Who Owns the Answer? Mechanistic Evidence for State Arbitration Between Written Reasoning and Internal LLM Computation

Anonymous ACL submission

Abstract

Scratchpads turn written intermediate reasoning into context that can compete with a large language model's own computation. We study this competition as scratchpad arbitration. In deterministic symbol tasks, a corrupted scratchpad and the rule table imply different answer options, yielding matched trace-owned and rule-owned runs. In Qwen3-4B, the trace-implied state, rule-implied state, and ownership-relevant variables are represented before the answer token. Causal patching then switches which state controls the answer. A sparse pathway transfers trusted targets toward the rule answer and skeptical targets back toward the trace answer. The effect replicates on held-out items, is necessary under ablation, and localizes to layer-25 heads L25H29/H30/H31 with a smaller layer-36 head L36H26 contribution. Source, attention, and temporal decompositions place the decisive signal at the options interface and the answer position, with little causal contribution from the local corrupted trace step. Written reasoning thus behaves as an external state proposal whose control can be decoded, localized, and causally reassigned.

1 Introduction

Large language models (LLMs) increasingly condition on artifacts written during the same interaction. Chain-of-thought rationales, scratchpads, retrieved passages, tool outputs, critiques, and memories all enter the context window before later tokens are generated (Nye et al., 2021; Wei et al., 2022; Guu et al., 2020; Lewis et al., 2020; Yao et al., 2023; Shinn et al., 2023; Madaan et al., 2023). These artifacts expose text that can be inspected as reasoning evidence and supply tokens that can alter subsequent computation.

A prompt can contain a scratchpad whose intermediate state implies one answer while the underlying task implies another. Surface behavior identifies the selected answer, but it does not identify whether the model encoded both candidates, where the decision was made, or which components granted control to one state. We say that the state determining the answer token *owns* generation.

Prior work provides closely related tools and benchmarks. Faithfulness work examines whether written rationales reflect the computation behind the answer (Turpin et al., 2023; Lanham et al., 2023; Paul et al., 2024). Perturbation and process-verification work studies corrupted reasoning chains and erroneous steps (Cobbe et al., 2021; Lightman et al., 2023; Zheng et al., 2025; He et al., 2025; Aravindan and Kejriwal, 2026). These framings usually take a written rationale, a perturbed step, or a verifier judgment as the object of analysis. We analyze state ownership. A written trace, a recomputed answer, a retrieved passage, or a tool result can each propose a state. When such proposals conflict, the mechanistic target is the state that obtains control over generation.

We introduce *scratchpad arbitration* as a controlled setting for this target. Each example contains deterministic symbol-transition rules, answer options, and a scratchpad trace. In conflict examples, the visible trace follows a self-consistent wrong branch, while the rule table implies a different final option. The construction assigns three usually coupled quantities to distinct observables. The trace state is written in the scratchpad, the rule state is recoverable from the task, and ownership is expressed by the answer. We use Qwen3-4B as a tractable model for mechanism discovery (Yang et al., 2025).

The resulting mechanism is sparse and reversible. Behaviorally, the same conflict can be trace-owned or rule-owned depending on framing. Representationally, linear probes recover the rule state, trace-following decision, and ownership-relevant variables before the final answer token. Causally, answer-time patching decides which state is read

out. Patching skeptical-source states into trace-owned targets raises rule ownership from 15.2% and 16.0% to 60.2% and 61.3%. Patching trusted-source states in the reverse direction lowers rule ownership from 66.7% to 22.1% and raises trace following to 75.8%. The path moves control in both directions.

Mechanistic localization identifies the relevant pathway. The dominant writers are layer-25 heads L25H29/H30/H31, with a smaller layer-36 head L36H26 contribution. Ablating these heads collapses decisive ownership in both rule-owned and trace-owned regimes, while matched control heads leave behavior nearly unchanged. The strongest source of the L25 signal is the options interface where answer candidates are present, with little contribution from the local erroneous trace step or the rule-conflict row. Temporal patching places the decisive causal operation at answer readout. The model can carry ownership-relevant information earlier, but answer-time control determines which represented state becomes the next token. The pathway acts after candidate states are available, when they are converted into answer logits.

