MANY WRITERS, ONE READER: A CAUSAL HANDLE ON RULE POLARITY IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that language models represent whether a rule permits or prohibits an action in a form that causally controls the verdict. We test this across four instruction-tuned models (Llama, Qwen, Mistral, Gemma) in a controlled anti-structuring task based on the Bank Secrecy Act. At a late decision layer, a single direction separates otherwise matched rules that permit or prohibit the same action. Removing this coordinate sharply degrades the decision, while repeatedly replacing the same coordinate through depth recovers 73–93% of the opposite decision. The computation that produces this coordinate is distributed: many late-layer components write this coordinate, while a single direction provides a compact causal readout. This effect is rule-selective. Across matched control tasks, it cannot be explained by either a generic binary-decision direction or the SATISFIED/EVADED answer tokens alone. The model's own computation also moves this coordinate: injecting a false premise and the conclusion it implies into a genuine violation shifts the internal state toward the same coordinate occupied by genuine permission. Finally, we test the same causal signature in FAA drone-altitude rules. It reproduces strongly in Llama and Gemma and more weakly in Qwen; Mistral does not form a sufficiently strong FAA decision signal for the downstream causal test.

1 INTRODUCTION

Consider a rule a compliance system might be asked to apply: file a report for any cash transaction above $10,000, and do not split a transaction into smaller deposits to stay beneath that line. Applying it takes two things, working out what the rule does and working out whether its condition holds, and behavioral accuracy cannot tell us whether a model does both. A model that answers most cases correctly might be evaluating the condition, or might be recognizing a familiar situation and reading off an answer it already associates with it. Which of those is happening matters because they are different computational failures. This paper isolates the first part: whether the model's verdict causally uses its representation of what the stated rule permits or prohibits.

We study semantic rule polarity: whether a stated rule permits or prohibits an action, with the scenario held fixed. We extract a permit-minus-prohibit direction v_r by difference-of-means and test it causally, first validating the pipeline by replicating the known refusal direction (App. A.1). We establish the mechanism on anti-structuring regulation under the US Bank Secrecy Act (BSA), then test whether its causal signature transfers to US Federal Aviation Administration (FAA) drone-altitude rules.

Across four instruction-tuned families, polarity becomes causally effective on the decision only in a late band. Ablate the single direction v_r and the permit/prohibit decision degrades sharply; overwrite that coordinate through the band and it recovers most of the opposite decision, even though many components upstream write it (§4). Inject into a genuine violation a false premise and the conclusion it implies, and the internal state moves onto the coordinate real permission occupies (§5).

The key test of what v_r represents is specificity (§6): we build the reference direction each simpler explanation predicts—a generic yes-or-no axis and an answer-word channel—and v_r's causal effect follows rule content rather than either. The same signature is strong in Llama and Gemma in the FAA domain and weaker in Qwen; in Mistral the stated rule barely moves the verdict at all (§7). We scope the paper to the rule's qualitative half deliberately: changing the stated number produces little

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

verdict-level movement and therefore no usable verdict-normalized patching target for this procedure (App. A.10). We therefore make no claim here about whether threshold-dependent information is represented elsewhere in the model.

Contributions.

1. **A direction the model decides with, not one we can merely read.** Many late-layer components write permit/prohibit polarity, while a single direction provides a compact causal readout: removing that direction sharply degrades the distinction, while replacing its coordinate through depth recovers 73–93% of the opposite decision (§4).
2. **Injected arguments move the same coordinate that decides the verdict.** A false premise and the conclusion it implies move the internal state onto the coordinate genuine permission occupies: a causal lever, not a change confined to the printed answer (§5).
3. **The effect follows rule content and is not explained by the task or answer words alone.** We build the reference direction each simpler explanation predicts—a generic yes-or-no axis and an answer-word channel—and test all three on matched tasks (§6).
4. **The mechanism generalizes beyond the original regulation.** It is strong in Llama and Gemma on FAA drone-altitude rules and weaker in Qwen; in Mistral the stated rule barely moves the verdict (§7).

2 Related Work

Representation engineering. Concepts and behaviors can be linearly represented and controlled through residual-stream directions (Zou et al., 2023; Arditi et al., 2024; Park et al., 2024), including directions extracted as a difference of means between contrastive examples (Rimsky et al., 2024), the same construction we use for v_r. We apply the same paradigm to rule polarity and, because decodability need not imply causal use (Elazar et al., 2021; Burns et al., 2023), test whether the extracted direction participates in the model's own decision.

Circuit discovery and causal patching. Where representation engineering asks *where* a concept lives, circuit discovery asks *how* it is computed by causally testing components. *Activation patching*, established contemporaneously by Wang et al. (2023) on indirect-object identification and Meng et al. (2022) via causal tracing of factual recall, replaces one component's activation from a clean run with its value in a corrupted run and measures the change in a task output; we adopt the single-token logit-difference metric from Wang et al. (2023). *Path patching* (Goldowsky-Dill et al., 2023) extends the idea to edges, and Conmy et al. (2023) automates the component search; TransformerLens (Nanda & Bloom, 2022) provides the activation hooks our interventions use. Golgoon et al. (2024) bring activation patching to a fair-lending detector (GPT-2 Small, head localization); we work in instruction-tuned models with difference-of-means steering and a counterfactual rule-vs-prior design.

Rule-following versus reciting. Whether a model applies an abstract rule or reproduces memorized associations is a standing question. Wu et al. (2024) and RuleBench (Sun et al., 2025) separate reasoning from reciting with counterfactual task and rule variants, behaviorally; the closest mechanistic analog, Kim et al. (2025), finds a syllogistic-inference circuit "contaminated" by pretraining world knowledge: a rule mechanism competing with a learned prior, in formal logic. Our counterfactual polarity test is a causal test of this paradigm on real rules; the present paper establishes the causal, rule-selective representation itself. As domain grounding rather than a benchmark, legal and finance rule-application datasets (Guha et al., 2023; Holzenberger et al., 2020) and phrasing sources (Islam et al., 2023; Koreeda & Manning, 2021) informed prompt construction, but none isolate a single rule per example at a fixed token position with matched contrastive pairs, which clean direction extraction requires.

3 Method

The same extraction, intervention, and scoring machinery is used unchanged for the anti-structuring rule here and for the aviation domain in §7.

Setup and notation. We work with the residual stream: the running hidden-state vector that each transformer layer reads from and writes to. Write $h_l \in \mathbb{R}^d$ for that vector at layer l, read at one chosen token position: the last token of the prompt, i.e. the position at which the model is about to emit its verdict. We study the computation underlying this immediate constrained verdict, not reasoning elicited through chain-of-thought generation; this design produces no generated sequence within which such reasoning would be expressed. A *contrastive pair* is two closely matched prompts constructed to vary the target experimental contrast while holding the scenario fixed: for anti-structuring, the same proposed action appears under a prohibiting versus permitting rule (Appendix A.2 documents the remaining wording and authority differences; holding them fixed leaves rank-1 recovery at 0.69–0.85, App. A.3). Collecting many such pairs gives a set $\mathcal{D}$ of last-token states, each labelled *permit* or *prohibit*. Throughout, *rule polarity* denotes the permit-versus-prohibit distinction induced by the stated rule with the scenario held fixed; v_r (below) is the residual *direction* we extract for it, and SATISFIED/EVADED its output *readout*.

Extracting a direction (difference of means). The "rule-polarity direction" is simply the vector from the average prohibit state to the average permit state,

$$v_r \;=\; \mathbb{E}_{x\in\mathcal{D}_{\text{permit}}}[\,h_l(x)\,] \;-\; \mathbb{E}_{x\in\mathcal{D}_{\text{prohibit}}}[\,h_l(x)\,].$$

This difference-of-means estimator is deliberately cruder than a trained probe, which could separate the classes by latching onto incidental correlates of the label (a topic word, sentence length); it is also the estimator prior steering-vector work uses (Zou et al., 2023; Arditi et al., 2024). A logistic probe serves as a cross-check, but all interventions use v_r.

The decision metric. The prompt is built so the next token is the decision; we read the logits for two answer tokens (SATISFIED and EVADED) and take their difference,

$$\Delta L \;=\; \text{logit}(\text{SATISFIED}) - \text{logit}(\text{EVADED}).$$

Positive ΔL leans "permitted," negative "violation." Because the two tokens carry different baseline frequencies, we report not raw ΔL but the intervention-induced change $\Delta\Delta L = \Delta L_{\text{intervened}} - \Delta L_{\text{baseline}}$, so per-token priors cancel. The single-token compression follows indirect-object-identification circuit work (Wang et al., 2023). On held-out examples, snap verdicts agree with step-by-step verdicts in 90.0–98.4% of cases for Llama and Gemma versus 55.6–78.9% for Qwen and Mistral; lower BSA accuracy reflects systematic over-flagging rather than random decisions (App. A.5).

