# COMMITTED BY THE FIRST TOKEN: HOW AN INJECTED ANSWER MIGRATES INTO CHAIN-OF-THOUGHT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Chain-of-thought is usually read as the process by which a model works out its answer. We ask what a chain does with an answer that is already present when it starts. We prefill the assistant turn with graded yes/no commitments and, by holding the reasoning chain fixed while varying what sits in context, decompose the resulting shift in the elicited answer distribution into a *direct* path, through which the commitment is read at answer time, and a *mediated* path, carried by the chain the commitment induced. On BoolQ with Qwen3-4B a plain commitment overturns about half of the correct answers it contradicts, and the mediated path dominates: a chain written under a plain "no", with the commitment itself removed, displaces the answer by 10.9 logits, against 5.0 for the commitment alone. Step-level probing shows the hand-off directly: the direct pull decays over the chain while the mediated pull grows, the two crossing within five steps. Three interventions confirm the split: masking the commitment from attention during generation drives the mediated effect to zero; withholding the supporting passage leaves the direct path unchanged and roughly doubles the mediated one, so the chain is not carrying the commitment by quoting evidence for it; and a chain generated under the opposite commitment overturns the commitment still beside it, driving compliance from 0.85 to 0.24. The model cannot report an unmarked commitment as foreign, and being told does not reduce compliance; directing the chain does. A forced verification prefix lowers compliance with a wrong commitment in every condition, and attenuating attention to the commitment to a tenth of its weight recovers 79% of the net damage while keeping the commitment readable. On Ministral-8B the effect is larger, not smaller, and the commitment's strength—legible before reasoning begins—no longer orders the outcome by the end of the chain.

## 1 INTRODUCTION

A reasoning chain is usually read as the place where a model works its answer out. That reading is under pressure from two directions. Faithfulness studies show that stated reasoning can omit what actually drove a prediction (Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025), and commitment studies show that models often settle on an answer early, with the remaining steps leaving it unchanged (Scalena et al., 2026; Boppana et al., 2026). Both lines study commitments the model reaches by itself. They leave open a question that matters whenever the start of an answer is not written by the model—in prefilling, in agent pipelines that seed a response, in multi-turn settings where an earlier conclusion sits in the assistant's own turn: what does a reasoning chain do with an answer that is already there when it starts?

Two answers are plausible. The chain could work as a *check*, so that a commitment the evidence does not support is gradually overridden. Or it could work as a *vehicle*, so that the commitment shapes what is written and the written chain then carries it forward. The two predict the same end-of-chain answer and differ only in where the commitment lives by the time the answer is produced. Distinguishing them requires separating the commitment in context from the chain it produced, which end-of-chain flip rates cannot do.



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.

We make that separation experimentally by splicing chains across conditions: placing a commitment in front of a chain it did not produce, or removing it from in front of one it did. Probing the elicited yes/no distribution under each combination decomposes the total displacement into a direct and a mediated part, and probing at every step boundary shows how the two evolve.

**Contributions.**

- **A fixed-chain decomposition of injected commitment** (§3). Splicing varies the commitment in context and the chain it induced independently, separating a direct effect (commitment present, clean chain) from a mediated one (chain present, commitment removed).
- **The commitment migrates into the chain** (§4.1–4.3). Before any reasoning all displacement is direct (7.3 logits for "yes", 17.7 for "no"); by the end of the chain the direct pull has decayed to 4.0 and 5.0 while the mediated pull has risen from zero to 7.0 and 10.9. Masking the commitment during generation removes the mediated effect.
- **The mediated path is not evidence selection** (§4.7). Without the passage, direct effects stay in a narrow band ($-0.1$ to 3.4 logits) while mediated effects roughly double (to 10.0–17.3). A chain generated under the opposite commitment overturns the commitment in context (compliance 0.85 to 0.24).
- **Self-report fails; directing the chain works** (§4.4, 4.5, 4.6). The model never reports an unmarked commitment as foreign, and labelling it does not change compliance. A forced verification prefix lowers compliance with a wrong commitment in every condition, and attenuating attention to the commitment to a tenth of its weight turns a loss of 603 questions into 124 without hiding it or breaking the output format.
- **The mechanism is not one model's** (§4.8). On Ministral-8B-Instruct a plain "no" overturns 0.87 of the correct answers it contradicts (Qwen: 0.56); strength is read at the start of the chain and lost by its end; the mediated path is at least as large as the direct one in every condition, though it does not dominate as in Qwen.

## 2 Related Work

**Self-correction.** Whether a model can repair its own errors without external feedback has been contested since Huang et al. (2024) reported that intrinsic self-correction generally fails, motivating training-time remedies (Kumar et al., 2025; Qu et al., 2024). That line studies second-attempt revision. We study whether a single chain can escape a commitment present when it begins, and find that the chain is where the commitment ends up rather than where it is examined.

**Faithfulness.** Turpin et al. (2023) showed that biasing features shift both prediction and explanation without being mentioned; Lanham et al. (2023) and Tanneru et al. (2024) formalised interventions on the chain, and Chen et al. (2025) found that reasoning models verbalise a hint they use less than 20% of the time. Closest to us, Yee et al. (2024) injected errors *inside* a chain and found recovery more frequent when context supplies more evidence. Our injected content is a conclusion rather than a premise, precedes the chain, and is carried forward by text the model itself writes, so there is no seam for the chain to find; their evidence finding has a counterpart here with the expected sign and an unexpected mechanism (§4.7).

**Answer commitment.** Scalena et al. (2026) identify a commitment boundary after which the final answer is essentially fixed; Boppana et al. (2026) decode the answer from activations before it is verbalised; Bogdan et al. (2025) find that a few sentences carry outsized influence. These characterise *endogenous* commitment. We inject an *exogenous* one and track how its influence moves from the injected tokens into the generated ones.

**Prefilling.** Assistant-side prefilling is a priming attack (Vega et al., 2023) and a standard jailbreak primitive (Andriushchenko et al., 2025; Struppek et al., 2026); Qi et al. (2025) argue that alignment mainly adapts the first few output tokens. Our step-level results add a complementary picture: the first tokens matter less for what they say at answer time than for what they cause to be written. Because prefill wording swings outcomes (Dotsinski & Eustratiadis, 2026), we treat strength as a designed factor and test lexical variants within a level. Wang et al. (2026) report that detection of

| condition | prefill text | role |
|---|---|---|
| none | — | reference |
| neutral | *Let me think about this step by step.* | content-free control |
| hedge | *I think the answer is probably yes/no.* | strength 1 |
| assert | *The answer is yes/no.* | strength 2 (plain) |
| strong | *The answer is definitely yes/no.* | strength 3 (emphatic) |
| revise | *The answer is yes/no, but let me verify this carefully.* | strength 2 + licence |
| lexical | *The answer is {certainly, obviously, clearly, undoubtedly, without a doubt} yes.* | strength 3, wording |
| check | hedge or assert + *Let me check the passage.* | forced verification |

Table 1: Commitment conditions.

a prefill and resistance to it are decoupled; for an unmarked assistant-turn prefill we find detection at floor. Kumaran et al. (2026) find that a choice-supportive bias disappears when an answer is attributed to another model; our within-turn attribution results point the other way (§4.5). Our damage/rescue split mirrors Laban et al. (2023).