Our key contributions are as follows:

- We introduce *state ownership* as a unit of mechanistic analysis for scratchpad-conditioned reasoning, treating written traces and internally computed endpoints as competing state proposals.
- We design a controlled identification setting for ownership: a deterministic conflict task that independently labels the trace state, the rule state, and answer ownership.
- We show that an LLM can represent competing state proposals before selecting one. In Qwen3-4B, matched framings switch the same corrupted scratchpad between trace-owned and rule-owned answers while both candidate states remain linearly readable.
- We localize the causal ownership pathway to answer time. A small set of L25 heads, with a smaller L36 component, transfers ownership in both directions, replicates on held-out items, and is necessary for decisive natural behavior.
- We derive and test a first-order margin-control account of arbitration. Bidirectional patching, graded interpolation, ablation, source replacement, and temporal localization support an answer-time rule–trace margin read mainly from the answer-option interface.

2 Method and Experimental Setup

2.1 Controlled State-Conflict Task

We use synthetic deterministic tasks so that arbitration is not confounded with open-ended mathematical difficulty. Each example contains a start symbol, a table of deterministic transitions, a multiple-choice answer set, and a written scratchpad trace. Applying the rule table from the start symbol yields a unique final state. Reading the scratchpad as written yields another final state when the scratchpad is corrupted.

The main manipulation corrupts the written scratchpad while leaving the original rules and answer options fixed. In the main condition, `corrupted_no_final`, the first scratchpad transition is corrupted and the remainder of the trace follows the resulting wrong branch. The trace and rule states therefore point to different options. The variant removes the final answer line from the trace, so the model cannot answer by copying a final written symbol. All main mechanistic experiments use depth-2 tasks in the `rules_only` prompt style and Qwen3-4B as the target model (Yang et al., 2025).

2.2 States and Labels

The generator provides labels independently of model behavior. The *rule state* is the endpoint implied by the deterministic transition table. The *trace state* is the endpoint implied by the scratchpad as written. A model output is rule-owned when the selected option matches the rule endpoint, and trace-owned when it matches the trace endpoint. These labels identify which candidate state controls the answer and do not depend on model explanations.

We also train probes for ownership-relevant variables. The `correct_rule` label marks whether the answer follows the rule state, `follow_trace` marks whether it follows the trace state, and `prefers_rule` marks whether the rule option is preferred over the trace option. These labels distinguish candidate-state identity from the decision that grants control to one candidate.

2.3 Matched Framing Conditions

We instantiate each conflict item under trusted, neutral, untrusted, and skeptical instruction framings. The transition table, trace tokens, and answer options remain fixed across framings. The paired construction changes ownership pressure while holding task content constant. For trace-to-rule transfer,

skeptical source activations are patched into trusted or neutral targets. For rule-to-trace transfer, trusted source activations are patched into skeptical targets. Whenever possible, source and target runs use the same item, so the contrast attributes ownership changes to framing and activation source, with task content held fixed. Unless otherwise stated, behavioral and module-level analyses use 256 examples, source-decomposition analyses use 384 examples, and attention-source measurements use 96 neutral conflict examples.

2.4 Metrics

Let s_{rule} and s_{trace} be the answer-token logits assigned to the rule and trace options. We report rule ownership, trace ownership, and the arbitration score (ASI)

$$\text{ASI} = s_{\text{rule}} - s_{\text{trace}}.$$

Positive ASI indicates a rule-state advantage, and negative ASI indicates a trace-state advantage. For interventions we report ΔASI relative to the matched baseline. Rule and trace ownership rates need not sum to one because the model can choose a distractor option.

2.5 Mechanistic Analyses

We collect residual-stream activations at `rules_end`, `trace_end`, `question_end`, and `answer`. Linear probes are trained to predict state and ownership labels, with both within-distribution splits and cross-framing splits (Elazar et al., 2021). The cross-framing setting trains on trace-heavy framings and tests on rule-heavy framings, measuring whether the ownership variable transfers across prompt regimes.

