REPRESENTED BUT NOT COMPOSED: HOW NUMERICAL COMPARISONS ENTER LANGUAGE-MODEL DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that a language model can represent a rule's numerical comparison without that representation causally controlling the verdict. We test this across four instruction-tuned models (Llama, Qwen, Gemma, Mistral) in two controlled rule-following settings. In a simple US Federal Aviation Administration (FAA) altitude rule, changing the stated ceiling strongly changes the verdict in the three models with a sufficiently strong FAA decision signal. In a more complex anti-structuring rule, where an amount–threshold comparison must combine with how the transaction is split, changing the threshold barely changes the verdict even when only one comparison changes. We contrast this numerical comparison with rule polarity (whether the rule permits or prohibits the action). This is not a representation failure: threshold-dependent information reaches the decision layer even when the gold answer is held fixed. In Llama, this representation is graded and causally movable, yet moving it across the threshold shifts the verdict by only 2.0–2.3% of the polarity-control effect; the identical intervention on the polarity decision direction serves as a positive control and flips the verdict completely. Across all four models, the late-layer components that drive the decision respond strongly to qualitative action structure but negligibly to the stated threshold. The same dissociation appears behaviorally: in Llama, telling the model that the full amount is below the threshold repairs 0 of 115 over-flagged cases, while telling it the resulting qualitative judgment that the split does not evade reporting repairs 115 of 115. Gemma matches this exactly; Qwen shows the same asymmetry less completely; Mistral shows partial repair. The numerical comparison is represented, graded, and causally movable, yet barely controls the verdict in this task: it is represented but not composed.

1 INTRODUCTION

A broader question motivates this work: whether language models can apply stated rules reliably using computation already inside the model, rather than relying on an external tool to carry out the inference. Answering it requires more than checking whether a model represents the facts a rule requires: a model can represent a condition required for a rule without that representation causally composing into the decision, so a behavioral error need not mean the information was unavailable. We call this causal composition: represented information must acquire control over the decision that depends on it. We test this requirement in two settings that pose the same causal question but differ in what determines the answer. In FAA drone-altitude rules, a single comparison (altitude vs. stated ceiling) is the rule: changing the ceiling moves the verdict more than the model's own permit/prohibit signal, in every model with a sufficiently strong FAA decision signal (three of four). In the anti-structuring task below, the same manipulation barely moves the verdict: an amount–threshold comparison must combine with how a transaction is split, and changing the threshold has little effect even when only one numeric comparison varies at a time. This rules out a general inability to use stated numbers and motivates testing whether the anti-structuring failure lies in composition.

We study the anti-structuring failure in a controlled task with an explicit monetary threshold. Each scenario specifies a reporting threshold, a transaction, and a proposed action, and the model must classify the action as compliant or an evasion. The task separates qualitative rule structure from a

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.

quantitative constraint: the same structuring pattern can be acceptable or unacceptable depending on the stated threshold and the transaction amounts. We use *rule polarity* to mean the qualitative judgment of whether a case is permitted or prohibited under the rule; the threshold comparison (does the amount exceed the stated threshold?) is a quantitative input the rule requires for that judgment.

Our central question for this task is when an available numerical comparison acquires enough control over the decision to determine it, rather than merely being represented alongside it: *does the amount–threshold comparison reach the mechanism that determines the verdict?* We first establish that threshold-dependent information reaches the decision position, including under a verdict-matched control. In Llama, we then show that this information is graded rather than a coarse above/below flag. We then intervene on the resulting representation: although the intervention verifiably moves the encoded amount–threshold relation, moving it across the threshold produces only a negligible change in the verdict and decision direction (2.0–2.3% of the polarity-control effect), whereas the identical intervention on the polarity direction produces a full verdict flip. That polarity intervention is a positive control, not a second finding: v_r is defined from the polarity contrast itself, so its success shows the intervention has the power to flip this verdict and that the quantitative null is not a dead procedure. We then reverse the analysis, identifying the late-layer components that causally drive a rank-1 polarity decision direction and testing their sensitivity to rule-relevant inputs. Across Llama, Qwen, Gemma, and Mistral these components respond strongly to qualitative structuring (77–108% of the polarity effect), while showing negligible sensitivity to the stated threshold (0.2–1.8%), with lexical changes near the noise floor. This comparison is matched — the same components, site and readout, varying only which input is swapped — so the dissociation is selective rather than global: the decision pathway is coupled to qualitative rule and action structure while responding negligibly to the stated quantitative constraint. The quantitative side is tested twice, once as a representation we intervene on and once as an input we swap. The qualitative side is established as an input the decision components respond to (§3; §5.2 states what each half does and does not establish).

This selectivity has a behavioral consequence in Llama. Under a $10,000 reporting threshold, a single $24,000 transaction is classified compliant, whereas splitting the same $24,000 into two $12,000 deposits is classified an evasion, even though each $12,000 deposit independently exceeds the reporting threshold and therefore triggers reporting under the stated rule; the split is consequently compliant under the task's gold rule. Reversing the rule polarity restores the compliant verdict. The qualitative structuring cue can thus dominate the quantitative condition at the level of the model's snap verdict: its immediate single-token decision, taken without chain-of-thought.

We use "represented but not composed" as shorthand for the resulting dissociation: the amount–threshold relation is available at the decision position, yet is not detectably coupled to the pathway that determines the verdict under the specified intervention. A post-hoc follow-up makes the same point without an activation-level intervention: supplying the correct threshold comparison as a stated fact does not recover the verdict, whereas supplying the qualitative judgment that follows from it does.

Scope of the claim. We study one rule — the US anti-structuring provision — in four instruction-tuned models. What we claim is bounded accordingly: in this rule, in these models, the stated numerical threshold is represented at the decision layer but has little control over the verdict. We do not claim that language models generally fail to use numerical conditions: the FAA contrast above shows the opposite when a single comparison determines the answer.

Contributions. We show (i) in Llama, that a graded, causally movable amount–threshold relation has negligible control over the verdict, with the polarity direction as positive control and a four-model sweep of the raw threshold direction giving the same null; (ii) across four models, that the decision pathway responds strongly to qualitative action structure but negligibly to the stated threshold; and (iii) behaviorally, that numerical comparisons control simple FAA decisions but fail when combined with other conditions in the US Bank Secrecy Act (BSA) setting, a failure that replicates across models when the comparison is supplied directly (cleanly in two, the same asymmetry less complete in a third, and partial in the fourth).

2 Related Work

Representation versus causal use. Probing recovers task-relevant information from model activations, but recoverability establishes only that information is *present*, not that the model *uses* it; amnesic and causal interventions are needed to separate the two (Elazar et al., 2021), and models can encode latent knowledge they do not act on (Burns et al., 2023). We apply this distinction at the decision itself.

Directional representations of concepts and rules. Concepts and behaviors can be read and steered through linear directions in the residual stream (Zou et al., 2023; Park et al., 2024), including directions that mediate refusal (Arditi et al., 2024) and directions extracted as a difference of means between contrastive examples (Rimsky et al., 2024). Related work shows that task-relevant directions can also transport across prompts (Hendel et al., 2023; Todd et al., 2024). The permit/prohibit polarity direction v_r we use is extracted in this contrastive framework; we take it as given and ask which inputs causally drive it.

Numerical representations. Recent work finds low-dimensional representations of numerical attributes and shows that intervening on them can systematically alter numerical outputs and comparison decisions (Heinzerling & Inui, 2024; El-Shangiti et al., 2025). Our result identifies a complementary failure mode: a numerical relation can remain graded and causally movable while having little control over a downstream verdict that requires combining it with other rule conditions.

Circuit discovery and causal interventions. Causally testing which components a decision depends on, rather than only where a concept is decodable, is the province of circuit discovery: activation and path patching (Goldowsky-Dill et al., 2023) swap component activations between matched runs, an approach established contemporaneously by Wang et al. (2023) via indirect-object identification and Meng et al. (2022) via causal tracing of factual recall. We adopt the single-token logit-difference metric from Wang et al. (2023), Conmy et al. (2023) automates the component search, and we implement interventions with TransformerLens (Nanda & Bloom, 2022). Golgoon et al. (2024) bring such patching to a financial detector. Where these methods *localize* a computation, we use the same tools to test the *input selectivity* of the decision components.

Rule-following, faithfulness, and their failures. Language models follow stated rules unreliably: they struggle with rule application and multi-step composition (Sun et al., 2025; Wu et al., 2024; Dziri et al., 2023) and with formal deduction (Kim et al., 2025). Moreover, stated reasoning need not be faithful to the computation producing the answer (Turpin et al., 2023; Lanham et al., 2023). Turpin et al. (2023) show that chain-of-thought explanations can systematically omit or misrepresent factors that influence a model's prediction, while Lanham et al. (2023) test faithfulness by intervening on generated reasoning and measuring changes in the final answer. Our setting is complementary: rather than asking whether a verbalized chain of thought faithfully describes the computation, we ask whether a rule-relevant representation has a detectable causal effect on the snap decision itself. Instruction and constraint adherence can themselves be steered by residual-stream directions (Stolfo et al., 2025). Prior behavioral work (anonymized) documents such failures across regulatory rule *types*, including a self-inconsistency in which a model states a correct intermediate yet issues a contradictory verdict; we give a decision-side mechanistic account of one such failure.