## 3 SETUP

**Model and data.** We use Qwen3-4B with thinking disabled and greedy decoding (at most 1,024 new tokens). Questions are the first 2,000 items of the BoolQ training split with their passages; the gold "yes" rate (0.617) matches the full split. The prompt asks for one short reasoning step per line, prefixed "Step $N$:". Without any prefill the model answers 84.9% correctly. §4.7 repeats all conditions with the passage withheld (4,017 questions, a superset; unprefilled accuracy 63.9%), and §4.8 on Ministral-8B-Instruct (unprefilled accuracy 83.3%).

**Commitment conditions.** A condition is a string placed at the start of the assistant turn, from which the model continues (Table 1): three strengths per polarity, a revision licence appended at the plain level, a content-free neutral prefill, and five alternative intensifiers at the emphatic level.

**Probe and displacement.** After each step boundary we append a blank line and *Final answer (Yes or No):*, run one forward pass, and read the logits of " Yes" and " No"; the margin is $m = \ell_{\text{Yes}} - \ell_{\text{No}}$. Probes where the two tokens carry less than half the probability mass are dropped (33 of about 70,000 in the main grid; the model answers on a True/False surface). For a question $q$ under a condition of polarity $s \in \{+1, -1\}$, end-of-chain displacement is $d_T(q) = s\,(m_T(q) - m_T^{\text{none}}(q))$, so positive values always mean movement toward the prefilled answer; $d_0$ is the same quantity before any reasoning. Each question is compared with itself, so no between-question variance enters. *Compliance* is the fraction of questions whose final answer matches the commitment; the *overturn rate* is compliance on questions the model answers correctly without a prefill and the commitment contradicts.

**Fixed-chain decomposition.** Let $X$ be a commitment and $c_X$ the chain generated after it. Splicing gives four cells per $X$:

$$\text{total} = d_T(X, c_X), \qquad \text{direct} = d_T(X, c_{\text{none}}),$$
$$\text{mediated} = d_T(\text{none}, c_X), \qquad \text{interaction} = \text{total} - \text{direct} - \text{mediated}.$$

The direct cell is a controlled direct effect: the commitment is present, the chain is the one the model writes without it. The mediated cell removes the commitment and keeps its chain. Splicing produces sequences the model would not generate; §4.3 validates the decomposition with an intervention that does not splice. Identical prefill strings give identical results to four decimal places, so the noise floor is zero.

**Damage and rescue.** Pooled accuracy hides opposite movements on questions where the commitment is right and where it is wrong, which with a 62% "yes" base rate can cancel almost exactly, and on questions where the commitment is correct it equals compliance. Against the unprefilled run,

| | Qwen3-4B, with passage | | | Qwen3-4B, no passage | | | Ministral-8B, with passage | | |
|---|---|---|---|---|---|---|---|---|---|
| condition | total | direct | mediated | total | direct | mediated | total | direct | mediated |
| hedge_yes | 4.07 | 0.01 | 4.41 | 11.52 | −0.10 | 13.38 | 2.87 | 0.82 | 2.40 |
| revise_yes | 3.74 | 0.37 | 3.87 | 8.28 | −0.10 | 9.98 | — | — | — |
| assert_yes | 11.04 | 3.96 | 6.98 | 17.90 | 3.39 | 15.66 | 3.52 | 1.97 | 2.23$^{\dagger}$ |
| strong_yes | 9.76 | 3.44 | 7.34 | 17.91 | 2.86 | 17.28 | 3.69 | 1.94 | 2.42 |
| hedge_no | 7.72 | 2.39 | 8.78 | 12.04 | 1.78 | 14.53 | 3.67 | 1.00 | 3.85 |
| revise_no | 8.92 | 3.53 | 8.68 | 11.14 | 2.80 | 12.02 | — | — | — |
| assert_no | 12.48 | 4.98 | 10.91 | 14.08 | 3.28 | 15.47 | 4.52 | 2.10 | 4.00 |
| strong_no | 11.08 | 4.74 | 10.40 | 14.51 | 3.24 | 16.69 | 4.09 | 1.75 | 3.93 |

Table 2: Fixed-chain decomposition of end-of-chain displacement (logits; interaction = total − direct − mediated). $n = 1{,}997$ per cell with the passage, $4{,}017$ without (3,694 for revise), and $\approx 1{,}995$ for Ministral. Every direct-versus-mediated difference is significant by paired Wilcoxon ($p < 10^{-16}$ with passage, $p < 10^{-170}$ without, $p < 10^{-95}$ for Ministral) except $^{\dagger}$assert_yes on Ministral ($p = 0.55$) and strong_yes on Ministral ($p = 0.004$). Ministral's logits are on a different scale (§4.8) and are not comparable across columns.

*damage* is the fraction of correctly answered questions that become incorrect with a prefill, and *rescue* the fraction of incorrectly answered questions that become correct. Net effect is rescued minus damaged questions. Of the 2,000 questions, 1,697 are answered correctly without a prefill and 303 incorrectly.

# 4 Results

## 4.1 Most of the displacement is carried by the chain

A plain commitment overturns 43% ("yes") and 56% ("no") of the answers the model gets right unprompted and the commitment contradicts, about 51% pooled; the accuracy losses are 6.0 and 24.8 points. In every condition the mediated effect exceeds the direct one (Table 2, left). For a plain "yes" the chain alone carries 6.98 logits and the commitment alone 3.96; for "no", 10.91 against 4.98. The gap is widest at low strength: a hedged "yes" has no measurable direct effect (0.01 logits) while the chain it produces carries 4.41.

The decomposition is close to additive for "yes" (interaction $-1.0$ to $+0.1$) and not for "no", where the interaction is consistently about $-3.4$: for hedged and revision-licensed "no" the mediated effect alone exceeds the total, so once the chain has absorbed the commitment, keeping the commitment in context slightly weakens the pull. We therefore report components in absolute terms (§4.7 shows this non-additivity belongs to the commitment, not the evidence).

Strength is not monotonic. Pooled over polarity, displacement is 5.9 logits for hedged, 11.8 for plain and 10.4 for emphatic commitments (Spearman $\rho = 0.14$); the plain commitment exceeds the emphatic one in total, direct and (for "no") mediated displacement in both polarities. Five alternative intensifiers span 2.79–3.54 logits of direct effect (s.d. 0.30), so the emphatic level is a level and not the idiosyncrasy of one word (Appendix A); the non-monotonicity does not survive removing the passage (§4.7). Appending a revision licence nearly eliminates the direct path for "yes" (3.96 to 0.37) and, pooled over polarity, reduces displacement from 11.8 to 6.3 logits (Mann–Whitney $p < 10^{-90}$), raising accuracy from 0.695 to 0.762.