Interventions. All interventions are forward-time edits to h_l implemented as hooks on the frozen network (TransformerLens (Nanda & Bloom, 2022)), never a weight edit or a gradient-based update: this keeps every intervention a test of what the model, as already trained, computes with, rather than a demonstration that some modified or fine-tuned version of it could produce the same output. v_r is always extracted at $l = L^*$. The intervention site depends on the experiment: the primary sufficiency test replaces its coordinate across the scanned layers through L^*, while the timing sweep intervenes at one layer at a time. We use three, matched to three questions:

- *Additive steering* $h_l \leftarrow h_l + \alpha v_r$: nudge the state along the direction. Used to ask "does moving along v_r move the decision, and how much?" The scalar α is swept.
- *Ablation* (projecting v_r out of h_l): delete the direction's component. Used for *necessity*: if the decision causally depends on v_r, removing it should systematically degrade the permit/prohibit distinction.
- *Projection-replace / activation patching*: overwrite the v_r-component (rank-1, i.e. just that one coordinate) with the value that same quantity takes when the model processes a matched prompt with the opposite label, then measure whether $\Delta\Delta L$ follows; used for direction-level *sufficiency* (specificity). A second variant overwrites a whole cached component activation the same way, used for circuit localization (§4).

We separately validate this intervention pipeline on the established refusal direction of Arditi et al. (2024) (App. A.1).

Two guards we carry throughout. First, because a direction's causal role changes across depth, we select L^* as the *earliest* layer in the late high-recovery band at which patching the full clean residual state into the corrupt run recovers more than 90% of the decision-logit gap, then freeze $v_r = v_r^{(L^*)}$ there for every downstream analysis. We choose this way rather than by separability (how cleanly

the two classes split) because separability peaks trivially at the very first layer, before the model has done any computation on the input — a layer being separable does not mean the model uses that layer's representation to decide anything. Second, we scale every intervention to the residual-stream norm at the target layer, since a unit vector with a fixed coefficient is negligible against a residual stream one-to-two orders of magnitude larger. Further dataset-design pitfalls we controlled for (telegraphing of the verdict, an over-flagging readout bias, and a risk-salience confound) are detailed in Appendix A.2.

Decodability is not causal mediation. A direction that linearly *decodes* the label need not be the one the computation *uses*, so every claim about v_r below rests on intervention, not probe accuracy; Appendix A.6 shows why this distinction mattered empirically in this experiment.

Sample sizes. Extraction and evaluation splits are disjoint for the cross-fitted estimates; the timing and component analyses are descriptive and reuse them. Each task or contrast is matched between its two conditions; exact counts per experiment are in Appendix A.4.

4 Localizing the Rule-Polarity Representation

This experiment asks whether the steering vector is part of the model's own decision computation, using activation patching.

Setup. Construct minimal clean/corrupted prompt pairs that differ only in the rule-relevant token(s), e.g. (the two prompts below describe the same split deposit under two different stated rules; what makes the clean case compliant is that its rule pre-authorizes the split, not that splitting is inherently fine):

- P_{clean} (the stated rule *permits* the action): *"Rule: pre-authorized deposits kept below the reporting threshold are exempt. Proposed action: splitting $12,000 into three separate deposits. Verdict:"* $\rightarrow$ target completion SATISFIED.
- P_{corrupt} (the stated rule *prohibits* the same action): *"Rule: deposits structured to stay under the $10,000 reporting threshold are a violation. Proposed action: splitting $12,000 into three separate deposits. Verdict:"* $\rightarrow$ target completion EVADED.

As in §3, we read the verdict as $\Delta L = \text{logit}(\text{SATISFIED}) - \text{logit}(\text{EVADED})$ on the clean run and report the baseline-normalized patching effect $(\Delta L_{\text{patched}} - \Delta L_{\text{corrupt}})/(\Delta L_{\text{clean}} - \Delta L_{\text{corrupt}})$ (clean $= 1$, corrupt $= 0$), so per-prompt token priors cancel and the metric reflects the decision rather than token preference.

Distributed write, compact readout. Both necessity and sufficiency interventions give the same picture: the permit/prohibit decision becomes recoverable across a mid-to-late layer band in all four models (Table 1). The polarity coordinate is not written by a small set of attention heads: across the four BSA models, the five largest head effects account for only 15–39% of the total absolute head effect, and 22–121 heads are needed to reach 70% (Fig. 1, left; per-model values in Table 1). Across heads, stronger writing along v_r is associated with larger causal effects ($r = 0.35$–0.57); we treat this correlation descriptively, because heads within a band are not independent samples. Thus many heads write the decision-relevant coordinate.

Yet by the late decision layers, intervening on that direction alone can recover a substantial fraction of the decision. This is what we mean by a *compact readout*: the computation that builds the decision is spread across many components, but much of its effect can be reached through a single direction late in the network, where changing it directly changes the verdict. This does not make v_r an exclusive bottleneck: fixing it reduces but does not eliminate writer-dependent decision variation (App. A.9). We use "reader" throughout for this compact causal handle.

Linking the circuit to the direction. Writes at the decision token accumulate monotonically along v_r across the late band (Appendix Fig. 4). The direction thus identifies the causal write without implying a sparse circuit.

Is the direction necessary, sufficient, and available only late? We test the direction as a whole rather than individual heads (Table 1). First, removing the v_r component across the band shifts held-out clean cases toward EVADED in every model. Second, repeatedly replacing only the v_r-aligned coordinate at the scanned layers through L^* recovers 73–93% of the clean–corrupt decision

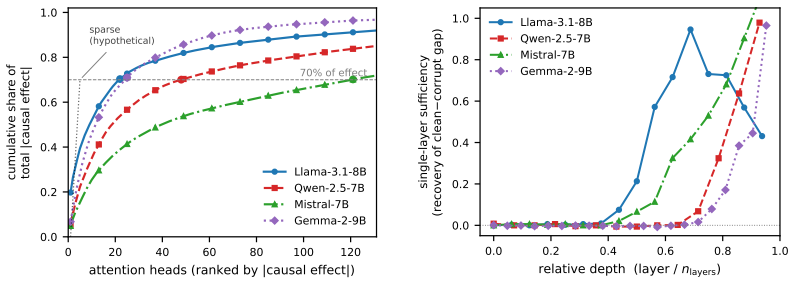


Figure 1: **Distributed write and late readout. Left:** cumulative absolute causal effect as attention heads are ranked by effect magnitude; no model's effect is concentrated in a handful of heads, and the dotted line shows a five-head reference distribution. **Right:** recovery from single-layer projection-patching of v_r across depth; causal efficacy emerges only in the late crystallization band.

difference on held-out scenarios. A single direction therefore carries substantial causal control over the verdict.

We then apply the replacement one layer at a time. It has almost no effect in early layers and becomes effective only in the late band, reaching 0.92–1.06 of the clean–corrupt decision difference across the four BSA models (Fig. 1, right); Mistral reaches 1.055 at L30. The normalization uses the clean–corrupt difference as a reference scale rather than an upper bound, so recovery can exceed 1. A layer-independent output-pointing vector would recover at any layer; instead the effect appears only late. This localizes where the decision becomes controllable through v_r, not the earlier computation that builds it.

Table 1: Causal localization across four BSA models ($n = 36$ matched pairs per model: 12 scenarios $\times$ 3 surface variants, distinct from the held-out set used for compliance-task evaluation elsewhere). Necessity and sufficiency use leave-one-scenario-out evaluation; brackets are 95% delete-one-scenario jackknife intervals. For the two head-concentration statistics, brackets instead show the range across scenario omissions, not sampling uncertainty. 'Crystallizes' is the band over which the decision becomes recoverable; necessity magnitudes are not comparable across models. Recovery is normalized to the clean–corrupt decision-logit gap, which is a reference scale rather than an upper bound, so values above 1 are possible. Columns are defined in §4.

Model	crystallizes	top-5 head share	heads→70%	head r	necessity $\Delta\Delta L$	suff. recovery	timing (early→late max.)
Llama-3.1-8B	L16–22	0.394 [0.375–0.398]	22 [21–26]	0.35	−4.3	79% [78, 80]	0.00 → 0.92 (L22)
Qwen-2.5-7B	L18–26	0.223 [0.213–0.234]	48 [41–58]	0.57	−1.6	93% [90, 96]	0.00 → 0.99 (L26)
Mistral-7B-v0.1	L16–28	0.154 [0.140–0.170]	121 [93–143]	0.53	−1.0	78% [77, 78]	0.00 → 1.06 (L30)
Gemma-2-9B	L28–40	0.283 [0.280–0.284]	25 [24–25]	0.51	−7.7	73% [71, 74]	0.00 → 0.97 (L40)

5 Activation-Level Polarity Convergence

Building on this, we ask a sharper question about what v_r actually *represents*, using a manipulation we call `both_wrong`. v_r is estimated only from permit/prohibit pairs; no `both_wrong` example enters its estimation. We take a genuine violation (a structured deposit that should be flagged EVADED) and inject into the prompt both a false premise ("this $12,000 does not exceed the $10,000 threshold") *and* the false conclusion it implies ("…so it need not be reported"). The model duly flips its output to the permit token. The surface answer has changed; the question is whether anything *inside* the model changed, or whether we merely overwrote the last word.