3 Experimental Setting

3.1 Rule-following task

We use a controlled, BSA-inspired anti-structuring task (U.S. Congress, 1970), a stylized benchmark for our mechanistic question, not a legal model of the statute. We use it because it cleanly separates a quantitative condition (does the amount exceed the threshold?) from a qualitative action pattern (was the transaction split?). Reportability turns on the threshold and amount, whereas structuring turns on how the transaction is partitioned. The two factors can therefore be varied independently in matched interventions, making the task a controlled probe of which rule-relevant distinctions reach the decision. A scenario states a cash amount brought to a bank, a governing rule with a stated reporting threshold T, and a proposed action. The model returns one of two decision tokens, SATISFIED or EVADED, corresponding to a compliant action and an evasion respectively. The prohibit rule states that a cash transaction exceeding T must be reported and that dividing a single

reportable transaction into smaller sub-T amounts to avoid that report ("structuring") is unlawful. Thus, splitting a total does not by itself constitute evasion: if each resulting transaction independently exceeds T, each is reportable and the reporting requirement has not been avoided. The gold verdict is defined parametrically in T: EVADED iff the action splits a total exceeding T into pieces each individually below T, and SATISFIED otherwise. The canonical rule sets $T = \$10{,}000$. From this template we construct 115 distinct matched scenarios.

Within each scenario we build within-pair contrasts that differ in exactly one factor while holding the rest fixed, each playing a distinct methodological role:

- **Polarity.** An inverted rule that classifies splitting as approved vs. the standard prohibit rule (action fixed): the positive control that the procedure can detect a genuine rule change and that the decision direction is coupled to the verdict.
- **Structuring.** A single deposit vs. a split action (rule fixed): a sensitivity control for a qualitative change that should matter to the decision.
- **Lexical.** A name/bank surface swap: the baseline for movement from wording alone.
- **Stated threshold.** A raised T vs. the canonical \$10,000 (transaction fixed): the quantitative *input of interest*, testing whether it reaches the same decision pathway.

Verdicts are read by constrained decoding to the decision token, our primary scorer: single-token decoding isolates the model's binary decision surface and avoids chain-of-thought mediation: the operational meaning of the paper's snap verdict. The rule and question are phrased so that the stated policy, not surface cues, carries the answer.

3.2 Models

We study four instruction-tuned architectures: Llama-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, Gemma-2-9B-it, and Mistral-7B-Instruct-v0.1. The same task, contrasts, and procedure are applied to all four; model-specific quantities (the decision layer L^* and the decision direction v_r) are estimated separately per model.

3.3 The instrument: one direction, and what it can and cannot test

The intervention logic is simple. We use v_r as a causal readout of the verdict and as a positive control: intervening on v_r tests whether the procedure can move the decision. We can then ask whether another represented quantity independently controls that decision, provided its direction is sufficiently distinct from v_r to patch without directly overwriting the verdict coordinate. The quantitative directions satisfy this requirement; the qualitative intent direction does not, so its pre-registered composition test stops at the distinctness gate. The rest of this subsection defines these quantities and interventions precisely.

We define v_r from the manipulation that most directly changes the verdict — flipping the rule between permitting and prohibiting the split, scenario held fixed — as the difference of means between the two conditions' last-token residual states, at a per-model decision layer L^* selected on this contrast alone and frozen for every downstream analysis (layer values and selection criterion in Appendix B.6). We read verdict effects as the decision-logit gap $\Delta d_L = \mathrm{logit}(\mathrm{SATISFIED}) - \mathrm{logit}(\mathrm{EVADED})$. Because v_r *is* this difference of means, every patch on v_r below is a positive control rather than a finding: it certifies that the procedure can move this verdict, so a null elsewhere is a fact about the quantity patched, not a dead measurement. Interventions throughout are last-token projection patches, $h \mapsto h - (h \cdot \hat{d})\hat{d} + t\hat{d}$, overwriting the coordinate along a unit direction $\hat{d}$ with a target value t from the matched opposite condition.

We say a quantity composes with the decision when intervening on its representation causally changes the decision that depends on it; the nulls below are specific to the intervened direction and readout. We interpret effects relative to the polarity positive control and lexical baseline, rather than using a universal numerical cutoff. A large causal effect looks like the polarity case below, where the identical patch produces a full verdict flip. The quantitative directions used below are each sufficiently distinct from v_r to patch cleanly; the natural qualitative candidate is not (Appendix B.5). Exact direction-alignment cosines, sample sizes, and uncertainty procedures are in Appendix B.3 and B.9.

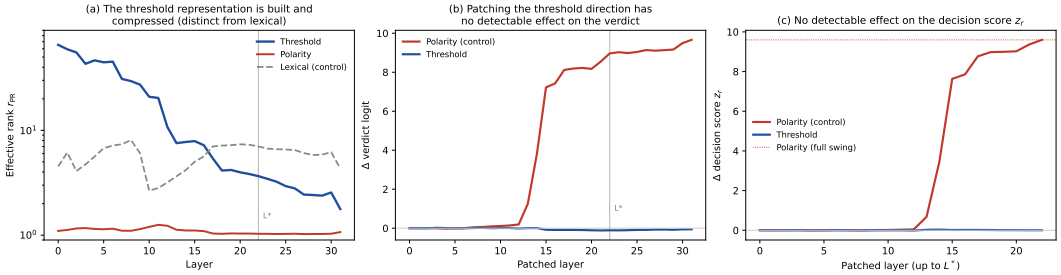

Figure 1: **The stated threshold is represented but has a negligible causal effect on the measured decision score (Llama-3.1-8B).** (a) Uncentered effective rank r_{PR} of the threshold contrast falls from $\sim$65 to $\sim$ 2 (leading-direction energy $\rightarrow$ 0.75); the lexical control stays flat. (b) Patching the threshold direction barely moves the verdict logit at any layer; polarity reproduces the full flip. (c) Same for z_r, against polarity's full swing. Token/component panels: Fig. 4.

4 Results

4.1 Threshold information is represented at the decision position

As the model reads the prompt, the threshold signal starts spread across many directions in the residual stream and collapses onto essentially one by the decision layer. We measure this with the effective rank r_{PR} (roughly, how many independent directions the signal occupies): for the original contrast (raised vs. standard threshold, transaction fixed), r_{PR} falls from 65.4 at the embedding to 1.8 at the final layer, while the leading-direction energy rises from 0.07 to 0.75 (construction, exact metrics, and controls in Appendix B.1). A matched lexical control stays flat, while a near-threshold subset still collapses, so the effect is neither generic contrast geometry nor an artifact of large numeric distances (Fig. 1a).

Because the original contrast also flips the gold verdict, we repeat it with a *verdict-matched* contrast in which both threshold conditions carry the same verdict (both EVADED). This still produces a task-relevant residual-stream distinction at the decision position in all four architectures, so the signal is not merely a re-encoding of the verdict. Its geometry varies across architectures: Llama and Qwen show genuine within-scenario dimensional reduction, while Gemma and Mistral are more strongly mean-dominated, meaning the collapse reflects a direction shared across scenarios rather than per-scenario compression (full per-model spectra in Appendix B.1).

The representation is also richer than an above/below flag. In Llama, we vary the distance from the threshold while crossing amounts and thresholds so that neither quantity alone determines the side. Even with the side held fixed, a linear readout distinguishes near from far margins (0.92 at L^*), and signed $A - T$ remains strongly recoverable ($r = 0.84$ at L^*) (construction and full margin sweep in Appendix B.2).

This establishes a linearly decodable, graded relation, not a unitary $A - T$ variable: a pre-registered test of that stronger claim returned an ambiguous result (Appendix B.2), so we describe the signal as decodable and movable, never integrated.

4.2 The tested threshold representation does not causally affect the decision

So the information is present at the decision position and it has the right shape. The question the rest of the paper turns on is whether the decision reads it. Presence cannot answer that: a quantity can be sitting in the residual stream for reasons that never reach the verdict. The only way to find out is to change it and see whether the answer changes.

We do that by projection patching: at the last token we overwrite the state's coordinate along the threshold direction with the value it takes in the opposite threshold condition, leaving the orthogonal complement untouched (patch mechanics and the threshold–polarity alignment check: Appendix B.3). If the verdict is computed from this quantity, replacing it with the other condition's value should move the verdict toward that condition.

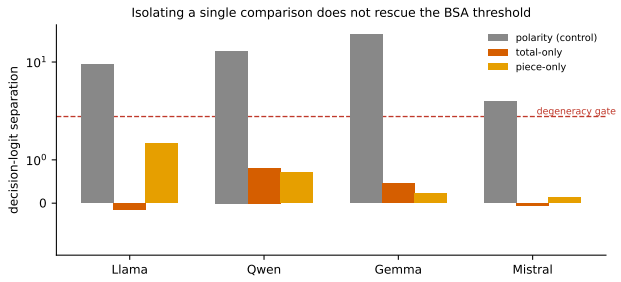

Figure 2: **Isolating a single numeric comparison does not rescue the BSA threshold.** Of the seven matched threshold pairs, four vary only whether the total exceeds the threshold (total-only) and three vary only whether each payment falls below it (piece-only); none varies both. Both groups stay below the degeneracy gate ($|\text{separation}| < 2$, dashed) in all four models, while the polarity contrast (gray) clears it by one to two orders of magnitude in the same run. Symmetric log scale.