## 4.2 The commitment migrates into the chain

Probing at every step boundary shows how the decomposition arises (Figure 1). At step 0 the total and direct cells coincide, since neither contains reasoning yet, and the mediated cell is exactly zero because its prompt is the unprefilled one. From there the direct pull decays—from 10.2 at step 1 to 4.0 at the end for "yes", from 17.7 to 5.0 for "no"—while the mediated pull rises from zero to 7.0 and 10.9. The two cross between steps 4 and 5. The total grows for "yes" (7.3 to a peak of 14.8) and

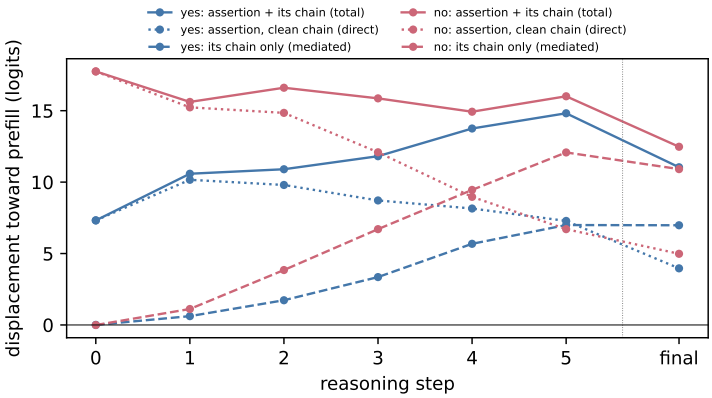



Figure 1: Displacement at each reasoning step, relative to the unprefilled run at the same step (Qwen3-4B, plain commitments, with passage). Solid: commitment and its own chain (total). Dotted: commitment with the unprefilled chain (direct). Dashed: the commitment's chain with the commitment removed (mediated). "Final" uses each chain's last step; step 5 includes only chains at least that long.

declines for "no" (17.7 to 12.5); what is common is its location: by the end most of it is carried by text the model wrote.

### 4.3 Masking the commitment during generation

Splicing places a commitment in front of a chain it did not produce. To test the decomposition without that artefact, we regenerate each chain with the commitment in the token stream but removed from the attention mask, with position ids adjusted so the hidden tokens leave no gap. The commitment is present at answer time but could not be read while the chain was written.

The resulting chains carry no displacement of their own (mediated 0.05 and $-0.05$ logits for "yes" and "no"; $-0.11$ for a control running the same code path without masking). With the commitment in context, total displacement falls to 3.97 and 4.95 logits, within 0.03 of the direct effects measured by splicing (3.96 and 4.98). Two methods that share no machinery agree on where the displacement sits. Masked chains are no more similar to unprefilled chains than a regeneration of the unprefilled condition is (69.8% versus 72.2% verbatim agreement): masking is equivalent to never having shown the commitment.

### 4.4 The commitment shapes the form of the chain, and directing the form helps

The prompt asks for one step per line, and without a prefill the model complies in all but 0.1% of chains. A prefill changes this, and the weaker the commitment the larger the change: after a hedged commitment the model runs its steps together in 59–74% of chains and drops step labels altogether in 11–18%, writing prose instead; a plain commitment affects 1–5% (Appendix B). The prose chains are not a formatting accident: 85–93% of them open with *Let me check the passage*, and when the hedged commitment is wrong the model follows it in 16% of them against 30% ("yes") and 51% ("no") of stepwise chains, a difference that holds within every quartile of unprompted confidence (logistic coefficient $-1.34$, $p < 10^{-10}$).

Chains the model chose to write are not the only ones that resist. Prefixing *Let me check the passage.* to the commitment lowers compliance with a wrong commitment in every condition (Table 3): from 0.346 to 0.193 for a hedged "no" and from 0.555 to 0.431 for a plain "no". At the hedged level the forced check recovers essentially all of the spontaneous effect (0.16 for chains that check of their own accord); at the plain level it recovers part. The forced chains typically quote the passage once

| condition | $n$ | no check | + check | McNemar $p$ |
|---|---|---|---|---|
| hedge_yes | 1834 | 0.202 | 0.162 | $8.6 \times 10^{-6}$ |
| hedge_no | 2555 | 0.346 | 0.193 | $1.1 \times 10^{-64}$ |
| assert_yes | 1908 | 0.420 | 0.356 | $1.0 \times 10^{-9}$ |
| assert_no | 2623 | 0.555 | 0.431 | $8.7 \times 10^{-37}$ |

Table 3: Compliance with a wrong commitment with and without a forced verification prefix, on questions present in both runs and answered correctly without any prefill (exact McNemar test on paired discordant counts).

and then resume the committed line: the check works to the extent that it changes what the chain says, and a stronger commitment leaves less of the chain to change.

### 4.5 Neither ownership nor source governs the effect

**Detection.** After the chain we ask, in a new user turn, whether the first line of the reply was the model's own, in two wordings of opposite polarity so that a yes-bias cancels (Appendix I). For an unmarked commitment the model never says it is not its own (0 of 2,000; 0 of 2,000 for its genuinely own text). When the commitment is preceded by a line stating that another model wrote it, 14.1% of answers call it foreign, so the probe is sensitive when the information is present. Being told does not reduce compliance (0.556 versus 0.562). Within the labelled condition, detection after the chain strongly predicts non-compliance after controlling for the unprompted margin (logistic coefficients $-1.17$ and $-0.81$, both $p < 10^{-15}$), while the same probe asked before any reasoning carries no signal of the expected sign and is uncorrelated with the post-chain one ($r = -0.07$; Appendix I). The sense that the commitment is foreign forms during reasoning alongside the refusal to follow it, rather than preceding it.

**Attribution and position.** Labelling an assistant-turn "yes" commitment as another model's lowers its direct effect from 3.96 to 2.89 logits, but a length-matched neutral line that makes no claim about authorship lowers it to 2.75: the label's content contributes nothing. Moving the commitment into the user turn increases its immediate pull ($d_0 \approx 12.6$ against 7.3), and attributing it to a person rather than a model changes it by 0.70 logits. Across all five placements compliance stays between 0.556 and 0.562 (Appendix C): with the chain held clean, none of these manipulations reaches the final answer.

**A content-free prefill compresses confidence.** The neutral prefill biases neither direction (mean change $-3.9$ logits on "yes" questions, $+2.3$ on "no"; both toward zero) but shrinks mean absolute margin from 21.9 to 17.8 logits (19%), against 0.4% for regenerating the unprefilled condition through a separate code path. The compression flips 4.8% of answers with a passage and 14.7% without (Appendix H). Occupying the assistant turn is not neutral, but its effect is on certainty, not direction.

### 4.6 Attenuation

Masking shows how much is lost when the commitment is read during generation, but removing it entirely is not always an available intervention. We instead scale its attention weight: a bias of $\log \alpha$ is added to the pre-softmax scores of the commitment's positions at every generation step, so that after normalisation they carry $\alpha$ times their usual weight; $\alpha = 1$ reproduces unmasked generation (mediated 6.79 logits against 6.98) and $\alpha = 0$ the hard mask.

