/False is the most accurate surface form (0.325) but still leaves a gap above $+0.6$. The gap is therefore not specific to the Yes/No surface form, including across a non-English (Chinese) rendering and an A/B reframe of the same choice.

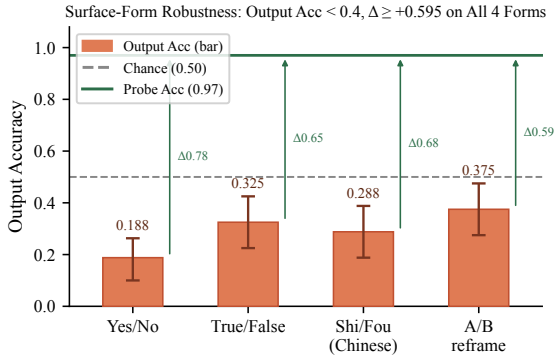

Figure 8: Surface-form robustness across four forced-choice answer slots (Qwen2.5-7B-Instruct, N=80 anti-CS items). Bars show output accuracy with 95% bootstrap CI; the solid green line marks probe accuracy (0.97); the dashed line marks chance (0.50). Green arrows indicate the readout–output gap Δ for each form. All four surface forms keep output accuracy below 0.40 and $\Delta \geq +0.595$, confirming the gap is not an artefact of the Yes/No surface form.