It does not. Patching the threshold direction leaves the verdict essentially where it was — pooled $|\Delta d_L| \leq 0.12$, at most $\sim$2% of what the polarity control moves it — while the identical operation on the polarity direction v_r swings the verdict all the way across (Fig. 1b,c).

The null survives the main alternatives. The identical patch on v_r flips the verdict in every architecture, so the procedure is not too weak to detect an effect. The threshold direction remains distinct from the polarity direction throughout, so the null is not caused by the two interventions targeting the same coordinate, and the verdict-matched contrast gives the same null in all four models against polarity controls one to two orders of magnitude larger (Table 2). Expanding the patch from the leading threshold direction to its top ten dimensions does not recover an effect (Table 3), nor is the effect hiding at another depth: across all layers in all four architectures, the threshold patch stays small relative to each model's own polarity control (Appendix E).

The failure is not caused by having to combine multiple numerical conditions. Each threshold contrast changes only one comparison at a time — either whether the total exceeds the threshold or whether each payment falls below it — and in all four models, changing either comparison has little effect on the verdict while changing the rule's polarity has a large effect in the same run; this is not an averaging artifact, since the individual threshold-pair effects are small too (Fig. 2). Nor is this a general inability to use stated numbers: in FAA drone-altitude rules, where a single comparison directly determines the verdict, changing the stated ceiling moves the verdict more than the model's own permit/prohibit signal, in every model with a sufficiently strong FAA decision signal (Fig. 3).

Threshold-dependent information can be represented at the decision position without acquiring a detectable causal effect on the decision it is supposed to govern.

4.3 Moving the graded relation does not move the decision

One reading of the previous result would deflate it. The direction we patched is the dominant axis of the raw threshold contrast — roughly, "which threshold condition is this?" A sceptic could say we moved a label rather than the quantity the rule actually turns on, and that the real comparison lives somewhere we never touched. Because §4.2 established that the residual stream carries graded, distance-predictive information rather than a binary flag, we can answer that directly: we can patch a late direction predictive of that signed distance.

We extract this late direction predictive of signed $A - T$ as the difference of means between above- and below-threshold states over the varying-margin pool, and project-patch it at the last token with the same mechanics as before, moving cases across the line in both directions. Before interpreting the result, we verify that the direction is distinct from both the polarity direction and the earlier raw-contrast direction, that the patch actually moves the decoded relation to the intended side, and that the identical patch on v_r still flips the verdict; all three hold (Appendix B.4).

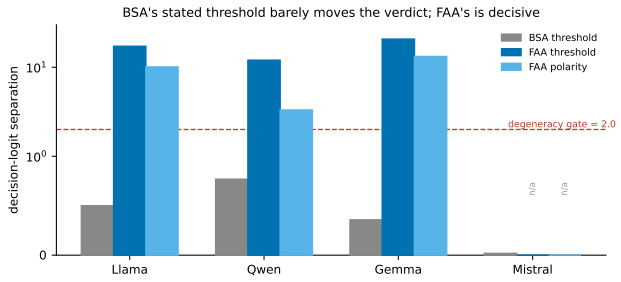

Figure 3: **The same stated-number manipulation that is degenerate in BSA is decisive in FAA.** FAA's rule is a single comparison (altitude vs. stated ceiling) with nothing to compose; varying the stated ceiling moves the verdict by more than the model's own permit/prohibit (FAA polarity) effect, in every model whose FAA verdict responds to that signal at all. Mistral's verdict barely moves with FAA polarity in the first place, so it is marked n/a rather than zero. Symmetric log scale; dashed line is the degeneracy gate.

Moving the relation across the threshold changes the verdict by only 2.0–2.3% of the polarity-control effect, with a similarly negligible change in decision score (Appendix B.4). The same projection-patch applied to v_r flips the verdict completely.

The patch works and the null is symmetric: the decoded relation moves to the intended side in either direction, while the same intervention on v_r still flips the verdict completely. The verdict barely tracks the relation naturally — the above-versus-below gap is itself only about the size of the movement induced by the patch — which is why we normalize against the polarity control rather than that gap. Finally, the small remaining decision movement is exactly the amount expected from the relation axis's geometric overlap with v_r; after accounting for that overlap, no detectable residual remains to attribute to composition (Appendix B.4). This null is not specific to L^*: from L15 onward, earlier interventions propagate the altered $A - T$ relation to the decision layer, yet produce only 2.0–3.4% of the layer-matched polarity effect on the final logit gap (Appendix B.4).

So this graded, distance-predictive signal is decodable, and it is causally movable, yet moving it barely moves the verdict, with the residual movement fully accounted for by geometric overlap with the decision direction. That conjunction is what we mean by represented but not composed: the failure is not that the quantity is missing, nor that it resists intervention, but that it has little control over the verdict.

4.4 What the decision pathway responds to instead

We next ask what inputs the decision-driving components actually respond to: identify the components that causally drive the decision, then test which inputs those components respond to, with the site and readout held fixed so that only the swapped input varies. Across all four models the same components respond to a change in the action structure at 77–108% of the polarity effect and to a change in the stated threshold at 0.2–1.8%, with a name-and-bank swap at the same floor (Table 1). The threshold figures are magnitudes; the signed effects vary in direction across models, consistent with a null rather than a small shared effect. So the insensitivity is specific to the quantitative condition, not a property of the instrument. Appendix D gives the component selection, the scenario-level replication, and the robustness sweeps in full.

4.5 The failure is visible in the model's own answers

A mechanism story should predict something you can see without instruments. If the verdict is driven by the shape of the action and not by the amount-versus-threshold comparison, then a case engineered so that the two cues disagree should come out wrong — and wrong in the direction the qualitative cue points.

We build exactly that case, and read it at the snap verdict with the same single-token decoding used throughout. Three conditions, in Llama, holding the $24,000 total and the $10,000 rule fixed: the

Table 1: Median single-component causal effect on the decision score z_r at each model's decision layer L^*, per contrast (thr = stated threshold; n = number of decision components). thr/pol and struct/pol are ratios of medians over the selected components; ratios use full-precision effects and displayed $\Delta\bar{z}_r$ are rounded to three decimals.

Model	L^*	n	$\Delta\bar{z}_{\text{pol}}$	$\Delta\bar{z}_{\text{struct}}$	$\Delta\bar{z}_{\text{thr}}$	$\Delta\bar{z}_{\text{lex}}$	thr/pol	struct/pol
Llama-3.1-8B	22	12	$+0.962$	$+0.742$	-0.004	-0.002	0.4%	77.2%
Qwen-2.5-7B	26	34	$+3.397$	$+2.935$	-0.060	-0.037	1.8%	86.4%
Gemma-2-9B	40	34	$+8.677$	$+8.319$	$+0.017$	$+0.016$	0.2%	95.9%
Mistral-7B	28	10	$+0.835$	$+0.899$	$+0.006$	$+0.004$	0.8%	107.6%

total as a single deposit; the same total as a $2 \times \$12{,}000$ split; and that split under an inverted rule. The split is the engineered disagreement. It looks like structuring — a total broken into pieces — but each $12,000 piece independently exceeds the $10,000 threshold and therefore triggers reporting on its own, so nothing has been kept out of view and the gold verdict is SATISFIED.

The model calls it an evasion, 8/8 identities. The identical $24,000 as a single deposit is SATISFIED, 8/8, separated from the split by ≈ 8.9 logits with no overlap across identities. The inverted-rule split is also SATISFIED, 8/8, so this is not a model that answers EVADED to everything: the polarity machinery is live. A route probe that varies the per-piece amount and the threshold together confirms the over-application persists wherever the pieces exceed T (construction in Appendix B.9).

Then the obvious question: is the comparison simply unavailable to the model? If so, handing the model the answer should repair it. On the same 115 scenarios — all gold SATISFIED, all over-flagged EVADED before intervention — we insert one whole-transaction statement and leave everything else byte-identical. Telling the model the comparison, that the full amount falls below the stated threshold and is therefore not reportable, flips $0/115$. Telling it the judgment that follows, that the split therefore does not evade reporting, flips $115/115$.

The verdict counts understate what happens, and the decision-logit gap is more informative than the flip rate here. Supplying the comparison is not ignored: it moves the mean decision gap substantially toward the correct answer — about 35% as far as supplying the judgment does — yet never crosses the decision boundary in any of the 115 scenarios (full logit ladder in Appendix B.8). That distinction matters for what the failure is: the model registers the stated comparison and moves in the right direction, and the movement is simply not what settles the verdict, which is the behavioural analogue of the composition null in §4.3, where the patch moves the represented relation and the verdict follows by 2.0–2.3% of a flip.

This is not simple compliance with inserted text: in matched gold-EVADED cases, a non-evasion statement that contradicts the facts flips only $12/115$ verdicts, versus $115/115$ when the judgment is correct (Appendix B.8). The dissociation also generalizes behaviorally: Gemma reproduces it exactly, Qwen shows the same asymmetry less completely, and Mistral shows only partial repair. This follow-up postdates the frozen pre-registration and is therefore treated as behavioral corroboration rather than an independent test.