Displacement falls smoothly with $\alpha$ rather than switching at zero (Table 4). Without intervention the two commitments together cost 603 questions; at $\alpha = 0.1$ the loss is 124, so 79% of the damage is recovered while the commitment stays readable. At $\alpha = 0$ generation degrades: 14.1% of chains abandon the step format and some emit stray thinking tags, whereas none does at any $\alpha \geq 0.1$. Even a weakly visible commitment anchors the format of the reply.

| $\alpha$ | 0 | 0.1 | 0.25 | 0.5 | 0.75 | 1 |
|---|---|---|---|---|---|---|
| "yes": total $d_T$ / mediated | 2.94 / $-0.44$ | 5.71 / 1.81 | 7.08 / 3.28 | 8.79 / 4.99 | 9.84 / 5.99 | 10.85 / 6.79 |
| "yes": net questions | $+6$ | $+14$ | $+7$ | $-43$ | $-70$ | $-111$ |
| "no": total $d_T$ / mediated | 4.39 / 0.67 | 6.46 / 1.67 | 7.76 / 3.94 | 9.69 / 7.01 | 11.19 / 9.25 | 12.53 / 11.06 |
| "no": net questions | $-82$ | $-138$ | $-200$ | $-307$ | $-402$ | $-492$ |
| combined net | $-76^*$ | $-124$ | $-193$ | $-350$ | $-472$ | $-603$ |

Table 4: Attenuation sweep for the plain commitment. Net questions are rescued minus damaged relative to the unprefilled run, with the commitment in context. *At $\alpha = 0$ about 14% of chains lose the step format and are excluded, so that column is not comparable. Damage and rescue curves are in Appendix D.

**Why the polarities differ.** A "no" commitment damages a third of the questions the model answers correctly (0.334) and rescues a quarter of those it gets wrong (0.248); a "yes" commitment damages 0.169 and rescues 0.578. The difference follows from the model's own errors: of its 303 errors, about 219 are "yes" questions answered "no", so a "no" commitment can rescue at most about 0.28, which it nearly reaches. The same asymmetry decides which path does the harm. A chain written under "yes" is by itself net beneficial at every $\alpha$ (it corrects more "no"-leaning errors than it introduces), so all of the "yes" commitment's damage comes through the direct path at answer time. A chain written under "no" compounds the model's existing bias and accounts for 448 of the 492 questions lost (91%). Whether a displaced chain hurts depends on whether it is displaced in the direction the model already errs.

### 4.7 Removing the passage

Every result so far is on a task where the answer is available in context, which allows an alternative reading of the mediated effect: the chain may carry the commitment by selectively quoting the passage in its favour. We therefore repeat the entire decomposition with the passage withheld; questions, commitments, probe, splicing and model are unchanged (Appendix F).

**The direct path is unchanged; the mediated path doubles.** The ablation acts on one path and not the other (Table 2, middle). Direct effects are, if anything, slightly smaller without the passage (3.39 against 3.96 for a plain "yes", 3.28 against 4.98 for a plain "no") and stay within $-0.10$ to 3.39 logits across all eight conditions. Mediated effects roughly double: 15.66 against 6.98, 17.28 against 7.34, 15.47 against 10.91, and 13.38 against 4.41 for a hedged "yes". This rules out the evidence-selection reading in the direction opposite to the one it predicts: if the chain carried the commitment by quoting supportive material, removing the passage would take the mechanism away. Instead the chain's grip strengthens while the commitment's own grip at answer time does not. A model with evidence in context has something to write against; a model reasoning from parametric knowledge alone has only the commitment, and writes a chain that follows it more completely. For a hedged and a revision-licensed "yes" the direct effect is $-0.10$ logits while their chains carry 13.38 and 9.98: whatever a weak commitment does, it does entirely by changing what gets written. Overturn rates rise accordingly (64.8% and 74.6% for plain "yes" and "no", against 42.5% and 55.6% with a passage). The polarity asymmetry in the mediated effect disappears (15.47 against 15.66 at the plain level) while the interaction stays more negative for "no" ($-3.7$ to $-5.4$ against $-1.2$ to $-2.2$), so non-additivity on the "no" side is a property of the commitment. The emphatic level now matches or exceeds the plain one (17.91 against 17.90; 14.51 against 14.08): an overstated commitment appears to invite verification, which is only actionable against something the passage supplies.

**The hand-off replicates.** Step-level probing reproduces the migration of §4.2 (Figure 7). For a plain "yes" the direct pull falls from 11.1 to 3.4 logits over the chain while the mediated pull rises from 3.5 to 15.7 (emphatic: 10.9 to 2.9 against 3.5 to 17.3; hedged: 8.1 to $-0.1$ against 3.5 to 13.4). The curves cross between the third and fifth decile in all cases, where they cross with the passage.

**A chain can overturn the commitment beside it.** Crossing commitments with chains of the opposite polarity gives a control the with-passage run lacks. A "no" commitment placed in front of a chain generated under "yes" displaces the answer $-8.53$ logits at the hedged level and $-7.83$ at

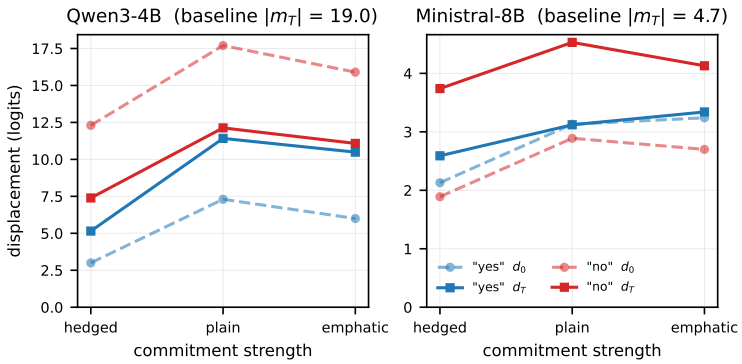



Figure 2: Displacement before reasoning ($d_0$, dashed) and after the chain ($d_T$, solid) against commitment strength. Panels do not share a $y$ axis. Ministral orders the strength levels at $d_0$ and not at $d_T$.

the emphatic level: the answer ends up on the side the *chain* argues for, against the commitment the model can still read. Compliance with the in-context commitment falls to 0.235 and 0.304 in these cells, against 0.845 and 0.908 when commitment and chain agree (Appendix F). Given a direct conflict between an explicit commitment and the text of the reasoning, the text wins.

**Masking agrees, with a stronger control.** Regenerating with the commitment masked from attention, masked chains again carry no displacement ($-0.08$ and $+0.08$ logits), and with the commitment in context the total falls to 3.27 and 3.46, within 0.13 of the spliced direct effects (3.39 and 3.28); masked and unprefilled trajectories agree to within 0.6 logits at every decile. The control here runs the same code path with the same commitment present and the mask switched off, so the only difference is whether the commitment could be read during generation: that chain carries 15.46 logits of mediated displacement against $-0.08$ when masked. The generation machinery contributes nothing; reading the commitment while writing contributes all of it.