Causal interventions replace module outputs or attention-head outputs from a source run into the matched target run (Geiger et al., 2021; Vig et al., 2020; Goldowsky-Dill et al., 2023; Conmy et al., 2023). The candidate path contains layer-25 heads L25H29/H30/H31 as an upstream group and layer-36 head L36H26 as a downstream contributor (Wang et al., 2022). We test sufficiency with bidirectional source-to-target patching, necessity with shuffle ablations, source-token dependence with category-restricted replacement, and timing with position-specific patches. Source-restricted interventions apply the same replacement only to activations associated with specified prompt regions, such as the options interface, the rule table, or the corrupted trace step. Interpolation experiments scale the L25 question/options component before insertion into the target run. Control groups use other heads in layer 25 and an earlier layer-12 group.

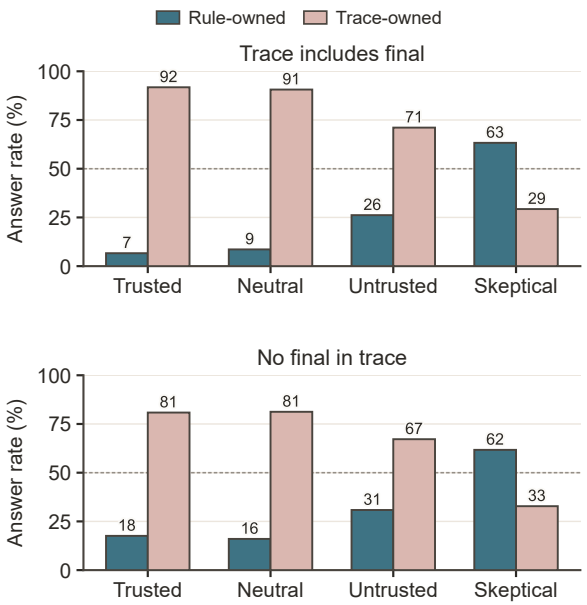

Figure 1: Behavioral arbitration under fixed corrupted traces. The answer distribution shifts between trace ownership and rule ownership as the framing changes, even when the rule table, trace tokens, and answer options are fixed.

3 Results

3.1 Behavioral Arbitration Under Fixed Conflicts

Figure 1 reports the framing effect before any causal intervention. The examples are matched across trusted, neutral, and skeptical framings while the visible trace and answer options remain fixed. Trusted and neutral framings yield predominantly trace-owned answers. Skeptical framing shifts the same items toward rule ownership. When the trace includes a final answer line, trace ownership is 92% under trusted framing and 91% under neutral framing, while skeptical framing raises rule ownership to 63%. Removing the final trace line lowers direct copying but preserves the ordering, with skeptical framing still reaching 62% rule ownership.

Under matched task content, a corrupted trace can therefore produce either trace ownership or rule ownership. The remaining analyses test whether the relevant states are represented and which internal pathway controls the ownership decision.

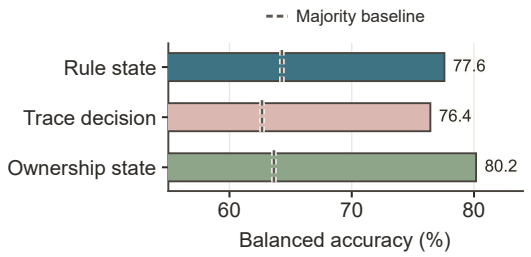

Figure 2: Answer-position probes recover rule state, trace decision, and ownership state above majority baselines.

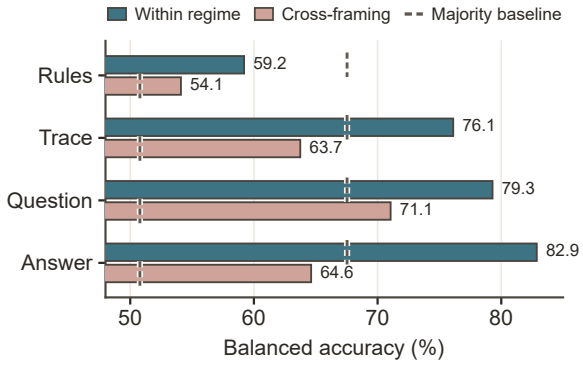

Figure 3: Pre-answer ownership probes. Balanced accuracy exceeds majority baselines at trace, question, and answer positions, and the question-boundary signal transfers across framing regimes.

3.2 State and Ownership Variables Are Readable

Answer-position probes recover the rule-implied state, the trace-following decision, and the ownership variable that predicts generation. Figure 2 shows balanced accuracies of 77.6%, 76.4%, and 80.2%, each above its majority baseline. The answer position therefore preserves readable information about both candidate states and the decision over which state controls generation.