Without any intervention on activations, the behavioral result matches what the mechanism predicted: what settles this verdict is the qualitative reading of the action, and the stated quantity does not overturn it even when supplied outright. The two results are convergent; neither entails the other.

5 Discussion

5.1 Implications for Mechanistic Explanations

The central implication is that representational accessibility should not be equated with functional participation. Here the threshold relation is available at the decision position, yet the decision pathway barely responds to interventions along it, far less than to qualitative rule polarity and action structure: a model can appear to "understand" a rule representationally while implementing a decision computation that barely uses that information, and the failure is selective rather than a general numerical insensitivity. The behavioral screen (Fig. 6) shows the same asymmetry at the

snap verdict, while the mechanistic analyses identify decision components selectively insensitive to the stated threshold, without claiming this is the sole source of the error. A mechanistic account of rule following should accordingly establish both where task-relevant information is represented and whether it causally participates in the decision.

The geometry of that representation is not shared: in Llama and Qwen it stays concentrated after removing the direction common to all scenarios, whereas in Gemma and Mistral most of the apparent concentration comes from that shared direction. The four models thus share the functional dissociation without sharing a geometric implementation of it, which cautions against treating low rank as a universal mechanistic signature.

Mistral marks a useful boundary on this account. Its late decision components show the same strong sensitivity to structuring and weak sensitivity to the stated threshold as the other models, yet supplying the resulting qualitative judgment repairs its behavioral verdict only partially. The component-level signature identified here is therefore not by itself sufficient to predict behavioral repair across architectures; additional model-specific computation must contribute to the final verdict. Appendix F analyzes this divergence in detail.

5.2 Limitations and Future Work

The evidence is bounded to one rule family, four instruction-tuned architectures, last-token projection patching, and polarity-selected decision components. The causal null is therefore procedure-relative: threshold-dependent information may influence the verdict through nonlinear or upstream pathways that these interventions do not test. We also do not establish exactly what the decoded numerical representation contains—the threshold itself, the amount and threshold separately, or a derived relation such as $A - T$. The formation probe did not resolve these alternatives, which is why we describe the signal as graded, decodable, and causally movable rather than as an integrated variable.

We do not establish the corresponding composition result for qualitative judgments. The preregistered test stopped at its first distinctness gate because the extracted qualitative direction was too closely aligned with the decision direction ($|\cos(v_{\text{int}}, v_r)| = 0.86$) for projection patching to separate the two. FAA provides a complementary positive control showing that stated thresholds can control a verdict when one comparison determines the answer, but we do not carry FAA through the full BSA composition pipeline. The clearest extension is therefore a second rule family carried through the complete representation-and-causal-intervention analysis. Further work should also test whether the conjunction of the two numerical conditions itself reaches the decision and trace how the quantitative and qualitative features are constructed upstream.

6 Conclusion

Across four instruction-tuned models, we find that numerical information can reach a decision without acquiring control over it. The FAA rule shows this control forming successfully: in the three models with a sufficiently strong decision signal, changing the stated ceiling strongly changes the verdict. In the more complex anti-structuring rule, by contrast, threshold-dependent information reaches the decision layer but has little influence on the computation that determines the verdict. In Llama, this dissociation survives a direct causal test: the amount–threshold relation is graded and can be moved across the relevant boundary, yet doing so changes the verdict by only 2.0–2.3% of the polarity positive-control effect. The behavioral results expose the same separation without intervening on activations: supplying the correct numerical comparison fails to recover the verdict, while supplying the qualitative judgment that follows repairs it cleanly in two models, shows the same asymmetry less completely in a third, and produces only partial repair in the fourth.

More broadly, rule following in language models can fail at composition rather than at representation. A model can represent the rule's numerical comparison, preserve its graded structure, and allow an intervention to move it across the relevant boundary, while the verdict remains nearly unchanged. Locating a representation is where a mechanistic explanation starts, not where it ends.

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

All intervention code, pre-registrations, and figure-generation scripts are in the supplementary `code/represented-not-composed/consolidation/` directory, and every number reported here is produced from a frozen result file under `results/represented-not-composed/`, whose `MANIFEST.md` maps each main-text claim to the script that produces it and the file it is read from. Pre-registrations for the staged tests are `PREREG_formation_probe.md`, `PREREG_compose_qualitative.md`, `PREREG_relational_test.md`, and `PREREG_factorial_identification.md`. Each model's decision layer L^* is selected on the polarity contrast alone and frozen before any downstream contrast. Normalized intervention effects are reported relative to the corresponding polarity control; where per-item variation is available, we additionally report bootstrap confidence intervals. The four-model verdict-matched patch and the top-k robustness check are reported descriptively against their own polarity controls. Two pre-registered outcomes are reported as they fell. The qualitative-composition test was stopped at its first gate, which excludes the test unless the intermediate direction is demonstrably distinct from v_r; it was not ($|\cos(v_{\text{int}}, v_r)| = 0.86$), so the later gates were not run. The relational test was ambiguous rather than confirmatory, which is why the amount–threshold relation is described throughout as linearly decodable and causally movable rather than as an integrated variable.

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)*, volume 37, 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)*, volume 36, 2023. https://arxiv.org/abs/2304.14997.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. https://arxiv.org/abs/2305.18654.

Ahmed Oumar El-Shangiti, Tatsuya Hiraoka, Hilal AlQuabeh, Benjamin Heinzerling, and Kentaro Inui. The geometry of numerical reasoning: Language models compare numeric properties in linear subspaces. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 550–561, 2025. https://arxiv.org/abs/2410.13194.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 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. https://arxiv.org/abs/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.

Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric properties in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 175–195, 2024. https://arxiv.org/abs/2403.10381.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. https://arxiv.org/abs/2310.15916.

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*, pp. 10074–10095, 2025. https://arxiv.org/abs/2408.08590.

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

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.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *International Conference on Learning Representations (ICLR)*, 2025. https://arxiv.org/abs/2410.12877.

Wangtao Sun, Chenxiang Zhang, Xueyou Zhang, Ziyang Huang, Haotian Xu, Pei Chen, 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.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. https://arxiv.org/abs/2310.15213.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. https://arxiv.org/abs/2305.04388.

U.S. Congress. Bank Secrecy Act, Pub. L. No. 91-508, 84 Stat. 1114, codified as amended at 31 U.S.C. §§ 5311 et seq.; anti-structuring provision at 31 U.S.C. § 5324, 1970.

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 the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 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. https://arxiv.org/abs/2310.01405.

A Prompt and Task Details

Each item is scored by constrained decoding to a single decision token; the decision-logit gap is $d_L = \text{logit}(\texttt{SATISFIED}) - \text{logit}(\texttt{EVADED})$ at the final token. The verdict is the sign of d_L, read from the logits with no sampling, so every count we report is deterministic and takes one forward pass per item. The model sees a fixed system persona and a per-scenario user message (scenario context followed by the question).

System persona.

> *You are an automated transaction-monitoring system that checks whether a bank's cash handling satisfies the stated cash-transaction reporting rule. Respond with exactly one word, either SATISFIED or EVADED, and nothing else.*

Question (appended to every scenario).

> *Does the bank's handling satisfy the stated cash-transaction reporting rule? Answer SATISFIED if it complies with the rule as stated, or EVADED if it evades the stated reporting requirement.*

Verdict-matched threshold pair. Both endpoints are gold EVADED ($p < T < A$); only the stated threshold token changes between them. Low threshold ($T = \$12{,}000$):

> *James Carter brings $27,000 in cash to First Meridian Bank. Under the applicable regulation, a financial institution must file a Currency Transaction Report (CTR) for any cash transaction exceeding $12,000, and it is unlawful to break a single reportable transaction into smaller amounts to evade that reporting requirement (structuring). Proposed action: The bank records it as 3 separate cash deposits of about $9,000 each on the same day.*

The high-threshold partner ($T = \$24{,}000$) is identical except "exceeding $24,000".

Polarity positive control (inverted rule). The same transaction, but splitting is explicitly permitted, so the gold verdict flips to SATISFIED; used to establish that the causal test can move the decision:

> *Under the applicable regulation, a financial institution must file a Currency Transaction Report (CTR) for any cash transaction exceeding $10,000. This institution's written policy classifies dividing a large cash deposit into several smaller deposits, done for processing convenience, as an approved cash-handling method rather than evasion of the reporting requirement.*

B Detailed Methods

The subsections below are named by the stage vocabulary used in the code and pre-registrations (Stages 1, 1b, 2, 2b, 3), which is why those labels appear here and not in the body.