We can tell these apart because v_r gives us a coordinate to read. On the clean data the last-token residual sits at a high projection onto v_r for a genuine permission (the SATISFIED end) and a low projection for a genuine violation (the EVADED baseline). For each `both_wrong` case we read where its residual lands on that same axis and report a *recovery* fraction: 0 means "still at the violation baseline — only the output token moved," and 1 means "all the way onto the activation state a genuine permission occupies along v_r." A recovery > 1 denotes overshoot *past* that state.

Prompt provenance control. Earlier permit prompts stated their own verdict, allowing v_r to reflect conclusion copying rather than computation. We therefore rewrote all prompts as classification-only and reran the experiment; all results below use the repaired prompts.

Result. The `both_wrong` state converges substantially toward the genuine-permission state along v_r: recovery is 0.76 (Llama), 0.91 (Qwen), 1.00 (Gemma), and 1.11 (Mistral) (95% CIs within ± 0.02, and Mistral's interval $[1.10, 1.12]$ excludes 1.00; the overshoot is not noise). Correct cases remain at baseline, while injecting the false number alone produces only partial movement. Random-direction controls reach only 1–2% of v_r's separating power. Mistral's recovery exceeds 1, landing *past* the mean genuine-permission state; we do not have an account for the overshoot.

We read this as substantial convergence on the same causal polarity *coordinate*, not evidence of an identical upstream circuit (which would require edge-level analysis).

6 Is v_r Really a *Rule* Direction?

Two deflationary accounts could explain every result above, and each must be ruled out before calling v_r a rule representation:

- **Generic decision polarity.** Perhaps v_r is not about *rule polarity* at all but about which way *any* binary decision points. The same direction would then also push "is this statement true?" or "did X happen?", and we would merely have rediscovered the model's general yes/no axis.
- **Output-token identity.** Perhaps v_r only encodes "make the word SATISFIED more likely than EVADED": a fact about the answer vocabulary with no semantic content, which would look causal in any experiment that reads those two tokens.

Our claim, by contrast, is that v_r specifically carries the rule's permit–prohibit polarity. This section tests that claim against both alternatives. The key test is therefore not whether v_r can move the original permit/prohibit output (that is already established) but whether its causal effect follows the *rule content* rather than the generic decision task or the answer vocabulary.

Five tasks, matched in form. We build five binary conditions that vary along two axes: the *task* judged and the *answer vocabulary* (Table 2). Task D is free of numbers and time-ordering, so it cannot re-use threshold or temporal machinery, and task E crosses the truth task with the rule task's answer words (the key task×vocabulary cell). Each task is built from matched $50/50$ pairs (a true vs. near-identical false sentence; a relation-holds vs. relation-absent passage) and split into disjoint *extraction* and *evaluation* halves; nothing used to *define* a direction is ever *measured* on.

Table 2: The five matched specificity tasks (task × answer-vocabulary). Task B is included for completeness; it is non-diagnostic on three of four models (§6).

	Task judged	Question	Answer words
A	rule-following	does the handling satisfy the rule?	SATISFIED/EVADED
B	rule-following	(same judgment)	YES/NO
C	truth	is the statement true?	TRUE/FALSE
D	neutral relation	does the passage state the relation?	YES/NO
E	truth	(truth, in rule vocabulary)	SATISFIED/EVADED

Two reference directions. A random direction tells us only "not arbitrary." We test the two alternatives above by constructing the direction each one predicts, and let the data say which (if either) v_r resembles:

- d_D (the subscript names the hypothesis it tests, not Task D): a *generic binary-polarity* direction extracted from the truth task (Task C), the difference of means between held-out true and false statements. If v_r were really tracking generic decision polarity, it should behave like d_D.
- d_T, a *vocabulary-readout* reference direction: the gradient, taken at the same decision layer, that most raises the SATISFIED-minus-EVADED logit, built from neutral content with no rule-following labels. If v_r were really tracking output-token identity, it should behave like d_T.

Directional sensitivity. For a unit direction d, we nudge the decision-layer residual by $\pm\delta d$ and measure how much the task's decision logit changes: in other words, how strongly task t reads direction d. The exact definition, intervention scale, and linearity checks are in Appendix A.12.

What we find (Figure 2). The two reference directions behave exactly as their hypotheses predict, which is the validation of the instrument: d_D moves *both* non-rule judgments (truth C and relation D), while d_T moves *only* the two tasks whose answer words are SATISFIED/EVADED (A and E) and nothing else. Against that backdrop v_r behaves like *neither* reference direction under the cross-task causal tests. It has near-complete sensitivity on the rule task (it captures essentially all of that task's steerability) and little to none on truth or relation: at the floor on Qwen and Gemma, small on Mistral, and nonzero but clearly smaller than on the rule task for Llama. The signed rule-minus-truth gap $S_A - S_C$ is positive on all four models, and the random direction sits at zero throughout.

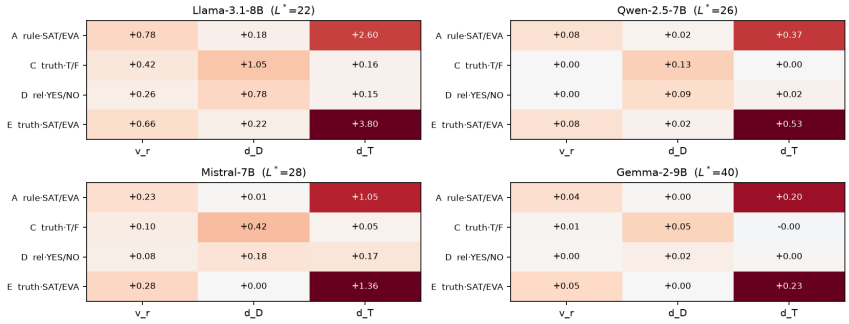


Figure 2: **Task specificity of directional sensitivity.** Finite-difference sensitivity $S_t(d)$ for rule (A), truth (C), relation (D), and truth-in-rule-vocabulary (E) tasks. d_D affects C/D, while d_T tracks the SATISFIED/EVADED vocabulary in A/E; v_r is strongest on A and weak on C/D. Colors show the sign of S; magnitudes are not comparable across models. The grid is in Appendix Fig. 7.

A sufficiency corroboration (Appendix Fig. 5). Sensitivity is a small-signal probe; we corroborate it with the stronger projection-patching intervention used in the circuit analysis. For each control task we take its wrong-pole items and *overwrite* their v_r-coordinate with the value that coordinate takes in a genuine rule-permit state, then ask how much of the task's own gap that recovers. Forcing the rule-permit state recovers the *rule-following* decision (0.84–1.11 across models) but leaves truth and relation almost untouched (0.01–0.16). Two guardrails show this is genuine rather than an artifact of a violent edit: next-token entropy stays near its unpatched baseline on the content tasks, and the generated continuation stays coherent: on the truth task it is in fact *identical* whether we target the rule-permit state or the task's own state, i.e. the rule edit does not perturb the truth verdict at all. A lever-validity check (targeting the same coordinate with the natively-extracted rule-polarity direction v_C) recovers the rule task at 0.94–1.02, so the near-zero result on truth/relation is real non-transfer, not a dead patch.

6.1 TASK POLARITY VERSUS OUTPUT-TOKEN READOUT

Task E is the direct answer to the most likely objection: "maybe you only found a SATISFIED-vs-EVADED direction." When the truth task uses those answer words, v_r shows a sizable effect (sufficiency recovery 0.44–1.09), and the vocabulary reference d_T is *also* large on E, marking E as the *vocabulary* channel, not the content channel. A continuation check confirms the distinction: pushing a genuinely *true/false* item toward the rule-permit state under E makes Llama emit "Satisfied" and then, in the very next clause, explain "…the statement is false": the answer *token* flipped, the semantic content did not. The honest reading is thus a decomposition: v_r has a rule-selective causal effect *on the decision* (what A-versus-C/D establishes) *and* participates in the SATISFIED/EVADED *output channel*, which is why any task read out in those words shows some movement. We therefore describe v_r as rule-*content*-specific rather than fully vocabulary-invariant.

Condition B is non-diagnostic in three models: their yes/no rule answer is barely steerable at this layer even by its own direction, so it provides no useful test of vocabulary-invariance. Llama is also consistently the least specific model here. Its weaker specificity covaries with its lower convergence in §5, but these experiments do not establish a causal relationship between the two.