### 4.8 A SECOND MODEL FAMILY

To test whether the mechanism is a property of Qwen3-4B, we repeat the conditions and the decomposition on Ministral-8B-Instruct, a different family with a different tokenizer, chat template and alignment recipe; task, questions, prefix inventory, probe and splicing are unchanged.

**A prefill is not automatically continued.** Given a bare commitment in the assistant turn, Ministral emits an end-of-sequence token immediately, treating *The answer is yes.* as a finished reply. Appending a single newline is sufficient and minimal; the commitment text is unchanged and the model then writes a correctly formatted chain in every condition (mean 3.0–3.6 steps, unprefilled accuracy 0.833). Whether a commitment is adopted at all thus depends on whether the chat format leaves the turn open, not on resistance.

**The effect is larger, not smaller.** Overturn rates exceed Qwen's in every condition (Appendix G): 0.652 against 0.425 for a plain "yes", 0.868 against 0.556 for a plain "no", with the same polarity ordering; accuracy under a plain "no" falls from 0.833 to 0.451. The two models operate on different confidence scales (mean $|m_T^{\text{none}}|$ 4.70 against 19.0 logits): mean $d_T$ is smaller for Ministral (3.57 against 9.61) yet larger as a fraction of the baseline margin (0.76 against 0.51), hence the higher overturn rate. We compare models on the unit-free overturn and compliance, and report $d_T$ within a model only.

**Strength is read but not retained.** The dose–response relation appears at the start of the chain and is gone by the end (Figure 2). Before any reasoning, Ministral separates the three strengths in order: $d_0$ is 2.13, 3.13 and 3.24 logits for hedged, plain and emphatic "yes", and 1.89, 2.89 and 2.70 for "no". After the chain the levels have collapsed: $d_T$ is 2.59, 3.12 and 3.34 (a spread of 0.75

where $d_0$ spanned 1.11), and the overturn rates are 0.715, 0.652 and 0.718, not ordered at all. The commitment's strength is legible before reasoning begins and does not survive the chain, because by its end the displacement is carried by text the model wrote, and that text does not preserve how firmly the commitment was stated. Qwen shows the same collapse in weaker form (its emphatic level falls below its plain level in $d_T$ but not in $d_0$).

**The decomposition transfers, with one qualification.** The mediated effect is at least as large as the direct one in all six conditions and significantly larger in four (Table 2, right). For a plain "yes" the two paths are indistinguishable (1.97 against 2.23, $p = 0.55$) and for an emphatic "yes" the gap is small (1.94 against 2.42, $p = 0.004$); on the "no" side and at the hedged level mediated exceeds direct by 1.6 to 2.9 logits per question ($p < 10^{-95}$). So the weaker claim transfers and the stronger one does not: that the chain carries a substantial part of the commitment holds in both families; that it carries *most* of it is, on this evidence, a property of Qwen. Consistent with Figure 2, Ministral's chains absorb less of the commitment (direct share 0.27–0.56 against 0.19–0.23), and so also less of its strength.

## 5 Discussion

**Reasoning as a vehicle.** A chain that could correct a premature answer would have to examine it. Here it mostly transmits it. The commitment's direct pull does weaken over the chain—a model reading a clean chain beside the commitment moves back toward its own answer—but the chain the model actually writes has by then absorbed the commitment, and nothing downstream separates the two. When commitment and reasoning text point in opposite directions, the text decides. This is a stronger failure than unfaithfulness: the reasoning is not merely silent about what drove the answer, it is itself the channel. The second family qualifies how much the vehicle carries, not its direction: across two families and with and without evidence in context, no configuration we measured has the chain working against the commitment that produced it.

**What evidence in context is for.** The passage does not prevent a commitment from being adopted (direct effects are the same with and without it), but it is the only thing in context that competes with the commitment for what the chain will say; remove it and the mediated path doubles. On this reading, retrieval is a partial defence against a seeded answer, tasks answered from parametric knowledge alone are where a commitment is most durable, and interventions should act on generation rather than on the answer.

**What helps, and what does not.** Asking the model about its output presupposes that it can locate the problem; for an unmarked commitment it cannot, and for a labelled one only after the chain has formed. Directing the chain works instead: a forced verification prefix, the revision licence, attenuation, masking and supplying evidence all give generation something other than the commitment to follow.

## 6 Limitations

The mechanism is characterised on one model and one dataset; §4.8 tests transfer of the conditions and the decomposition only, the passage ablation varies evidence rather than domain, and attenuation is quantified for the plain commitment with a passage only. Displacement is bimodal (flipped answers move by 28–40 logits, unflipped by 0.4–3.9; Appendix H), so mean $d_T$ is close to the flip rate times a constant step, medians are much smaller than means, and the direct-to-mediated ratio should be read as a ratio of flip propensities; where the unprefilled margin is near its ceiling, displacement can turn slightly negative, so averages of $d_T$ are conservative. Logits are not comparable across models, and no cross-model claim rests on a $d_T$ comparison. Splicing creates sequences the model would not produce; masking addresses this in both runs and agrees with splicing to within 0.03 and 0.13 logits. How the assistant turn is opened matters on its own (a StrategyQA pilot moved the unprefilled "yes" rate from 0.12 to 0.02 by changing only the opening); all conditions here share one opening path. The detection probe measures a behavioural report, not introspective access.

## REFERENCES


Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *International Conference on Learning Representations*, 2025. arXiv:2404.02151.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which LLM reasoning steps matter? *arXiv preprint arXiv:2506.19143*, 2025.

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

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Asen Dotsinski and Panagiotis Eustratiadis. Sockpuppetting: Jailbreaking LLMs by combining prefilling with optimization. *arXiv preprint arXiv:2601.13359*, 2026.

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

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D. Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, 2025. arXiv:2409.12917.

Dharshan Kumaran, Stephen M. Fleming, Larisa Markeeva, Joe Heyward, Andrea Banino, Mrinal Mathur, Razvan Pascanu, Simon Osindero, Benedetto De Martino, Petar Veličković, and Viorica Patraucean. Competing biases underlie overconfidence and underconfidence in LLMs. *Nature Machine Intelligence*, 8(4):614–627, 2026. doi: 10.1038/s42256-026-01217-9. Preprint: arXiv:2507.03120.

Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure? challenging LLMs leads to performance drops in the FlipFlop experiment. *arXiv preprint arXiv:2311.08596*, 2023.

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

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, 2025. Outstanding Paper Award; arXiv:2406.05946.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching language model agents how to self-improve. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Daniel Scalena, Sara Candussio, Luca Bortolussi, Elisabetta Fersini, Malvina Nissim, and Gabriele Sarti. Beyond the commitment boundary: Probing epiphenomenal chain-of-thought in large reasoning models. *arXiv preprint arXiv:2606.13603*, 2026.

Lukas Struppek, Adam Gleave, and Kellin Pelrine. Exposing the systematic vulnerability of open-weight models to prefill attacks. *arXiv preprint arXiv:2602.14689*, 2026.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*, 2024.