Figure 3 shows that ownership is also readable before the answer token. Within conflict examples, a `prefers_rule` probe reaches 76.1% balanced accuracy at `trace_end`, 79.3% at `question_end`, and 82.9% at `answer`, compared with a 67.5% majority baseline. Under cross-framing transfer, training on trace-heavy framings and testing on rule-heavy framings, the same probe reaches 71.1% at `question_end`. The `correct_rule` and `follow_trace` probes reach 70.0% and 69.9% at the same position.

Cross-framing transfer weakens a wording-cue account. A probe trained where the trace usually wins still transfers to settings where the rule usually wins. The probe results show availability before final readout, while the patching results identify a sparse causal pathway that selects which represented state determines the answer.

3.3 A Reversible Ownership Switch

Figure 4 shows that the same path transfers ownership in both directions. In trace-owned trusted targets, skeptical-source patching raises rule ownership from 15.2% to 60.2%. In neutral targets it raises rule ownership from 16.0% to 61.3%. Reversing the source and target reverses the sign of the effect. Trusted-source patching into skeptical targets lowers rule ownership from 66.7% to 22.1% and raises trace ownership to 75.8%, with $\Delta\text{ASI} = -3.30$. The discovered head group gives $\Delta\text{ASI} = -3.28$ in this reverse direction, while L25H0/H1/H2 and L12H5/H6/H7 controls remain near zero.

Patch-scale interpolation gives graded control over the same decision. Figure 5 interpolates the L25 decision component in neutral targets. Increasing the question/options component transfers ownership from trace to rule. Rule ownership rises from 17.6% at $\alpha = 0$ to 62.5% at $\alpha = 1.5$, while trace following decreases from 81.2% to 34.4%. The no-question control remains nearly flat, isolating the question/options component as the causal source.

A one-sided skepticism detector would mainly increase rule following. Here, the same heads also make a rule-owned run follow the written trace when the trusted source supplies the corresponding state. The intervention changes which represented state has answer control.

3.4 Module and Path Localization

Module patching localizes the ownership transfer. Table 1 reports the largest answer-token effects. The layer-25 attention block produces the strongest shift, raising rule ownership to 57.0% and increasing ΔASI by 3.04. The next-largest module, the layer-35 multilayer perceptron (MLP), shifts ΔASI by 1.14. The gap motivates the head-level decomposition of the L25 attention block.

Module	Layer	Rule	Trace	ΔASI	ΔRule
Attn	25	57.0	39.5	3.04	38.7
MLP	35	32.8	66.8	1.14	14.5
MLP	29	31.6	68.0	0.89	13.3
Attn	31	30.1	69.9	0.83	11.7
Attn	22	28.9	71.1	0.70	10.5
MLP	26	25.8	74.2	0.58	7.4

Table 1: Largest answer-token module-patching effects. Attn denotes attention. Rule, trace, and ΔRule are reported as percentages or percentage points.

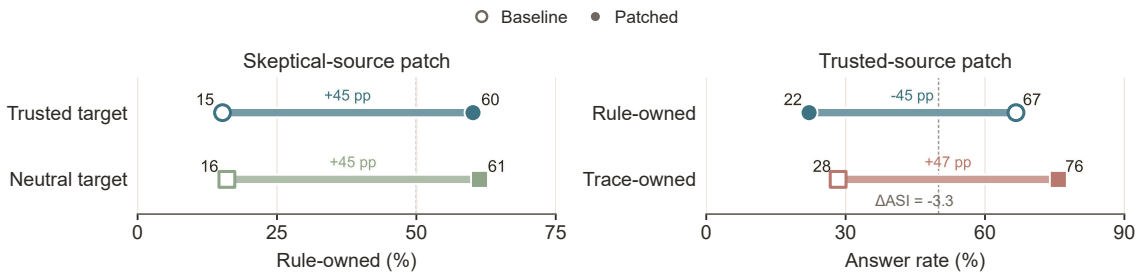

Figure 4: Bidirectional ownership transfer. Skeptical-source states move trusted and neutral targets toward the rule answer, while trusted-source states move skeptical targets toward the trace answer.