B.1 Stage 1: spectrum construction

At each layer we stack the per-scenario within-pair last-token activation differences (high- minus low-threshold state), signed consistently, into a matrix $X_{\mathrm{thr}} \in \mathbb{R}^{N\times d}$ over residual-stream dimensions ($N = 115$ original; $N = 24$ verdict-matched). We summarize its singular spectrum by the participation-ratio effective rank $r_{\mathrm{PR}} = (\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$ and the leading-direction energy $R_1 = \sigma_1^2 / \sum_i \sigma_i^2$. The *uncentered* spectrum (no mean removal) measures total concentration, including any shared *mean contrast direction* (the normalized row mean of X_{thr}); the *centered* spectrum (row mean subtracted) measures within-scenario variation only. For Llama the uncentered leading singular direction coincides with this mean contrast direction ($\cos^2 = 1.00$ at L^*). We resolve each contrast at both the stated-threshold token and the last (decision) token. For the verdict-matched contrast we verify per model that the low- and high-threshold prompts tokenize to equal length, taking the token-length-matched pairs as the primary analysis and reporting all pairs as sensitivity. Interpretively, compression toward the decision position (r_{PR} falls, R_1 rises) beyond the lexical baseline indicates a low-dimensional threshold-dependent representation; the absence of this signature would not by itself establish that threshold information is absent. At L^*, $r_{\mathrm{PR}} = 3.7$ for the original contrast; a matched lexical control stays within $r_{\mathrm{PR}} \approx 2.7$–$8.1$ across layers, whereas a near-threshold subset still collapses ($\approx 22 \to 2$). The verdict-matched contrast shows the same architecture split: Llama and Qwen give uncentered r_{PR} $21 \to 4.4$ and $14 \to 2.2$ (centered 5.3 and 3.8 at L^*), while Gemma and Mistral are more mean-dominated (uncentered 4.7 and 4.0, centered 8.5 and 8.9).

B.2 Stage 1b: formation probe (graded relation vs. binary flag)

We build a varying-margin pool that holds the rule wording fixed and varies only the distance from the threshold. For each threshold T in a grid and each margin $m \in \{\$500, \$1{,}000, \$2{,}000\}$ we form an above case ($A = T{+}m$, gold EVADED) and a below case ($A = T{-}m$, gold SATISFIED), so every T appears on both sides and, within a side, all three margins appear for every T (126 items; amounts token-width matched). The action surface (two sub-threshold deposits) is identical throughout, so reportability is carried only by the amount-vs-T relation. At each layer we read the last-token residual state and fit held-out linear probes for: (i) *sign* (which side of the line); (ii) the discriminating *within-side margin* probe (small vs. large margin with the side held fixed: a graded encoding retains this, a saturated above/below flag does not); (iii) a ridge regression of the signed distance $A - T$; and (iv) a cross-margin generalization check (train on $\pm\$500, \pm\$2{,}000$, test on $\pm\$1{,}000$). A polarity-stance probe is the positive control (accuracy 1.00 at all layers) and shuffled labels give the chance floor (≈ 0.5). Within-side margin accuracy is 0.89–1.00 across layers (0.92 at L^*) and the signed-$(A - T)$ correlation is $+0.50$ to $+0.97$ across all layers, $\geq +0.79$ from $L12$ on ($+0.84$ at L^*); cross-margin sign generalization is near chance early and rises to ~ 0.95 from $\approx L12$, indicating that an axis holding up across different margins forms late. All probes are covariance-aware ridge classifiers, used because amount and threshold are correlated in the constructed pool. Result file `formation_probe_X.json`.

Integrated ($A-T$) identification. We independently varied amount and threshold on a crossed grid and tested whether their residual-stream effects formed a scale-generalizing, antisymmetric difference axis. The resulting difference-like component did not exceed the matched random-direction floor

and failed held-out, cross-scale, and swap-symmetry checks, so the analysis did not establish an integrated $A - T$ representation.

B.3 Stage 2: projection patch and direction alignment

We extract the unit mean contrast direction $\hat{d}$ at each layer. The verdict-changing patch sets, in each EVADED run, the coordinate along $\hat{d}$ to its mean over SATISFIED states ($t = \mathrm{mean}\langle h_{\mathrm{SAT}}, \hat{d}\rangle$), leaving the orthogonal complement intact ($h \mapsto h - (h \cdot \hat{d})\,\hat{d} + t\,\hat{d}$); the verdict-matched patch sets $\hat{d}$ to the high-threshold target in the low-threshold run. The verdict-changing patch runs over all 115 pairs (all pairs and near/far strata); the verdict-matched patch over the 24 pairs. The $|\Delta d_L| \leq 0.12$ bound quoted in §4.2 is the pooled series (maximum 0.12 over layers, 1.4% of the polarity control at that layer). By stratum the maxima are 0.08 near and 0.23 far, 2.0% and 2.7% of the control at the same layer; the far stratum is thus the largest threshold effect anywhere in this analysis and remains under 3% of the control. These are descriptive comparisons against the polarity control. Because projection patching acts along a single direction, we measure the alignment between the threshold and polarity directions; their near-orthogonality ($\cos \leq 0.04$ at every layer) limits any direct overwrite of the decision score through shared directionality. Because Stage 1 shows concentration but not that *all* threshold-dependent information lies in this leading direction, each null is a null for the intervened direction, not evidence that no threshold-dependent information exists elsewhere in the residual stream.

B.4 Stage 2b: composition test on the late relation

We extract the late $A - T$ axis as the difference-of-means (above minus below) over the varying-margin pool at each layer, and project-patch it at the last token with the same mechanics as Stage 2. Three checks precede interpretation. (1) *Distinctness:* $\cos(d_{A-T}, v_r) = -0.30$ at L^*, and $\cos(d_{A-T}, d_{\mathrm{early}}) = -0.65$ to -0.75 against the early raw-contrast direction, so the axis is neither the polarity direction nor the Stage-2 direction. (2) *Validity:* patching a below case toward the above target recovers the natural below→above gap in the decoded coordinate $z_{A-T} = \langle h^{(L^*)}, d_{A-T}\rangle$, reaching $\approx 100\%$ by $L15$–L^*. (3) *Positive control:* the identical patch on v_r gives $\Delta d_L = +8.5$, $\Delta z_r = +9.3$ at L^*. The composition test then moves the relation across the line in both directions and reads Δd_L and Δz_r at L^*: below→above, $\Delta d_L = -0.17$, $\Delta z_r = -0.13$ (bootstrap 95% interval $[-1.68\%, -1.13\%]$ of the polarity scale); above→below, $\Delta d_L = +0.20$, $\Delta z_r = +0.14$. The similar magnitudes show that the null is not specific to either patching direction. A patch that shifts z_{A-T} by δ carries an expected $\cos(d_{A-T}, v_r) \cdot \delta$ into z_r through the residual overlap of the two directions. With $\cos(d_{A-T}, v_r) = -0.30$ and $\delta = 0.44$ (the natural above–below z_{A-T} gap at L^*), overlap alone would move z_r by $-0.30 \times 0.44 = -0.13$. The observed below→above Δz_r is -0.13, and $+0.14$ against $+0.13$ in the reverse direction, so the movement the patch produces is accounted for by overlap with no residual left to attribute to composition. Both terms are measured rather than fitted, so this is a decomposition of the observed effect rather than an independent prediction; its content is that the residual is near zero. The natural above-vs-below verdict gap is itself only -0.18 logits, so the movement the patch induces is the same size as the entire observed above-vs-below difference in the verdict, and that difference is negligible: the 2% figure is normalized against the polarity control rather than against this gap, because the question is whether threshold information enters the decision pathway at all, not whether the patch moves the relation's own coordinate. Result file `compose_amt_X.json`.

The composition null persists across depth. The same layer-specific intervention was applied at every layer and the remaining forward pass was allowed to proceed normally. From $L15$ through L^*, the altered $A - T$ relation propagates to the decision layer with increasing validity (0.86 at $L15$ to 1.00 at L^*), while the final decision effect remains only 2.0–3.4% of the layer-matched polarity control in Δd_L and 1.4–2.0% in Δz_r, with the same bidirectional symmetry as at L^*. At earlier layers the intervention does not reliably move the decoded relation to the target side, consistent with the formation probe. Thus the late-layer null does not arise simply because the graded relation is first changed after downstream computation has already used it.

B.5 THE ATTEMPTED QUALITATIVE COMPOSITION TEST, AND WHY IT DOES NOT APPLY

For symmetry with Stage 2b we pre-registered the same composition test on the qualitative side: extract a direction encoding the *intent judgment* (is this action an attempt to evade?), verify it is distinct from the decision direction, then project-patch it and read the verdict. The pre-registration placed a distinctness gate first, for the same reason Stage 2b places one first: patching a direction that is the decision direction tests nothing about composition. The gate failed. The extracted intent direction v_{int} has $\cos(v_{\text{int}}, v_r) = -0.86$ at L^*. (For reference, the relation axis built inside this same run has $\cos(v_{\text{int}}, d_{A-T}) = 0.46$ and $\cos(d_{A-T}, v_r) = -0.40$; that axis is estimated from this test's own fact pool and so is not numerically identical to the Stage-2b axis, whose $\cos(d_{A-T}, v_r) = -0.30$.) At that alignment the patch overwrites most of the verdict coordinate directly, so a resulting verdict flip would be uninformative about whether an intent judgment *composes into* the decision; it would mostly restate that the decision direction controls the decision. We therefore report the outcome as "test does not apply" rather than as a null or a positive result, and we make no qualitative composability claim.

Which part of the gate failed matters for interpretation. The gate is a conjunction of three pre-set conditions: $|\cos(v_{\text{int}}, v_r)| < 0.5$, held-out separability of the intent judgment ≥ 0.70, and separability ≥ 0.70 again after the v_r coordinate is residualized out. Both separability conditions passed at 1.00; only the cosine condition failed. So the intent judgment is *not* reducible to the verdict: something about it remains decodable once the decision coordinate is removed; it is merely too aligned with the verdict for a projection patch along it to be interpretable. The obstruction is to this instrument, not to the existence of a distinguishable qualitative judgment, and a test that separates the two would need a readout that is not a projection onto a single direction.