What the specificity tests establish. The two intervention tests agree across four families: v_r strongly affects rule-following decisions but has little effect on unrelated binary decisions. Although it also participates in the SATISFIED/EVADED output channel, its causal effect on rule following cannot be explained by a generic yes/no direction or by those output tokens alone, and we interpret v_r as a rule-selective causal handle. The next section asks whether the same functional signature survives a change in regulatory content.

7 GENERALITY: THE SAME SIGNATURE IN A SECOND DOMAIN

Does the rule-polarity mechanism of §4–§6 depend on the specific regulatory content (US cash-reporting law), or does it recur under a controlled change of content? We test a second domain with the same *rule structure* — a numerical threshold plus an explicit exception — and the same SATISFIED/EVADED output, changing only the subject matter from cash reporting to drone altitude.

Domain: FAA Part 107 (drone altitude). 14 CFR 107.51(b) sets a 400 ft ceiling for small unmanned aircraft, and 14 CFR 107.200 lets a waiver authorize deviation: the same threshold-with-explicit-exception structure as the BSA rule, in an unrelated domain. The two policies are **P400** (above 400 ft without a waiver is a violation) and **PWAIVER** (a valid 107.200 waiver covers the operation), applied to the *same* above-limit flight so only the stated rule flips the verdict. The rule text states no verdict; the same above-limit flight is held fixed while the presence of a valid regulatory waiver changes its status. The readout tokens are unchanged from BSA.

Preregistered protocol. Two stages, in order. First, an eligibility gate asks whether permit/prohibit separation clears a preregistered threshold (mean $|\Delta L_{\text{clean}} - \Delta L_{\text{corrupt}}| \geq 2.0$); models below it do not enter the battery. Second, for models that pass, L^* is selected by the same causal criterion used for BSA and frozen before the FAA battery, so the battery is out-of-sample with respect to layer choice. The gate is behavioural: it asks whether the verdict moves enough for there to be an effect to localize, not whether polarity is represented; the causal criterion then localizes it. Models were selected using this gate before the expanded evaluation. Qwen passed at selection time (3.31); its separation on the later 36-pair set is 1.91, consistent with the weaker FAA result below.

Selection outcome (3/4 eligible). This FAA-domain v_r is a separate difference-of-means extraction from the BSA one, sharing notation because it is extracted the same way from the same kind of contrast, not because it is the same vector. In three families the stated rule moves the FAA verdict well clear of the gate, and each receives a frozen L^*: Llama (sep $= 10.1$, $L^* = 18$), Qwen (3.31, $L^* = 26$), Gemma (13.2, $L^* = 32$): each a different layer, none carried over from BSA. Mistral's separation is 0.66, below the gate: its *FAA verdict changes too little between the permitting and prohibiting rules to support the planned localization*, so no causal battery is attempted. (The trivial early linear separability present in every model reflects the surface wording of the waiver clause; causal efficacy appears only in the late band, consistent with the timing control of §4.)

Table 3: FAA content-generalization. *Eligible* indicates whether the model passed the preregistered decision-separation gate (≥ 2.0) at selection time; Mistral does not clear it, so no downstream causal test is run. *Rank-1 recovery* is the fraction of the clean–corrupt decision-logit gap recovered by projection-replacing the FAA polarity direction at each model's L^*, reported for all three models that passed the eligibility gate regardless of the component-screen prerequisite below. Qwen's lower recovery accompanies a less selective effect: its ablation moves permit and prohibit cases by nearly the same amount rather than separating them (§7). The expanded ($n = 36$) component screen carries its own decision-separation prerequisite; Qwen's expanded set falls below it ($1.91 < 2.0$), so its component-screen result is not reported.

Model	Eligible	L^*	Component screen	Rank-1 recovery	Interpretation
Llama-3.1-8B	yes	18	yes	0.898	selective polarity control
Gemma-2-9B	yes	32	yes	0.705	selective polarity control
Qwen-2.5-7B	yes	26	—	0.358	less selective polarity control
Mistral-7B	no	—	—	—	downstream causal test not run

The core causal signature reproduces in Llama and Gemma (Table 3). At each frozen L^*, on the expanded $n = 36$ set, the same pattern holds in both: ablating v_r moves the verdict toward

violation (necessity $\Delta\Delta L = -2.5$ for Llama, -12.3 for Gemma), restoring its coordinate recovers a substantial fraction of the permission decision (sufficiency 0.898 [0.889, 0.906] and 0.705 [0.700, 0.710]), the effect is confined to the late band, no random direction reproduces it, and the false-argument condition moves the state toward the genuine-waiver coordinate (1.01 and 0.54).

Qwen is different. Rank-1 recovery is only 0.358 and varies substantially across scenarios (95% CI $[0.107, 0.609]$). Removing the direction also moves the permit and prohibit cases by almost exactly the same amount (-1.26 and -1.25). That is not the selective effect expected if this direction strongly controlled whether Qwen treats the rule as permitting or prohibiting the action, and it contrasts with Llama (-2.5 versus $+7.3$) and Gemma (-12.3 versus $+6.5$), where the two conditions move apart. Together with the low recovery, this makes Qwen's FAA result substantially weaker than Llama's and Gemma's. Its false-argument convergence (0.59) is measured on a separate 288-item set and is unaffected by this expansion.

The evaluable FAA replications preserve the same writer organization. Across the six evaluable model-domain pairs, the five largest head effects account for 15–41% of the total absolute head effect, with 16–121 heads required to account for 70%; the write-alignment/effect correlation remains positive throughout ($r = 0.35$–0.74). Qwen's FAA cell is not evaluable for this analysis (see Table 3).

What FAA establishes. The BSA causal organization reproduces strongly in Llama and Gemma, is weaker and less selective in Qwen, and cannot be tested as planned in Mistral, whose verdict barely responds to the rule. FAA therefore shows generalization across regulatory content while holding the threshold-plus-exception rule structure fixed. Generalization to structurally different rules (sequential, conjunctive, or conditional rather than threshold-based) remains open.

8 LIMITATIONS

These claims have several boundaries. The experiments identify a compact causal handle on the decision, not a one-dimensional computation or an exclusive bottleneck: fixing v_r reduces but does not eliminate writer-dependent effects (App. A.9). The specificity tests rule out generic binary polarity and answer-token identity as explanations of the effect, but do not establish an explicit symbolic or fully vocabulary-invariant representation of the rule; Llama is also less selective than the other models. Component-level patching can mix direct effects with interactions among components (Vaidyanathan et al., 2026).

The evidence also has limits of scope. Localization uses 12 scenarios with three surface variants each, and the cross-domain test changes the subject matter while preserving the same threshold-plus-exception rule structure. Whether the result extends to structurally different rules, whether the direction remains geometrically stable across models and domains, and how the mechanism relates to multi-token chain-of-thought reasoning (§3) remain open.

9 CONCLUSION

The mechanistic picture is a distributed computation with a compact causal handle: many late-layer components write the permit/prohibit polarity, while a single direction becomes necessary and substantially sufficient near the decision layer. That handle is *rule-selective*: its causal effect cannot be explained by generic binary-decision polarity or output-token identity alone. When we inject into a genuine violation a false premise and the conclusion it implies, the model's internal state moves onto roughly the same coordinate a real permission would occupy: the same coordinate that decides the verdict is therefore one the model's own computation moves, not merely reads. In FAA drone-altitude rules the same causal organization is strong in Llama and Gemma and weaker in Qwen.

A decodable or steerable direction is not thereby a mechanism: the critical question is whether the model's own decision causally reads it. Here it does. Many components write a rule's polarity; one direction provides a compact causal readout of it: many writers, one reader.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, implementing methods, and cleaning and reformatting datasets. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, or supporting qualitative and thematic data analysis and interpreting results. Providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures, creating or editing software code, and formatting references. We have reviewed all AI-assisted work; LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Layer selection (L^*) and the decision-separation and alignment gates (2.0 and 0.7) were fixed in advance of the corresponding causal measurements and recorded in timestamped protocol files; "preregistered" means commit-before-measurement discipline, not a public registry. L^* and v_r are produced by `code/rule_circuits/activation_patch.py`, wrapped by `code/runpod/run_path_patch.sh`: its `Lstar` field is the earliest scanned layer whose denoising recovery exceeds 0.9 and its `v_dir` field is the polarity difference-of-means at that layer, so the selection rule stated in §3 can be verified directly against the released code. The scan defaults to every other layer, which is why the component screen of §4 covers even-indexed layers; `--layers` overrides it. All code, data, prompts, and extraction/intervention scripts are included as supplementary material with this submission and will be released publicly on GitHub upon completion of the review process.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. https://arxiv.org/abs/2406.11717.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023. https://arxiv.org/abs/2212.03827.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. https://arxiv.org/abs/2304.14997.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9, 2021. https://arxiv.org/abs/2006.00995.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023. doi: 10.48550/arXiv.2304.05969.