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*, volume 36, 2023.

Jason Vega, Isha Chaudhary, Changming Xu, and Gagandeep Singh. Bypassing the safety training of open-source LLMs with priming attacks. *arXiv preprint arXiv:2312.12321*, 2023.

Andy Wang, Parv Mahajan, David Demitri Africa, Alexandra Souly, Jordan Taylor, and Robert Kirk. Prefill awareness in large language models. *arXiv preprint arXiv:2606.12747*, 2026.

Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. Faithful and unfaithful error recovery in chain of thought. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2405.15092, titled "Dissociation of Faithful and Unfaithful Reasoning in LLMs".

## A LEXICAL VARIANTS AT THE EMPHATIC LEVEL

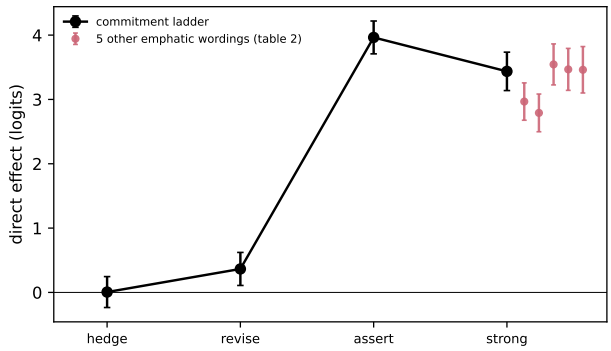



Figure 3: Direct effect along the commitment ladder, with five alternative emphatic wordings placed at the emphatic level.

| wording | direct $d_T$ | 95% CI |
|---|---|---|
| definitely (= strong_yes) | 3.44 | 0.30 |
| certainly | 2.97 | 0.29 |
| obviously | 2.79 | 0.29 |
| clearly | 3.54 | 0.32 |
| undoubtedly | 3.47 | 0.33 |
| without a doubt | 3.46 | 0.36 |
| assert (plain, for reference) | 3.96 | 0.26 |

Table 5: Direct effect of emphatic wordings. "Definitely" was run twice under separate labels and gave identical results to four decimals. The five intensifiers have not been rerun without a passage.

## B DEPARTURE FROM THE REQUESTED CHAIN FORMAT

Table 6 gives the format statistics summarised in §4.4.

| condition | not one step per line | no step labels |
|---|---|---|
| none | 0.1% | 0 |
| assert_yes / _no | 1% / 5% | 0 / 0 |
| revise_yes / _no | 12% / 14% | 0 / 0 |
| strong_yes / _no | 10% / 33% | 1 / 7 |
| neutral | 10% | 0 |
| hedge_yes / _no | 59% / 74% | 354 / 220 |

Table 6: Departure from the requested chain format with the passage ($n$ = 2,000 each). Endpoint measures are unaffected; step-level analyses use the plain conditions. Without a passage, steps are run together in 37.5% of hedged "no" chains, 19.5% of hedged "yes" and at most 2.8% of plain chains, and only one chain drops step labels.

## C ATTRIBUTION AND POSITION

Table 7 gives the placements and attributions discussed in §4.5. The label and the neutral tag are 16 and 15 tokens.

| placement of *The answer is yes.* | $d_0$ | $d_T$ | complied |
|---|---|---|---|
| assistant turn | 7.32 | 3.96 | 0.562 |
| assistant turn, labelled as another model's | 5.39 | 2.89 | 0.556 |
| assistant turn, length-matched neutral tag | 5.03 | 2.75 | 0.557 |
| user turn, attributed to another model | 12.53 | 4.97 | 0.556 |
| user turn, attributed to a person | 12.79 | 4.26 | 0.557 |

Table 7: Direct effect of the same commitment under different placements and attributions (clean chain, $n \approx 1{,}997$).

## D ATTENUATION: ACCURACY SPLIT BY WHETHER THE COMMITMENT IS CORRECT

Figures 4 and 5 expand Table 4 into damage, rescue and accuracy on the two subsets of questions.

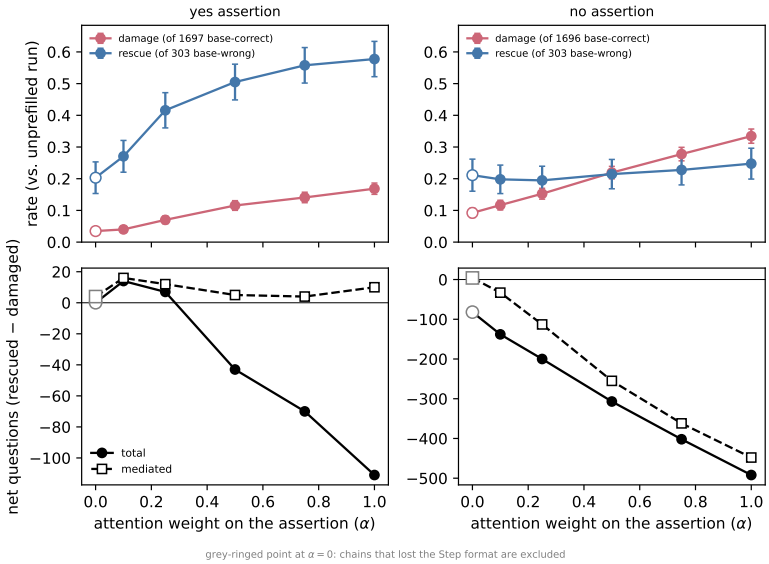



Figure 4: Damage and rescue rates (top) and net questions (bottom) as a function of attention weight on the commitment. Rates are relative to the unprefilled run; note the different pool sizes. Grey-ringed points at $\alpha = 0$ exclude chains that lost the step format.

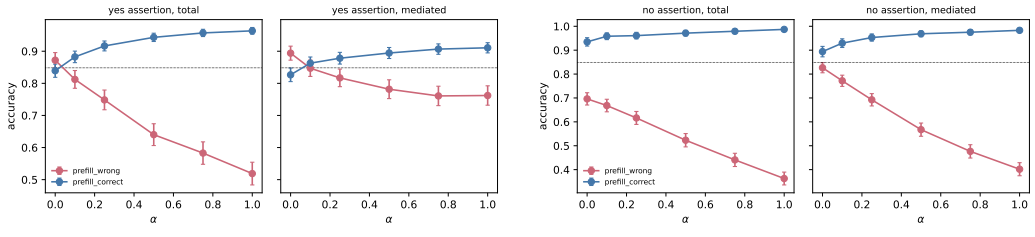



Figure 5: Accuracy under attenuation, separately for questions where the commitment matches the gold answer and where it does not (left: "yes" commitment; right: "no"). On the former, accuracy equals compliance by construction.

## E DISTANCE FROM THE START OF THE TURN

We inserted a neutral filler (*Let me take this one step at a time.*, repeated) between the start of the assistant turn and the commitment. The pre-reasoning margin falls from 4.02 to 0.50 logits as the filler grows to 61 tokens, then rises again to 2.06 at 241 tokens; compliance moves only from 0.562 to 0.556 across the whole range. The rise at long lengths indicates that the repeated filler is not neutral, so we do not interpret this curve as a distance effect.