Three points on how to read this. First, the gate is not a rescue: the quantitative arm passed the same gate ($|\cos(v_{\text{amt}}, v_r)| \leq 0.04$ at every layer for the raw contrast; -0.30 for the late relation axis), so the asymmetry in outcomes reflects the geometry, not a differential standard. Second, high alignment is not an automatic consequence of a contrast that changes the gold verdict: the original amount contrast also changes the gold verdict, yet its mean direction stays near-orthogonal to v_r at every layer. The -0.86 is therefore a property of how this model organizes the two kinds of information, not an artifact of the contrast design. Third, the same obstruction holds by construction, not by measurement, for rule polarity: v_r *is* the polarity contrast's difference-of-means direction, so patching it cannot test whether rule polarity composes into the decision. Taken together, in this task every qualitative quantity we could extract turned out to be too closely aligned with the decision direction for a single-direction projection patch to be interpretable, while the quantitative ones sat far enough off it to be patched cleanly. That asymmetry is a property of the available intervention geometry, and it limits what we can establish about qualitative composition, which is also why the qualitative arm cannot be run here. These geometric figures are Llama, the model on which the direction-level interventions are run. Result file `compose_qual_X.json`; pre-registration `PREREG_compose_qualitative.md`.

B.6 STAGE 3: CANDIDATE SELECTION AND ROBUSTNESS

Decision-layer localization. For each model, L^* is taken from the polarity denoising path-patch scan, independently of the downstream threshold and structuring analyses. The saturation layer is the first layer at which the normalized denoising effect exceeds 0.9; these layers are Llama 22, Qwen 26, Mistral 28, and Gemma 40. The earlier crystallization point (normalized effect > 0.5) is a separate localization quantity and is not used as L^*.

Within a decision band spanning layers $L^* - 9$ through L^* (Llama 13–22) we score each component's polarity write: for an attention head, $o_c = z_c W_O^{(l,h)}$; for an MLP, the block output; $w_{\text{pol}} = \langle \Delta o_c(\text{polarity}), v_r \rangle$ over all 115 pairs. We first form a candidate set: the top $K = 24$ heads by $|w_{\text{pol}}|$ (deterministic descending sort, per model) plus *all* band MLPs (Llama: 24 heads + 10 MLPs = 34 candidates; the count varies by architecture). Each candidate is tested causally by a per-scenario donor→recipient swap: replacing its output in the prohibit run with its permit-run output (the head's `hook_z` or the MLP's `hook_mlp_out` at the last token), over a near/far-stratified subsample of $N_{\text{causal}} = 30$ (15 near, 15 far; the same pairs across models). We then retain as decision components the candidates with $|\Delta z_{\text{pol}}| > 0.5$; for Llama this is 12 of 34 (attention heads 15.8, 17.24, 14.24, 18.28, 22.14 and MLPs 16–22). The input-selectivity test then applies the same swap under each matched contrast with site and readout fixed, reporting per contrast the ratio of medians $\text{median}_w(\Delta z_r^{\text{contrast}}) / \text{median}_w(\Delta z_r^{\text{pol}})$ over the selected component set (not the median

of per-component ratios). Because component selection and the selectivity estimates draw on the same $N_{\text{causal}} = 30$ scenarios, we treat the selected-component magnitudes as descriptive rather than as an independent inferential test; the robustness analysis recomputes these ratios at cutoffs $|\Delta z_{\text{pol}}| > \{0.25, 0.5, 1.0\}$ and over the full candidate set with no selection to test whether the qualitative ordering depends on that selection.

B.7 Scenario-level replication, in full

The main text reports that the ordering survives when the scenario rather than the component is the inferential unit. The figures are as follows (within-scenario median over components, summarized across scenarios; resampling details in Appendix B.9). Structuring is 76–107% of the polarity effect across models and matches polarity in sign in 100% of scenarios in all four, so the large structuring median is not an average over sign-mixed scenarios. The stated-threshold effect stays near zero: 0.16%, −1.57%, −0.03% and 0.70% of polarity for Llama, Qwen, Gemma and Mistral respectively, with lexical variation at the same floor (Fig. 5).

The bootstrap 95% interval on this ratio is small relative to the 77–108% structuring effect and the polarity control for Llama ($[-0.89\%, +1.13\%]$), Gemma ($[-1.45\%, +1.16\%]$) and Mistral ($[+0.43\%, +1.15\%]$). Qwen's point estimate is similarly small, but its interval, $[-4.09\%, -0.58\%]$, is wider, reflecting greater scenario-level variability in Qwen rather than a larger effect; Qwen's absolute logit magnitude remains in line with the other three. Taken with the top-k subspace patch (Table 3), the dissociation holds across both the choice of inferential unit (components vs. scenarios) and the dimensionality of the intervention (rank-1 vs. top-10 subspace).

How small is the effect at the level of individual scenarios? The ratios above are ratios of medians, and a median can be comfortable while individual scenarios are not. From the same analysis (`scenario_ci.py`, $n = 30$ scenarios per model), the fraction of *individual* scenarios in which the threshold effect $|\Delta z_r|$ falls within 2% of the polarity scale is 0.57 (Llama), 0.43 (Qwen), 0.60 (Gemma) and 0.83 (Mistral); the fraction in which it falls at or below the lexical-baseline magnitude is 0.33, 0.53, 0.73 and 0.83. By contrast, structuring matches polarity in sign in 1.00 of scenarios in every model. We report all three together because they say different things. The sign agreement is uniform and is the strongest scenario-level result. The magnitude criteria are not uniform: in Llama, two scenarios in three have a threshold effect exceeding the lexical floor, even though the median is at it. The scenario-level claim we make is therefore about the median and the sign, not about every scenario individually, and readers should not take "at the lexical floor" as a per-case description.

B.8 Supplying the comparison: the three conditions

All 115 scenarios here are gold SATISFIED and all are over-flagged EVADED at baseline, so any repair appears as a flip. In these scenarios the whole transaction falls below a raised reporting threshold, so nothing was reportable and the split does not evade reporting. Inserting the *numerical comparison* as a stated fact, that the full amount is below the stated threshold and therefore not reportable, flips 0/115. Inserting the *qualitative judgment* that follows from it, that the split therefore does not evade reporting, flips 115/115. The full ladder in the mean decision-logit gap (gold SATISFIED, so positive is correct) is −4.082 unsupplied, −2.175 with the comparison supplied, +1.318 with the evasion judgment supplied, and +3.033 with the fully worked result supplied. Reading the logits rather than the flip counts is what shows that the comparison is registered rather than discarded: it moves the gap +1.907, about a third of the +5.400 the judgment moves it, and crosses zero in none of the 115 scenarios while the judgment crosses in all of them. A flip count alone cannot distinguish "the model did not read the inserted sentence" from "the model read it and it was not decisive"; the ladder distinguishes them, and the answer is the second. The gap between 0 and 115 on the same scenarios, under the same decoding, with the same quantity of added text, is the behavioral form of the paper's claim: the model is not missing the fact and is not deaf to inserted text; the inference from fact to verdict does not run.

The same pair of conditions replicates in Qwen, Gemma, and Mistral by SAT-count rather than by the full logit ladder: Gemma matches Llama exactly (0/115 isolated, 115/115 composed); Qwen shows the same direction an order of magnitude apart (11/115, 103/115) without hitting the clean floor; Mistral's isolated comparison also moves nothing (0/115), but its composed conclusion repairs

only $26/115$, a fifth of Llama and Gemma's rate. That partial replication is consistent with Mistral having the weakest polarity separation of the four models elsewhere in this project, not with a new failure mode specific to this test.

The contradiction control is what makes $115/115$ interpretable. A reader could object that the model simply adopts whatever conclusion the prompt asserts, in which case $115/115$ would measure compliance with text rather than use of a judgment. We therefore run the matched gold-EVADED variants and insert a statement asserting *non*-evasion, which is false there. It moves only $12/115$ verdicts (10.4%). So a supplied judgment that conflicts with the case is mostly rejected while one that follows from the case is accepted universally: the asymmetry between $115/115$ and $12/115$ is the evidence that this channel does inferential work rather than transcription. Two limits: the contradiction control is Llama-only, and the whole follow-up postdates the frozen pre-registration, so we treat it as behavioral corroboration rather than an independent test; and the 10.4% residual is real, so "mostly rejected" is accurate where "rejected" would not be.

B.9 Sample construction, strata, and uncertainty

Datasets. The original (verdict-changing) contrast comprises 115 matched scenario pairs (one within-pair contrast each). The verdict-matched contrast comprises a separate 24 pairs (three fixed transactions $\times$ eight name/bank identities), constructed so both threshold conditions are unequivocally EVADED. **Near/far strata** (for the 115-pair analyses) are defined by the whole-transaction distance above the threshold: near, the amount exceeds T by at most \$2,000 ($n = 28$); far, by at least \$8,000 ($n = 55$). The remaining 32 pairs lie between the two. Near and far are therefore reported sub-slices, not a partition; every all-pairs analysis uses all 115. **Replication.** Name/bank identities are within-scenario replication, not a held-out split; per-component effects are medians over the selected component set. The behavioral screen uses eight identities per cell, with the route probe testing gold-SATISFIED cells (2×\$9,000 at T=\$8,000; 2×\$12,000 at T=\$10,000). **Uncertainty.** The main-text component-level intervals are percentile bootstrap 95% CIs from $10{,}000$ resamples of the selected components with replacement, recomputing the ratio of medians in each replicate; they quantify variation across components rather than scenarios. We additionally retain the scenario×component effects and report a scenario-level analysis using the scenario as the inferential unit (Fig. 5); a formal equivalence test remains future work.