Ashkan Golgoon, Khashayar Filom, and Arjun Ravi Kannan. Mechanistic interpretability of large language models with applications to the financial services industry. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF)*, 2024. https://arxiv.org/abs/2407.11215.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils

Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. https://arxiv.org/abs/2308.11462.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. A dataset for statutory reasoning in tax law entailment and question answering. In *Proceedings of the Natural Legal Language Processing Workshop (NLLP) at KDD*, 2020. https://arxiv.org/abs/2005.05257.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023. doi: 10.48550/arXiv.2311.11944.

Geonhee Kim, Marco Valentino, and André Freitas. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. https://arxiv.org/abs/2408.08590.

Yuta Koreeda and Christopher D. Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1907–1919, 2021. https://arxiv.org/abs/2110.01799.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. https://arxiv.org/abs/2202.05262.

Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022. Open-source library for mechanistic interpretability of GPT-style language models.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024. https://arxiv.org/abs/2311.03658.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. https://arxiv.org/abs/2312.06681.

Wangtao Sun, Chenxiang Zhang, Xueyou Zhang, Xuanqing Yu, Ziyang Huang, Haotian Xu, Shizhu He, Jun Zhao, and Kang Liu. Beyond instruction following: Evaluating inferential rule following of large language models. In *Proceedings of the 24th China National Conference on Computational Linguistics (CCL)*, pp. 1043–1066, 2025. https://arxiv.org/abs/2407.08440.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. *arXiv preprint arXiv:2606.27510*, 2026. https://arxiv.org/abs/2606.27510.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. https://arxiv.org/abs/2211.00593.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of NAACL-HLT*, 2024. https://arxiv.org/abs/2307.02477.

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, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023. doi: 10.48550/arXiv.2310.01405.

A Additional Experimental Details

A.1 Pipeline validation on a known causal direction

Before applying the intervention pipeline to rule polarity, we validated it on refusal, for which a causal residual-stream direction is already established (Arditi et al., 2024). We extracted a refusal direction using the same difference-of-means procedure and applied the same intervention machinery. Removing the direction reduced refusal on harmful prompts from 96% to 46% ($23/24$ to $11/24$), while additive steering increased refusal on benign prompts from 0% to 92% ($0/24$ to $22/24$ at $\alpha{=}4$). At larger intervention strengths, generations became incoherent and refusal markers disappeared entirely. This known-answer test verifies that the intervention pipeline can recover an established causal direction and also motivated the intervention-strength checks used in the rule experiments.

A.2 Pitfalls in extracting a clean rule direction

Getting a *clean* direction turned out to depend more on dataset and prompt design than on the interpretability machinery. We record the failure modes because they recur for anyone extracting a "concept direction" for a subtle, rule-governed judgment.

Telegraphing turns the direction into a surface artifact. If a violation example names its own illegality ("split the deposit *to avoid the report*", "keep each *under $10,000*"), the model can label it by spotting those words, and the difference-of-means direction then encodes *those tokens*, not the rule. A held-out split does not catch this, because the give-away phrasing appears on both sides of the split. We therefore state only the behaviour (one sum recorded as several sub-threshold deposits) and require the model to infer the rest.

De-telegraphing exposes an over-flagging readout bias. Removing the cues reveals the opposite failure: asked for a one-token verdict, the model exhibits a prior toward treating financially unusual actions as suspect, and flags $\sim 90\%$ of *compliant* actions. Read naively this looks like "the model doesn't know the rule." But the *same* model, allowed to reason step by step first, is 96% correct and cites the fragmentation-below-threshold mechanism explicitly. So the deficit is a biased *readout* at the decision token, not missing knowledge. Two fixes follow and are held fixed across all experiments: a neutral "transaction-monitoring" persona instead of a "careful auditor" one (which primes flagging), and the de-loaded token pair SATISFIED/EVADED instead of a scary word like VIOLATION that would bias the single-token read.

A rule direction can secretly be a "this-looks-risky" direction. Because violations can also appear more suspicious than permitted actions, a plain contrast can confound rule polarity with risk salience. We therefore used a 2×2 design crossing legal status with surface suspicion, including *suspicious-but-permitted* and *bland-but-violating* cases, and separately probed suspicion-related directions. These diagnostics motivated the neutral persona and matched contrasts used in the final experiments. They do not establish a model-general geometric separation between rule polarity and suspicion, so we make no such claim.

Our contrast holds the number fixed, but it changes more than just the permission. We record what our own pairs do and do not isolate. First, to be clear about which side of the pair varies, since §3 states it compactly: the action is held fixed (a given amount, split into a given number of deposits, each one below the reporting line) and the stated rule is what differs between the two prompts, one prohibiting that split and one permitting it. Both rule texts open with the same reporting clause word for word, including the same dollar figure, so nothing in the contrast varies the number. That is what licenses calling the extracted direction a polarity direction rather than a threshold one (App. A.10). The differing sentence also changes authority attribution, motive framing, and some wording of the action, including use of the technical term "structuring." These features can therefore ride along with the permission contrast. The specificity battery (§6) does not separate permission from this framing, so we test the confound directly in Appendix A.3: holding authority, motive and action wording fixed leaves rank-1 recovery at 0.69–0.85 across all four models. The FAA replication (§7) provides a further change of realization, using a regulatory waiver rather than an institutional exemption.

Choose the intervention layer by causal effect, not by separability. Class separation along v_r peaks trivially at layer 0 (it just reflects the input tokens) and is a poor guide to where intervening

changes behaviour. We instead pick the earliest layer at which patching the clean residual stream recovers more than 0.9 of the decision-logit gap, and freeze that layer (L^*) before any downstream analysis. Two details of that rule matter for anyone reusing it. It is a *first crossing*, not a maximum: the causal effect keeps rising toward the output, so in two of our four families a maximum would instead select a layer near the top of the network (L30 of 32 in both Llama and Mistral) rather than the layer at which polarity first becomes sufficient. And it scans the *full* residual stream, not a single direction; v_r is extracted only after L^* is fixed, so the layer choice does not presuppose the direction it will be used to test.

L^* lies within a broader late-layer plateau. Recovery crosses the 0.9 threshold gradually rather than at a sharp transition, so small differences between runs move the exact first-crossing layer substantially: on re-measurement the crossing shifted by 0–6 layers across the four families, while scanning every integer layer rather than every other accounted for at most one. Recovery nonetheless stays above 0.9 continuously through the late plateau, and every selected L^* remains inside it (recovery at L^*: Llama 0.973, Qwen 0.989, Mistral 0.940, Gemma 0.968). The downstream analyses therefore require L^* to identify a high-recovery late layer, not a uniquely privileged one.

Scale the intervention to the residual-stream norm. A unit vector added with a small coefficient is negligible against a residual stream whose norm is one-to-two orders of magnitude larger, which shows up as a spurious "steering does nothing" null. We scale strengths relative to the mean residual norm at the target layer. (This exact mistake, injecting a norm-1 vector, invalidated our first task-specificity attempt in §6; we flag it as a recurring trap.)

A.3 Holding authority, motive and action wording fixed

The rule texts above differ in more than the permission, so we repeat the localization on a matched contrast that removes those differences. Both arms attribute the classification to the same regulatory authority, neither supplies a motive, the act is described in identical words, and no loaded term appears on one arm only; the arms differ in whether an exemption applies and hence in how the act is classified. The prompts are otherwise byte-identical within a pair.

The contrast first has to move the decision at all. On the same 36 pairs it separates permit from prohibit by 4.80–27.53 logits, clearing the 2.0 eligibility criterion used elsewhere in all four models, and all 12 scenario-level separations in each of the four models (48 model-scenario cells) individually exceed it — so no model's pass is carried by a few scenarios. We then rerun the localization at each model's frozen L^*, with no reselection.

Table 4: Authority/motive control. Separation is the permit/prohibit decision-logit gap on the matched contrast; 2.0 is the eligibility criterion. Sufficiency is rank-1 projection-patch recovery, leave-one-scenario-out, with 95% delete-one-scenario jackknife intervals. Concentration is components reaching 70% of total causal effect.

		sufficiency		heads to 70%	
Model	separation	original	matched contrast	original	matched
Llama-3.1-8B	7.09	0.791	0.805 [0.796, 0.814]	22	25
Qwen-2.5-7B	27.53	0.930	0.853 [0.846, 0.860]	48	25
Mistral-7B	4.80	0.775	0.801 [0.795, 0.807]	121	35
Gemma-2-9B	23.98	0.726	0.689 [0.679, 0.699]	25	25

Replacing the rank-1 coordinate still recovers 0.69–0.85 of the decision in every model, the write stays distributed, and each model's recovery still emerges only in the late band at its frozen L^*. The effect is therefore not produced by the authority, motive, action-wording, or loaded-terminology differences carried by the original contrast.