## F THE NO-PASSAGE RUN

The no-passage run uses the same prompt with the passage block omitted and is otherwise identical: same model, decoding, probe, prefix inventory and splicing code. It covers 4,017 BoolQ training items in every cell (3,694 for the revision-licensed conditions, whose generation run was shorter), a superset of the 2,000 items used elsewhere, with a gold "yes" rate of 0.617. Unprefilled accuracy is 0.639 against 0.847 with the passage, and unprefilled confidence is comparable (mean $|m_T^{\text{none}}| = 20.2$ against 19.1 logits). Of about 145,000 probes, 19 are dropped for low yes/no mass. The revision licence remains the most effective prompt-level remedy, cutting pooled displacement from 16.45 to 10.36 logits (Mann–Whitney $p < 10^{-98}$) and reducing the direct path to zero on the "yes" side.

| condition | accuracy | complied | damage | rescue | net |
|---|---|---|---|---|---|
| hedge_yes | 0.691 | 0.811 | 0.238 | 0.564 | +103 |
| assert_yes | 0.680 | 0.865 | 0.278 | 0.606 | +82 |
| strong_yes | 0.672 | 0.894 | 0.304 | 0.630 | +65 |
| revise_yes | 0.680 | 0.710 | 0.185 | 0.442 | +82 |
| hedge_no | 0.496 | 0.844 | 0.382 | 0.280 | −287 |
| assert_no | 0.464 | 0.885 | 0.432 | 0.280 | −351 |
| strong_no | 0.458 | 0.899 | 0.448 | 0.290 | −364 |
| revise_no | 0.515 | 0.803 | 0.339 | 0.257 | −248 |

Table 8: No-passage run: accuracy and damage/rescue against its own unprefilled reference, on the 2,000 questions shared with the main run (1,279 correct and 721 incorrect without a prefill).

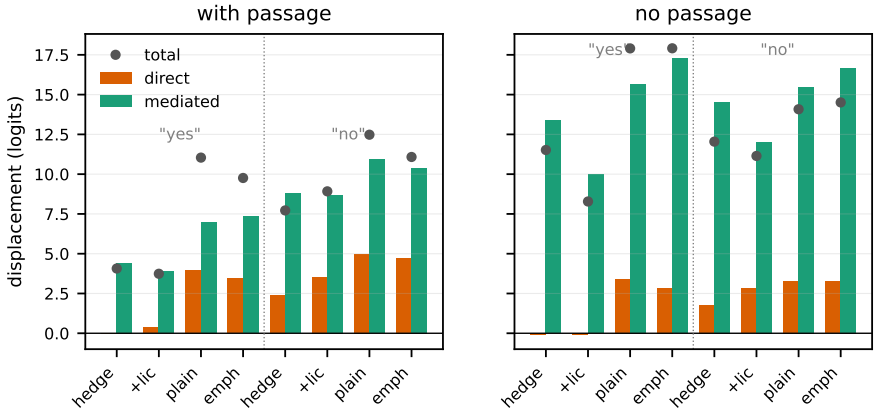



Figure 6: Direct and mediated displacement by condition, with and without the passage. Removing the passage leaves the direct path where it was and roughly doubles the mediated one.

**Cross-polarity cells.** Because the splicing tool crosses chain sources with contexts exhaustively, the grid also contains cells in which the commitment in context and the commitment that generated the chain have opposite polarity (Figure 8). Displacement is signed toward the commitment held in context.

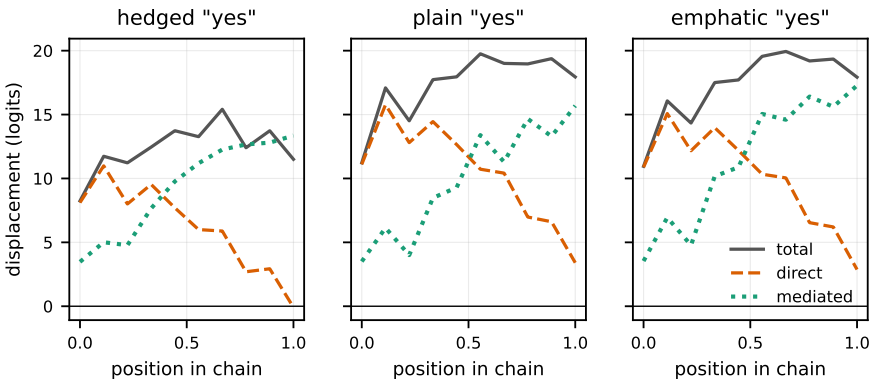



Figure 7: Step-level hand-off without the passage, for the three "yes" commitment strengths. The direct pull decays as the chain lengthens while the mediated pull grows, the two crossing near the middle of the chain.

| context | chain from | $d_T$ | complied |
|---|---|---|---|
| hedge_no | hedge_yes | −8.54 | 0.235 |
| hedge_yes | hedge_no | −11.88 | 0.156 |
| strong_no | strong_yes | −7.82 | 0.304 |
| strong_yes | strong_no | −10.44 | 0.111 |
| revise_no | revise_yes | −4.81 | 0.345 |
| revise_yes | revise_no | −9.96 | 0.206 |

Table 9: Cross-polarity cells, no passage. Negative $d_T$ means the answer moved against the commitment held in context, to the side its chain argues for.

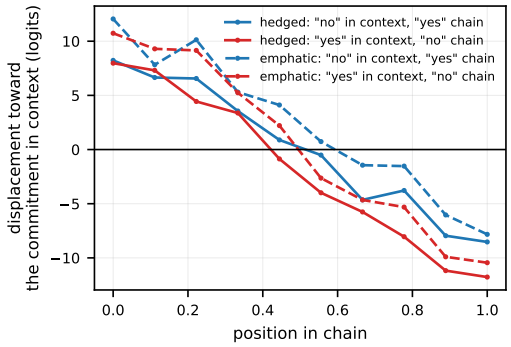



Figure 8: Cross-polarity cells: displacement toward the commitment held in context, when the chain was generated under the opposite commitment. Values below zero mean the answer has moved to the side the chain argues for. The trajectories cross zero near the middle of the chain, at the same place as the hand-off.

## G SECOND MODEL FAMILY: OVERTURN RATES

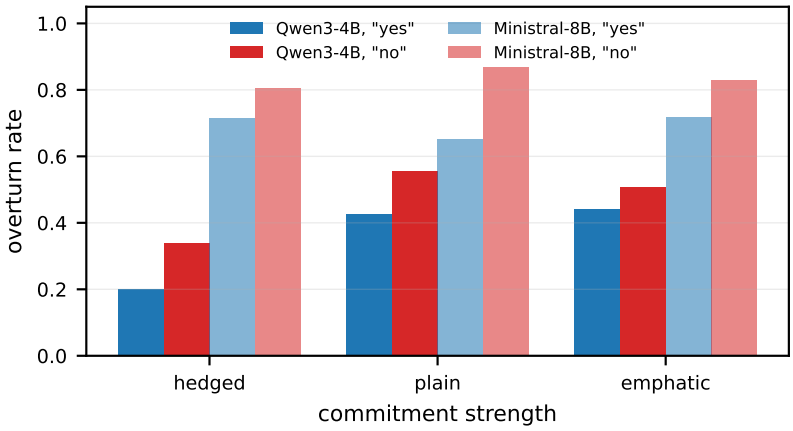



Figure 9: Overturn rate by condition and model, with the passage supplied ($n = 2{,}000$; the denominator is the questions the model answers correctly unprefilled and the commitment contradicts).