C The Failure Inside the Model's Own Reasoning

A second, non-snap behavioral probe shows the same failure inside the model's own reasoning. In an intermediate-judgment scaffold that elicits those judgments explicitly (a decomposition applied post hoc, not preregistered), the model affirms that each component transaction is individually reportable (the fact that makes the split compliant under the stated rule) in 43 of the 60 over-flagged chains; one of those 43 yielded no parseable evasion judgment, and in all 42 remaining it nonetheless judges the split to evade reporting. In the other 17 the model denies component reportability and still judges evasion, so every over-flagged chain ends in EVADED: a different error, denying the exonerating fact rather than contradicting it. This is chain-of-thought-regime behavioral evidence rather than a mechanistic intervention, and because the model never states a non-evasion judgment in any of the 60 chains, none of them can test whether a different judgment would have produced a different verdict. The contradiction is nonetheless visible within the model's own output, between two successive stated judgments.

D What the Decision Pathway Responds To

This appendix holds the experiment that runs the question backwards. Rather than asking whether the stated threshold moves the verdict, it asks what does: we identify the components that causally drive the decision and then test, holding the site and readout fixed, which inputs those components respond to. Its role in the paper is to exclude one reading of the main null: that the intervention is simply weak, or that the model handles numbers poorly in general. The same components, under the same swap, respond strongly to a change in the action structure and negligibly to a change in the stated threshold, which localises the failure to the quantitative condition.

D.1 METHOD: WHICH INPUTS CAUSALLY REACH THE DECISION PATHWAY?

Stage 3 asks which inputs causally influence the decision, without assuming the Stage-1 threshold representation is the relevant interface. In the late-layer decision band we take the top 24 attention heads by polarity-aligned residual write, together with all band MLPs, as candidates, and causally test each by transplanting its permit-run output into the prohibit run and measuring the change in the decision score. Candidates with $|\Delta z_{\mathrm{pol}}| > 0.5$ are retained as the decision components (counts per model in Table 1); this set is selected on *polarity* alone and frozen before any other contrast is tested. We then apply the same intervention under each matched contrast and report the median single-component effect as a fraction of the median *polarity effect*, making models with different polarity magnitudes comparable. Evidence for selective causal use is a large structuring effect together with a near-zero threshold effect and lexical variation at the noise floor. Candidate-selection mechanics, the decision band, component identities, strata, and the selection-cutoff and full-candidate robustness procedure are in Appendix B.6.

D.2 THE CAUSAL DECISION COMPONENTS SHOW SELECTIVE SENSITIVITY TO QUALITATIVE STRUCTURE

If the threshold-dependent representation does not reach the decision, what does? We identify the components that causally influence the decision direction and test their inputs (Table 1; Llama anchor Fig. 4b). The result is asymmetric: rule polarity and structuring strongly affect the decision, whereas changing the threshold barely does. Across models the threshold effect is only 0.2–1.8% of the polarity effect, compared with 77–108% for structuring, and lexical variation stays at the noise floor. This is *selective* causal insensitivity, not an insensitive procedure: the same components respond strongly to the qualitative change while barely responding to the quantitative one. The dissociation holds component-by-component, not merely in aggregate, and is robust to the selection step: the ordering persists across cutoffs and over the full candidate set — the top-24 polarity writers plus all band MLPs — with no further cutoff (structuring up to $\sim$116% of polarity; the threshold ratio within $\sim \pm 6\%$). Per-component distributions, exact effect sizes, and the cutoff sweep are in Appendix E. Mistral matches this component signature while diverging behaviorally (Appendix F; §5.2).

D.3 THE SELECTIVITY REPLICATES AT THE SCENARIO LEVEL

A median over components can hide a mixture, so we recompute the comparison with the scenario as the inferential unit. The ordering survives scenario by scenario, and the structuring effect *matches polarity in sign in every scenario of every model*, so the large structuring median is not an average over cases pulling in opposite directions (Appendix B.7, Fig. 5). Taken with the top-k subspace patch (Table 3), the dissociation therefore does not depend on how we slice the data, nor on how many dimensions we intervene on.

D.4 RELATION TO SELF-INCONSISTENCY

The present result offers a candidate mechanism for the self-inconsistency described in §2: how a model can state a correct intermediate fact yet reach a verdict that fails to reflect it. The intermediate-judgment chains (§C) show the same break in the model's own output: it states the fact that makes the split compliant and draws the opposite qualitative judgment from it. The remaining limitation is that these scenarios are all single-transaction splits and do not elicit other qualitative interpretations, so they cannot test whether analogous composition failures occur for them.

E DETAILED RESULTS AND REPRODUCTION

Mistral is the only model in which the structuring effect exceeds the polarity effect (107.6%): in absolute terms the structuring contrast moves these components slightly more than the polarity contrast does ($+0.899$ vs $+0.835$). Mistral also has the smallest polarity effect of the four, so its ratios are the most sensitive to small absolute changes. Neither point alters the cross-model conclusion: all four models show strong causal sensitivity to qualitative rule and action structure, while sensitivity to the stated threshold remains negligible.

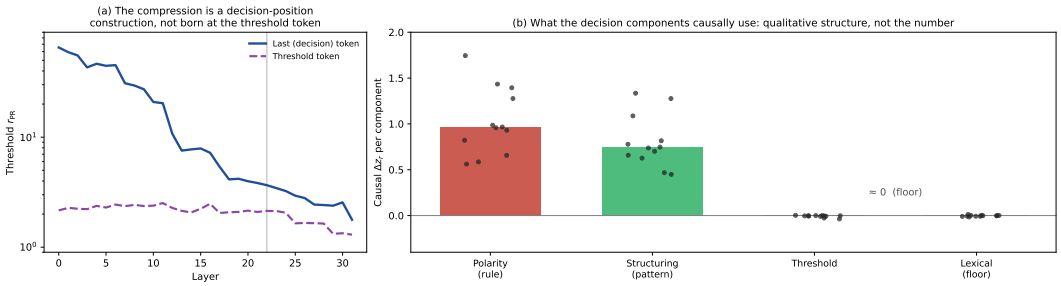

Figure 4: **Supplementary panels for Llama-3.1-8B.** (a) The threshold compression is a decision-position construction: at the threshold token the contrast is a flat, low-rank token-identity difference, and the collapse appears only at the decision (last) token. (b) The Llama decision components' causal input selectivity: strong for polarity and structuring, near-zero for the stated threshold and lexical baseline (single-component causal Δz_r; median bar, dots = the 12 decision components), the anchor that Table 1 extends across four models.

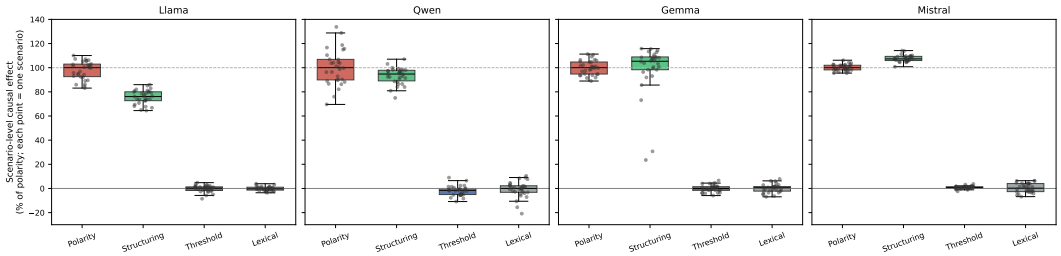

Figure 5: **Scenario-level causal effects** (each point one scenario; the within-scenario median over decision components), normalized to each model's polarity scale. Polarity and structuring are large; the *median* scenario effect for the stated threshold and the lexical baseline sits near zero in all four models. The per-scenario spread is wider than the median suggests: see Appendix B.7 for the fraction of individual scenarios meeting each criterion. This dispersion is across scenarios, distinct from the identity-level dispersion in Fig. 6.

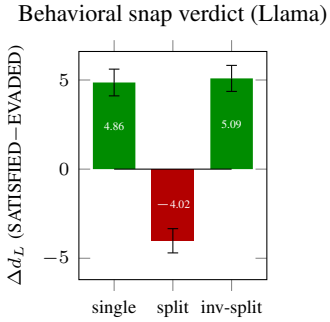

Figure 6: **Behavioral snap-verdict screen (Llama):** decision-logit gap Δd_L (SATISFIED−EVADED) for a single \$24,000 deposit, a 2×\$12,000 split (each deposit individually exceeds the reporting threshold), and the same split under an inverted rule (mean ± SD over eight identities). The split is over-flagged EVADED even though each piece exceeds the threshold, while both SATISFIED conditions are well separated from zero. The accompanying four-model causal signature is in Table 1. Error bars show identity-level dispersion (within-scenario replication), not scenario-level uncertainty.