Not every statistic is equally stable across the two realizations. Sufficiency moves little, but writer concentration does: Mistral goes from 121 components to 35 while its sufficiency changes by 0.026, and the four models' concentrations span 22–121 originally but 25–35 here. Necessity also shifts substantially in Qwen (-1.61 to -8.97). What transfers across rule realizations is the compact readout, not the detailed organization of what writes it.

Model	Domain	n	Agreement	Parse rate	Snap acc.	Step-by-step acc.
Llama	BSA	128	125/128 (97.7%)	100%	65.6%	68.0%
Llama	FAA	160	144/160 (90.0%)	100%	83.8%	93.8%
Qwen	BSA	128	101/128 (78.9%)	100%	45.3%	60.2%
Qwen	FAA	160	89/160 (55.6%)	100%	45.6%	88.7%
Mistral	BSA	128	76/128 (59.4%)	100%	65.6%	51.6%
Mistral	FAA	160	90/153 (58.8%)	96.3%	46.9%	85.7%
Gemma	BSA	128	126/128 (98.4%)	100%	65.6%	65.6%
Gemma	FAA	160	147/160 (91.9%)	100%	91.9%	100.0%

Table 5: Snap-versus-step-by-step scorer validation. Agreement counts and parse rate are both over the examples that ended with a parseable verdict (see App. A.5 for Mistral's FAA shortfall); accuracy columns are against the gold label.

For Qwen, the rank-1 and component analyses completed before the head-alignment reproduction gate; that gate failed (0.543 versus 0.539, tolerance 0.0015), so we report the completed results but withhold the subsequent head-r and timing analyses.

A.4 Sample sizes by experiment

Circuit localization patches 36 matched clean/corrupt polarity pairs per model and domain (12 scenarios $\times$ 3 identities; same action, stated rule flipped), evaluated leave-one-scenario-out. The result is insensitive to scenario composition: across the 12 leave-one-scenario fits, BSA sufficiency spans 0.790–0.794 (Llama), 0.921–0.935 (Qwen), 0.772–0.776 (Mistral) and 0.722–0.728 (Gemma). The head write-alignment correlation runs over all band-layer heads ($n = 336$–480). A separate threshold-varying contrast (App. A.10) uses 7 matched pairs and is degenerate in all four families. Direction extraction and the compliance tasks use the full-scenario split (BSA: 56 permit / 32 prohibit for extraction, 28 / 16 held-out; FAA: 40 / 40 and 20 / 20). The five specificity tasks are balanced $24/24$ except tasks A and B, which reuse the BSA held-out split (28 permit / 16 prohibit), with finite-difference $\delta = 0.10\times$ the mean residual norm. The activation-convergence test uses a 288-item control set and a 100-direction random null, and the refusal pipeline validation (App. A.1) uses 24 harmful and 24 harmless instructions.

A.5 Scorer validation: snap versus step-by-step generation

We compare the constrained snap verdict against the verdict reached after unconstrained step-by-step generation, on the same held-out examples used elsewhere (BSA $n = 128$, FAA $n = 160$; Table 5). The two are not interchangeable: for Llama and Gemma they agree on 90.0–98.4% of examples, but for Qwen and Mistral only 55.6–78.9%. Accuracy against the gold label is lower still and moves separately from agreement — on Llama's BSA examples the two verdicts agree 97.7% of the time while each is correct only 65.6–68.0% of the time, so high agreement here reflects two methods reaching the same answer, not two methods reaching the right one. Generations end with the required `Final verdict:` line in every case except Mistral's FAA condition (153/160; the remaining 7 include one line stating the verdict in different words, caught by a fallback parse, and six that loop through near-identical restatements of the prompt until the generation-length cap is reached without ever concluding). Low gold accuracy does not imply random snap decisions. On the BSA validation set all four models classify every violating case as EVADED; their errors are false positives on compliant cases, with Qwen over-flagging most strongly ($70/88$). Gold accuracy therefore measures something different from the rule-responsive decision contrast localized in §4. We do not treat snap decoding as a general proxy for step-by-step reasoning; the causal experiments in §4 concern the snap decision specifically, and this validation bears on how well that decision corresponds to a separately elicited judgment, not on the internal validity of manipulating it.

A.6 Decodability versus causal mediation: full disambiguation

A direction that linearly *decodes* the label need not be the thing the model *uses*: a probe can read a feature off the residual stream that the computation never consults. Our factorial probe shows the rule is trivially decodable at the embedding (≈ 1.0 at $L0$, in every model); whether v_r *causally*

mediates the verdict is a separate question, and an initial "ablate it and see if flagging stops" test came back null. That null is uninformative on its own, because it is consistent with several very different situations: a *saturated metric* (the binary verdict is already pinned at 100%, so removing one direction cannot push it across a distant boundary), a *wrong layer/scale* (the earlier norm bug), a *redundant subspace* (the concept spread over several directions, so deleting one is repaired downstream), or a genuinely *decodable-but-unused* direction. We separate these by (i) switching from the saturated binary verdict to the continuous $\Delta\Delta L$, whose small-α slope is a boundary-independent sensitivity measure, and (ii) selecting the layer by causal effect rather than separability. On anti-structuring the resolution is that the null was the saturated-metric-plus-wrong-layer artifact: at L^* a norm-scaled intervention along v_r moves the verdict strongly and oppositely while random and orthogonal-complement directions stay flat, so a causal, layer-localized rule direction does exist. This disambiguation is the backbone of the circuit-localization analysis (§4).

A.7 Rank-1 versus higher-rank polarity subspaces

Higher-rank interventions differ sharply across models. Starting from v_r, we add principal directions of the permit/prohibit differences orthogonal to v_r and compare them with random directions from the same residual span. Additional polarity dimensions increase recovery in Llama, increase it from rank 3 in Mistral, add little in Qwen, and reduce recovery through rank 5 in Gemma before increasing it at rank 10. Because these added dimensions can produce larger interventions than the random controls, we repeat the comparison after matching displacement in the added dimensions. The Llama and Mistral advantages remain ($G_{10} = 0.041$ and 0.017), Qwen falls near zero (0.003), and Gemma's negative low-rank effect becomes stronger. Thus v_r provides a strong rank-1 causal handle in all four models, but additional polarity-correlated dimensions improve recovery only in some models.

Table 6: Rank-1 versus higher-rank polarity subspaces, BSA, at each model's frozen L^*. R_1 is rank-1 projection-patch recovery; $\Delta_{\mathrm{pol}}(k)$ is the change from adding $k-1$ polarity directions. $G(k)$ subtracts the change produced by $k-1$ random directions drawn from the same residual span, so it isolates what the polarity ordering buys: $R(k)$ rises mechanically as the intervention approaches full-residual patching, and its slope alone is uninformative. G_{matched} repeats that comparison with the added dimensions' displacement matched to the random controls'.

Model	R_1	$\Delta_{\mathrm{pol}}(2)$	$\Delta_{\mathrm{pol}}(3)$	$\Delta_{\mathrm{pol}}(5)$	$\Delta_{\mathrm{pol}}(10)$	$G(10)$	$G_{\mathrm{matched}}(10)$
Llama-3.1-8B	0.791	$+0.023$	$+0.025$	$+0.028$	$+0.062$	$+0.044$	$+0.041$
Qwen-2.5-7B	0.930	$+0.005$	0.000	$+0.003$	$+0.005$	$+0.012$	$+0.003$
Mistral-7B	0.775	$+0.002$	$+0.019$	$+0.017$	$+0.027$	$+0.020$	$+0.017$
Gemma-2-9B	0.728	-0.011	-0.009	-0.019	$+0.036$	$+0.028$	$+0.033$

Gemma's negative effects at $k = 2, 3, 5$ survive displacement matching and become slightly stronger (-0.014, -0.012, -0.022), so its leading polarity directions are counterproductive under this intervention rather than merely producing larger patches. Subspaces are nested (v_r is in every one, treatment and control alike), so the comparison isolates the added dimensions and does not search for a subspace that replaces v_r. Estimation is leave-one-scenario-out, 32 random draws per fold, and the $k = 1$ arm reproduces each model's published rank-1 recovery to within 0.002.

At rank 10, random directions from the same residual span also increase recovery in Llama, Mistral and Gemma, and additional polarity directions improve recovery in Llama and Mistral. Thus v_r is a compact causal handle, not the only direction that can causally affect the decision.