Probe mass on Ministral is comparable to Qwen's in the unspliced run (0.75% of final probes below 0.9, against 0.11%). Spliced contexts push more probability off the yes/no pair than generated ones do (6.6% of final probes below 0.9 mass), but the decomposition does not depend on them: raising the mass threshold from 0.5 to 0.9 moves every cell mean by less than 0.02 logits. The interaction terms on Ministral are $-0.35$, $-0.68$ and $-0.67$ for hedged, plain and emphatic "yes", and $-1.18$, $-1.58$ and $-1.60$ for "no".

## H FLIP RATES AND THE BIMODALITY OF DISPLACEMENT

| condition | mean $d_T$ | median $d_T$ | flip rate | $d_T$ \| flip | $d_T$ \| no flip |
|---|---|---|---|---|---|
| *with passage* | | | | | |
| hedge_yes | 5.15 | 0.50 | 0.144 | 29.11 | 1.12 |
| assert_yes | 11.42 | 4.00 | 0.234 | 36.00 | 3.92 |
| strong_yes | 10.49 | 3.25 | 0.240 | 33.83 | 3.12 |
| revise_yes | 5.11 | 1.38 | 0.136 | 27.93 | 1.52 |
| hedge_no | 7.39 | 1.88 | 0.209 | 30.50 | 1.26 |
| assert_no | 12.13 | 3.75 | 0.319 | 33.04 | 2.36 |
| strong_no | 11.08 | 4.25 | 0.292 | 32.32 | 2.32 |
| revise_no | 8.15 | 2.50 | 0.238 | 30.12 | 1.27 |
| neutral | — | — | 0.048 | — | — |
| *no passage* | | | | | |
| hedge_yes | 12.45 | 1.50 | 0.355 | 34.28 | 0.41 |
| assert_yes | 17.90 | 5.62 | 0.396 | 39.58 | 3.69 |
| strong_yes | 17.79 | 5.00 | 0.421 | 38.91 | 2.40 |
| revise_yes | 9.48 | 1.50 | 0.278 | 31.14 | 1.14 |
| hedge_no | 12.79 | 2.38 | 0.345 | 35.21 | 0.95 |
| assert_no | 15.00 | 3.50 | 0.378 | 36.67 | 1.85 |
| strong_no | 15.82 | 4.25 | 0.391 | 37.16 | 2.11 |
| revise_no | 11.24 | 2.38 | 0.309 | 33.56 | 1.26 |
| neutral | — | — | 0.147 | — | — |

Table 10: End-of-chain displacement decomposed by whether the final answer changed from the unprefilled run ($n = 2{,}000$ per condition). The neutral prefill carries no answer, so displacement is not signed for it and only its flip rate is reported.

Displacement is not a graded quantity that a commitment pushes further or less far. Questions whose answer changes move by 28–40 logits in every condition and in both runs; questions whose answer does not change move by 0.4–3.9. The between-condition spread in these two columns is small next to the gap between them, so mean $d_T$ is close to the flip rate multiplied by a roughly constant step, and medians, which fall on the unflipped majority everywhere, are much smaller than means. The content-free neutral prefill flips 4.8% of answers with a passage and 14.7% without: whatever the passage supplies, it also stabilises the answer against a perturbation that asserts nothing.

On questions where commitment, gold answer and unprompted answer all agree, 68% show negative displacement but none change answer, and the negative values are concentrated in the most confident quartile.

## I DETECTION PROBE

| condition | "own" wording | "inserted" wording | called foreign |
|---|---|---|---|
| unprefilled | $-6.00$ | $-17.28$ | 0.000 |
| assert_yes, its own chain | $-6.02$ | $-14.53$ | 0.000 |
| assert_yes, clean chain | $-2.20$ | $-16.12$ | 0.000 |
| labelled as another model's | $-0.90$ | $-6.75$ | 0.141 |
| user turn, another model | $-4.33$ | $-15.14$ | 0.000 |

Table 11: Detection margins after the chain (positive = "not my own"). The two wordings are of opposite polarity. In the user-turn condition the first line of the reply is the model's own, so "not foreign" is correct.

Within the labelled condition, questions where the model calls the line foreign are those where the commitment conflicts most with the model's unprompted answer. Post-chain detection predicts non-compliance with logistic coefficient $-1.17$ where the commitment is correct and $-0.81$ where it is wrong (both $p < 10^{-15}$, adjusting for the unprompted margin); the pre-chain probe gives $+0.14$ ($p = 0.11$) and $+0.34$ ($p = 0.014$).

## J RECOVERY

Counting a trajectory as displaced if it moves at least 2 logits toward the commitment and as recovered if it ends within 1 logit of the unprefilled margin, recovery among displaced trajectories is rare: 0.6–2.1% for hedged and revision-licensed commitments, 2.1–5.8% for emphatic, 4.5% and 15.0% for plain "no" and "yes". It does not decrease with strength. We do not rely on this measure. Every recovery occurs only at the final step; the 1-logit window is narrow against margins of $\pm 20$; the plain-"yes" figure is inflated by questions whose unprompted margin was already near its ceiling; and varying the window from 0.5 to 5 logits moves the pooled rate by a factor of six without preserving the ordering across conditions. The logit scale supplies no natural unit for "close".

## K IMPLEMENTATION

The probe appends a blank line before the answer suffix; without it the suffix is concatenated to the last step and probe mass drops. The answer tokens are " Yes" (id 7414) and " No" (2308); " True"/" False" take the mass on the few dropped probes. Attenuated generation uses eager attention and a hand-written greedy loop carrying a four-dimensional additive bias; at $\alpha = 1$ it matches standard generation at the same rate as an independent regeneration of the same condition. Stray thinking tags emitted at low $\alpha$ are stripped before probing. Chains are split into steps before any *Step N:* label that follows whitespace, which also separates steps written on one line; concatenating the pieces reproduces the chain exactly. In the no-passage masking run the unmasked control uses the same commitment as the masked condition with the mask switched off, rather than an empty commitment as in the with-passage run; the two therefore differ only in whether the commitment was visible during generation. The no-passage generation run was interrupted and recovered from its own output, so its files end in a truncated record; the loader skips unparseable lines and the

counts reported exclude them. The forced-verification run (§4.4) covers 4,980–5,670 questions per condition rather than the full grid; Table 3 is computed on the intersection with the unprefilled run, and all comparisons there are paired per question.