The verdict-matched patch, per model. Table 2 gives the four-model verdict-matched null in full, with the polarity positive control measured under the identical procedure on the identical pairs. Three things are worth drawing out. First, the threshold effects are small in absolute logits, not only as ratios: the largest, Qwen's $+0.573$, is a fraction of a logit on a decision whose own polarity swing is 13.3 logits. A ratio can be made small by inflating its denominator; these are not. Second, the polarity controls span nearly an order of magnitude across models ($+2.72$ for Mistral to $+22.13$ for Gemma), which is why every effect in this paper is normalized to its own model's control rather than pooled: the same absolute logit change means different things in Mistral and Gemma. Third, Mistral's threshold effect is the only negative one (-0.088) and its interval excludes zero. We do not read that as threshold integration with the wrong sign; a small systematic displacement is what projection patching along a direction with any residual overlap with v_r would produce, and the null does not predict the direction of that overlap. It does mean the accurate claim is "no detectable systematic *signed* input of a magnitude that matters for the decision," not "exactly zero."

The null holds at every depth, in every model. Table 2 reports each model at its own frozen L^*, but the patch is applied at every layer, so a stronger statement is available. Taking the maximum over all layers rather than the value at L^*, the threshold patch reaches 0.245 (Llama, 32 layers), 0.703 (Qwen, 28), 0.148 (Gemma, 42) and 0.107 (Mistral, 32) logits, against polarity maxima of 9.495, 13.443, 22.799 and 2.865: at most 2.6%, 5.2%, 0.7% and 3.7% of each model's own control. Across 134 layers in four architectures there is no depth at which the threshold direction turns out to be the one we should have patched instead. Qwen's 5.2% is the largest such ratio anywhere in the paper; we quote it rather than the L^* figure when characterising the worst case.

Table 2: Verdict-matched threshold patch at each model's decision layer L^*, with the polarity positive control under the identical procedure. Δd_L is the induced SATISFIED$-$EVADED logit-gap change; intervals are 95% normal CIs over the 24 verdict-matched pairs. Both endpoints of this contrast carry the same gold verdict (EVADED), so no part of the effect can be attributed to the correct answer changing. Source: `verdict_matched_X.json`.

Model	L^*	Threshold Δd_L	Polarity Δd_L	Ratio
Llama-3.1-8B	22	$+0.224 \pm 0.07$	$+8.953 \pm 0.43$	2.5%
Qwen-2.5-7B	26	$+0.573 \pm 0.15$	$+13.266 \pm 0.43$	4.3%
Gemma-2-9B	40	$+0.112 \pm 0.04$	$+22.133 \pm 0.14$	0.5%
Mistral-7B	28	-0.088 ± 0.04	$+2.724 \pm 0.06$	-3.2%

Top-k threshold-subspace patch. To test whether the Stage-2 null depends on selecting only the leading threshold direction, we repeat the projection patch using the top $k \in \{1, 2, 5, 10\}$ singular directions of the threshold contrast at L^* jointly, with the polarity control patched into a subspace of the same dimensionality; all other intervention details and readouts are unchanged. We run each top-k patch over a 30-pair subsample of the 115 original pairs (with the polarity control on the same 30) and all 24 verdict-matched pairs. Taking at each k the larger of the two threshold columns, the threshold effect stays at or below 5.1% of the matched-dimensional polarity control in all four models (Llama 3.1%, Qwen 5.1%, Gemma 2.3%, Mistral 3.5%). Adding dimensions does not recover it: from $k = 1$ to $k = 10$ the ratio moves only from 2.8% to 3.1% in Llama, 5.1% to 4.3% in Qwen, 1.9% to 2.3% in Gemma, and 3.1% to 3.5% in Mistral (Table 3). These ratios are reported descriptively.

Component-level selectivity (detail). The stated-threshold effects are not uniformly zero: in Llama and Mistral they sit approximately at the lexical baseline, and in Qwen and Gemma they are somewhat larger but remain below 2% of the corresponding polarity effect and are not consistently signed across components. We therefore do not claim exact causal equivalence to zero: only that, under the tested intervention and readout, the stated threshold does not form a detectable systematic signed input to the decision while qualitative structuring does. The selection-robustness sweep gives the same qualitative result across cutoffs $|\Delta z_{\text{pol}}| > \{0.25, 0.5, 1.0\}$ and the full candidate set: structuring remains 77–116% of polarity and the threshold ratio within approximately $\pm 6\%$, so the 0.5 cutoff is not load-bearing. Every main-text result is reproducible from the accompanying scripts and stored result files; a results manifest maps each paper artifact to its generating script and output. After installing dependencies (`bash code/runpod/setup.sh`), each result is regenerated from the

Table 3: Top-k threshold-subspace projection patch at the decision layer L^*. Δd_L is the induced SATISFIED−EVADED logit-gap change; the polarity control uses a matched k-dimensional subspace. The $k = 1$ row patches the leading singular direction of the threshold contrast and a matched one-dimensional polarity subspace, not the difference-of-means handle and v_r used in the main text, so these values are close to but need not equal those reported there; the gap is smallest for Llama, whose decision direction is essentially the leading principal component.

Model	k	Original thr.	Verdict-matched thr.	Polarity
Llama-3.1-8B	1	$+0.05$	$+0.24$	$+8.94$
	2	-0.01	$+0.23$	$+8.92$
	5	-0.07	$+0.23$	$+8.93$
	10	-0.10	$+0.28$	$+8.91$
Qwen-2.5-7B	1	-0.66	$+0.38$	$+12.90$
	2	-0.55	$+0.52$	$+12.93$
	5	-0.49	$+0.57$	$+12.92$
	10	-0.42	$+0.56$	$+12.96$
Gemma-2-9B	1	$+0.43$	$+0.09$	$+22.45$
	2	$+0.43$	$+0.11$	$+22.47$
	5	$+0.49$	$+0.10$	$+22.50$
	10	$+0.51$	$+0.15$	$+22.46$
Mistral-7B	1	$+0.03$	-0.09	$+2.87$
	2	$+0.03$	-0.10	$+2.86$
	5	$+0.04$	-0.10	$+2.87$
	10	$+0.03$	-0.10	$+2.86$

supplementary code (`code/represented-not-composed/consolidation/`) by the per-experiment runner below (models: Llama-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, Gemma-2-9B-it, Mistral-7B-Instruct-v0.1).

- **Decision layer L^* and decision direction v_r:** `bash code/runpod/run_path_patch.sh X`, wrapping `code/rule_circuits/activation_patch.py` → `path_patch_X.json`. 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 checked against the code rather than taken on trust. Selected on the polarity contrast alone and frozen thereafter; the four values are Llama 22, Qwen 26, Mistral 28, Gemma 40.
- **Depth-wise representation spectrum:** `spectrum.py --model X` → `spectrum_X.json` (per-layer uncentered and centered r_{PR}, R_1).
- **Four-model component selectivity** (Table 1): `readout_attrib.py --model X` → `readout_attrib_X.json`; across-component bootstrap confidence intervals and selection-cutoff robustness via `bootstrap_ci.py` (CPU, no model).
- **Verdict-matched threshold control:** `verdict_matched_run.py --model X` → `verdict_matched_X.json` (Stage 1 spectrum at the decision and threshold token, token-length- matched pairs as the primary analysis with all-pairs as sensitivity; Stage 2 projection-patch Δd_L with a polarity positive control).
- **Snap behavioral screen** (Fig. 6): `reportability_precheck.py --model X` → `reportability_precheck_X.json`.
- **Formation probe** (Stage 1b, §B.2): `formation_probe.py --model X` → `formation_probe_X.json` (within-side margin, signed-$(A - T)$, cross-margin generalization).
- **Composition test** (Stage 2b, §4.3): `compose_amt.py --model X` → `compose_amt_X.json` (distinctness, validity, bidirectional $\Delta d_L/\Delta z_r$, carry).

All numerical results reported in the paper are generated from the result files listed above; the accompanying results manifest maps each paper artifact to its generating script and output.

F MISTRAL: MECHANISTIC GENERALITY DOES NOT IMPLY BEHAVIORAL COMPETENCE

Mistral's component signature matches the other three models (Table 1: structuring 107.6% of polarity; stated threshold 0.8%): the same component-level qualitative-over-quantitative signature is observed. This must not be read as Mistral behaviorally integrating the threshold.

A bidirectional control under the same intermediate-judgment scaffold is asymmetric: when the threshold comparison is supplied, the model corrects to SATISFIED across 60 over-flag items (60/60), but does not reverse to EVADED across 36 under-flag items (0/36). The over-flag cases have gold verdict SATISFIED because each component deposit exceeds the stated threshold, whereas the under-flag cases have gold verdict EVADED because each component is below it; in both, reportability and evasion are still elicited from the model rather than supplied. When the model derives the comparison itself instead of being given it, the over-flag figure is 57/60 and the under-flag figure is 0/36, so the asymmetry does not depend on supplying it. Of the three over-flag items outside 57/60, two answered EVADED and one stopped after the intermediate judgments without stating a verdict, leaving nothing to score.

Mistral is therefore not a clean threshold-integrating positive control; its behavioral prior differs from the others while its causal component mechanism is the same. Mechanistic invariance across architectures thus does not entail behavioral invariance.