A.8 Component activation-patching procedure and method sketch

Procedure (component activation patching). Run the clean prompt and cache every internal activation. Run the corrupted prompt. Rather than brute-forcing every attention head and every MLP at every layer, we first prune the search with a cheaper layer-level pass, then only drill into individual heads/MLPs within the layers that matter (an automated alternative is ACDC (Conmy et al., 2023)). Within the pruned space, one component at a time, overwrite that component's activation *during the clean run* with its cached value *from the corrupted run*, leaving everything else clean, and re-measure ΔL. A component whose patching causes a large drop in ΔL is causally load-bearing; a component

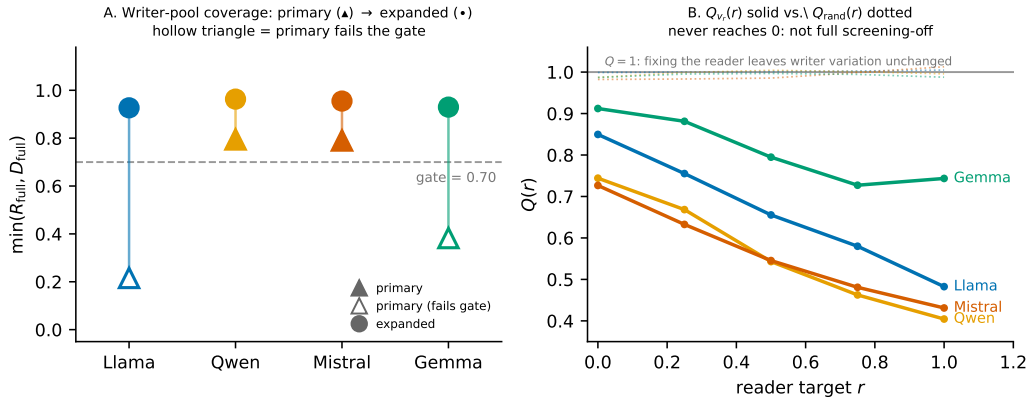


Figure 3: **A. Writer-pool coverage.** $\min(R_{\text{full}}, D_{\text{full}})$ for the primary (preregistered) and expanded within-band writer pools, per model; the gate is 0.70. Llama and Gemma fail on primary and pass only on expanded; Qwen and Mistral pass on primary. **B.** $Q_{v_r}(r)$ **against the norm-matched random-direction null** $Q_{\text{rand}}(r)$, each model at its coverage-qualified arm. Q_{v_r} falls well below the flat Q_{rand} at every r but never approaches 0: the writers use v_r, but the reduction is incomplete in every model.

whose patching has no effect is not, regardless of whether it correlates with the label. We run this on the anti-structuring domain and replicate across four instruction-tuned families (Llama-3.1-8B, Qwen-2.5-7B, Mistral-7B, Gemma-2-9B). We patch at the last (decision) token, which localizes *where the verdict is read out* rather than tracing the full sequence derivation.

A.9 Distributed writers use the polarity direction, but not exclusively

Distributed writers use the polarity direction, but not exclusively. Does the compact readout established above mean that upstream writers influence the decision *only* through v_r? We test this directly by independently overwriting only the v_r-aligned coordinate at the decision layer, leaving the orthogonal residual state unchanged, and asking whether writer-subset-dependent verdict variation survives. Norm-matched interventions along 20 random directions orthogonal to v_r provide an intervention-strength control.

Concretely, the v_r coordinate is clamped by the same projection-replacement operator used throughout this paper: $h' = h + (z_{\text{target}} - \langle h, \hat{v}_r \rangle)\hat{v}_r$, which satisfies $P_{\perp v_r} h' = P_{\perp v_r} h$ — the orthogonal residual state is provably untouched, not merely assumed to be. Let $D_p(S, r)$ be the decision-logit gap for matched pair p when writer subset S is restored and the v_r coordinate is clamped to reader target r, and $D_p^{\text{nat}}(S)$ the same quantity with no clamp. With $B_p(r) = \text{SD}_S[D_p(S, r)]$ and $B_p^{\text{nat}} = \text{SD}_S\big[D_p^{\text{nat}}(S)\big]$ (the standard deviation across writer subsets S of fixed size, for pair p), $Q_{v_r}(r) = B_p(r)/B_p^{\text{nat}}$, averaged over pairs: $Q = 1$ means fixing the reader leaves writer-dependent decision dispersion unchanged; $Q = 0$ means complete screening-off. Q_{rand} repeats the measurement along a norm-matched random direction orthogonal to v_r. Matching the perturbation norm row by row precludes a fixed target across subsets, so it is an intervention-strength control rather than a second screening-off test.

Across writer pools that pass the preregistered coverage criterion, fixing v_r selectively reduces writer-dependent decision variation, whereas norm-matched random-direction interventions leave it essentially unchanged (Table 7). The reduction is incomplete in every model: Q_{v_r} remains well above zero. Thus the distributed writers causally use v_r, but their influence on the decision is not confined to it.

The preregistered primary writer pool passes the coverage criterion in Qwen and Mistral (Fig. 3). Llama and Gemma require the prospectively specified expanded within-band pool: the original every-other-layer screen under-covers the jointly restorable pathway in those models. We therefore describe the population as *band-localized writers* rather than an exhaustive writer set.

Table 7: Writer–reader causal decoupling. Q is the writer-dependent decision dispersion under intervention on the indicated residual direction, normalized by its natural value ($Q = 1$ denotes no reduction; $Q = 0$ complete screening-off). We report the minimum Q_{v_r} across the tested reader targets and the corresponding norm-matched random-direction control. Primary denotes the preregistered writer pool; expanded denotes the prospectively specified expanded within-band pool used when primary coverage failed.

Model	Writer pool	$\min_r Q_{v_r}(r)$	Q_{rand}
Llama	expanded	0.483	≈ 1.00
Qwen	primary	0.404	≈ 1.00
Mistral	primary	0.431	≈ 1.01
Gemma	expanded	0.727	≈ 0.99

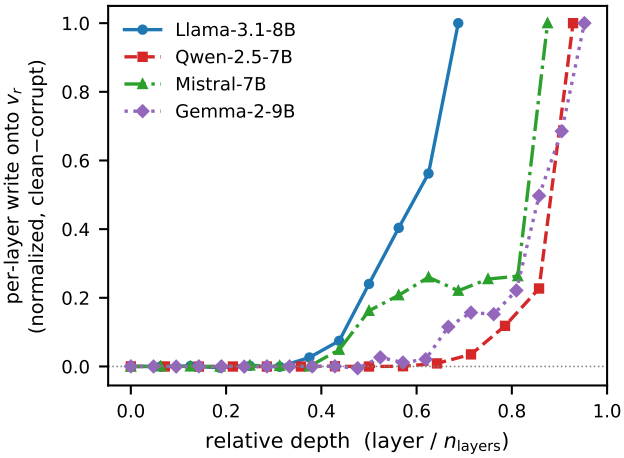


Figure 4: **Depth localization of polarity writing.** Attention and MLP outputs at the decision token are projected onto the normalized clean−corrupt direction v_r. Polarity-aligned writing is minimal in early layers and accumulates in the late crystallization band.

A.10 Polarity, not the numeric threshold

Polarity, not the numeric threshold. We also asked whether the same causal pathway carries the rule's *quantitative* threshold, running the patching procedure on a matched *threshold-varying* contrast of 7 pairs (action fixed, only the stated dollar amount changes, e.g. \$10k vs. \$15k) alongside the polarity contrast. This is a scope check on a separate contrast, not part of the expanded polarity analysis, and we leave it at its original size: the result is a uniform null across four families. Changing the stated threshold flips the *correct* answer: with the action held at, say, \$12k split into two \$6k deposits, a \$10k threshold makes the split structuring (the total is reportable and each piece is not) while a \$15k threshold does not (the total is itself below the reporting line). The model's verdict does not follow: across all four families, changing the stated threshold produces little verdict-level movement relative to changing rule polarity, with the decision-logit gap moving 0.02 to 0.77 under the threshold contrast ($n = 7$) against 2.9 to 23.2 under the polarity contrast ($n = 36$), and both prompts stay on the EVADED side. Path patching attributes a *change* in the verdict between two prompts to the components that carry it; with no change to attribute, the quantity the method estimates is undefined rather than small, so the contrast provides no usable verdict-normalized patching target for this procedure, which is why none appears in our artifacts. This does not imply that threshold-dependent information is absent from the model; it establishes only that changing the threshold has little causal leverage on the verdict in this task.

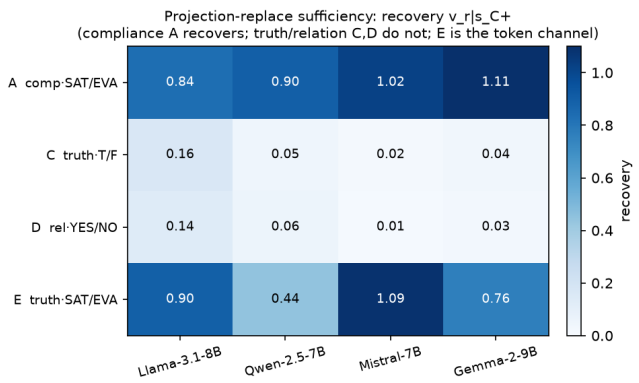


Figure 5: **Task specificity of projection-patch sufficiency.** Fraction of each task's clean–corrupt gap recovered by replacing the v_r coordinate with its value in a genuine rule-permit state. Recovery is substantial for rule-following (A) but small for the unrelated truth (C) and relation (D) tasks; task E uses the same SATISFIED/EVADED vocabulary as the rule task.

A.11 Depth localization of polarity writing

A.12 The finite-difference instrument: an earlier null

The instrument: directional sensitivity (and why not the obvious test). The obvious experiment is to add a rule-polarity direction to a control task and see whether the answer moves. Our first attempt did exactly that and was uninformative: the vector we added had been unit-normalized, so a norm-1 nudge against a residual stream of norm in the tens-to-hundreds barely perturbed anything; even the rule-following positive control failed to move. (A superseded implementation bug, not a scientific negative.) The fix is to measure a *sensitivity* rather than a one-shot recovery. For a unit direction d we nudge the decision-layer residual a small amount δ along d and an equal amount against it, and record how much the decision logit $\mathrm{DL} = \mathrm{logit}(\mathrm{word}_0) - \mathrm{logit}(\mathrm{word}_1)$ moves:

$$S_t(d) \;=\; \frac{\mathrm{DL}_t(h + \delta d) - \mathrm{DL}_t(h - \delta d)}{2\delta}.$$

This is a finite-difference derivative: "how strongly does task t's decision read direction d?" It has no task-difficulty term in the denominator, so a very confident task is not mechanically penalized, and δ is a chosen fraction of the residual norm ($0.10\times$ the mean residual norm at the layer), so magnitude is controlled rather than accidental. We use a single common δ for all tasks within a model and confirm on a small sweep that we are in the linear regime.

A.13 Task specificity of projection-patch sufficiency (figure)

A.14 Localization of rule-specific sufficiency

Where in the network does this live? (localization). A final robustness check asks *where* the rule-specific sufficiency becomes available. We transport the frozen v_r and apply the same projection-replace at each layer of a pre-specified late band, measuring rule-following (A) recovery and the off-target C/D recoveries per layer. The frozen L^* was fixed by the circuit causal-layer analysis (§4) *before* this sweep, so its relationship to the recovery curve is an independently specified comparison rather than a tuned choice. Across all four families the rule-following recovery is ≈ 0 through the early and middle layers and rises only within the late band, peaking at the independently-frozen L^* in three of four families (in Mistral the top is flat, 1.023 at L28 against 1.024 at L30), while off-target C/D recovery stays flat throughout and next-token entropy shows no pathology (Figure 6). The *shape* of the emergence differs by model: Llama shows a clean rise-and-fall centred on L^* (the clearest "localized to the decision-formation band" signature) whereas Qwen and Gemma show a late onset with L^* near the top of the swept range (only the rising side is visible), and Mistral shows a late plateau. We therefore state that the sufficiency is *concentrated in the late decision-formation region, with its maximum at or effectively tied with* L^*, not that it is sharply localized to a single

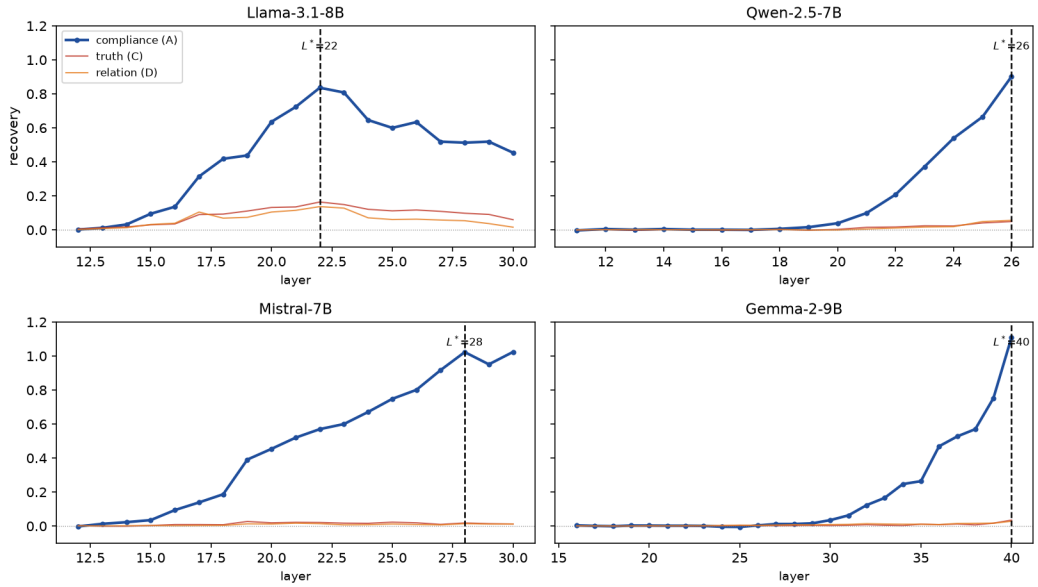


Figure 6: **Late-band localization of rule-specific sufficiency.** Recovery from projection-replacing the frozen v_r coordinate is swept across the pre-specified late band. Rule-following recovery (A) rises in the late band and reaches its maximum at or effectively tied with the independently selected L^* (dashed), whereas off-target truth (C) and relation (D) recovery remains near zero.

layer in every model. Crucially L^* is *not* re-selected from this sweep: it remains the pre-registered site, so its agreement with the recovery curve is a consistency check on the independently selected layer, not a post-hoc fit.

A.15 Task specificity of directional sensitivity (full five-direction grid)

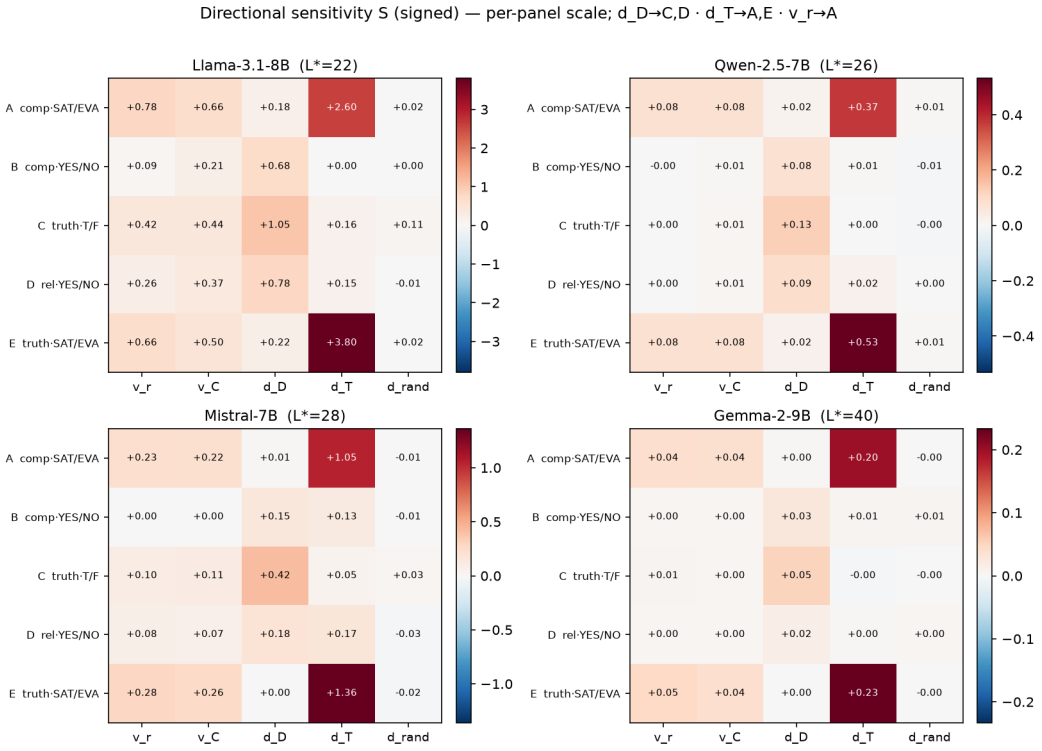


Figure 7: **Full task-specificity grid.** Signed finite-difference sensitivity $S_t(d)$ for all five tasks (A–E) and all five directions (v_r, v_C, d_D, d_T, d_{rand}), one panel per model (per-panel color scale: absolute S is not comparable across models). Body Figure 2 shows the A/C/D/E $\times$ $v_r/d_D/d_T$ subset. The native control v_C tracks v_r (lever validity); d_{rand} is ≈ 0 throughout; and row B (COMP YES/NO) is non-diagnostic on three of four models, so we do not lean on it.