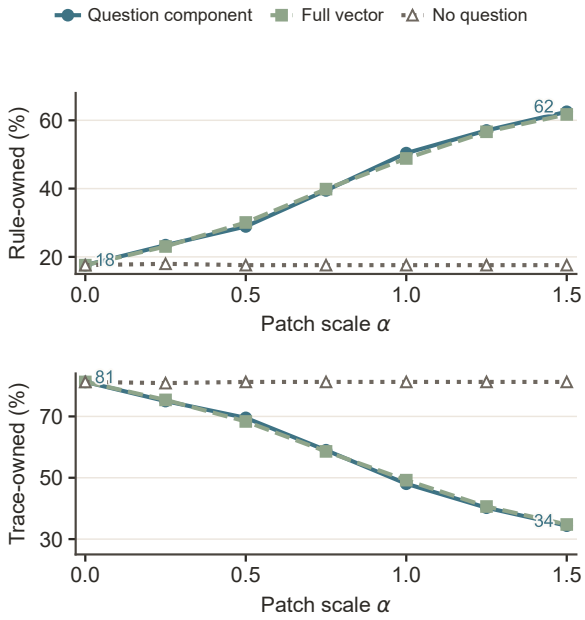

Figure 5: Continuous intervention on the L25 component. Scaling the question/options component produces a graded increase in rule ownership and a corresponding decrease in trace following, while the no-question control remains nearly unchanged.

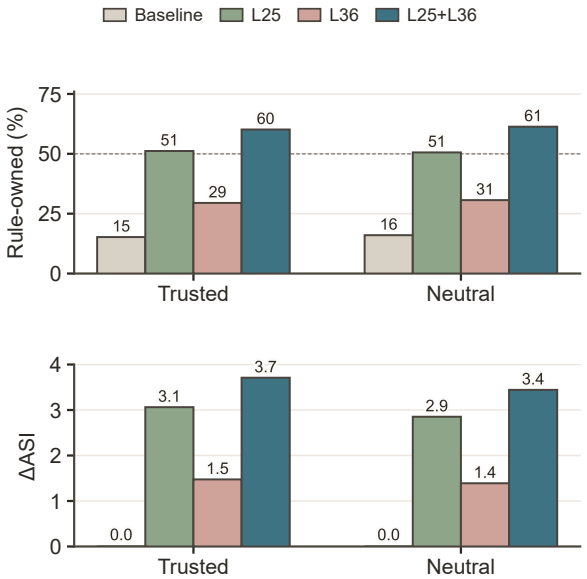

Figure 6: Held-out path patching. L25H29/H30/H31 provide the dominant ownership write, and L36H26 adds a smaller answer-side contribution.

Figure 6 confirms that the causal effect is stable and identifies the main writer. In trusted targets, L25H29/H30/H31 alone raises rule ownership from 15.2% to 51.2% with $\Delta\text{ASI} = 3.06$, L36H26 alone reaches 29.5% with $\Delta\text{ASI} = 1.47$, and the combined patch reaches 60.2% with $\Delta\text{ASI} = 3.71$. Neutral targets reproduce the ordering, with 50.6%, 30.7%, and 61.3% rule ownership. L36H26 adds a late readout contribution, while L25 writes most of the ownership transfer.

3.5 The Pathway Is Necessary in Natural Runs

Necessity ablations test whether the discovered heads are used in natural runs. Table 2 reports the results. In skeptical targets, shuffle ablation of L25H29/H30/H31 plus L36H26 reduces rule ownership from 68.0% to 27.1% and shifts ΔASI by -1.19. L25 alone accounts for most of the drop. In untrusted targets, where the baseline follows the trace, the same ablation reduces trace ownership from 67.2% to 28.5% and shifts ΔASI by 1.31. Matched L25 and L12 control groups leave ownership nearly unchanged.

The ablation reduces decisive ownership without pushing the model toward a fixed opposite answer. Rule-owned skeptical runs lose their rule advantage, and trace-owned untrusted runs lose their trace advantage. The effect supports a control role for the path, with no evidence for a head-specific bias toward either answer class.

3.6 The Decisive Source Is the Options Interface

Source-restricted replacement identifies the prompt region used by the L25 heads. A step-verification account predicts causal signal from the corrupted trace step or the rule row that exposes the conflict. Figure 7 shows a different pattern. For trusted targets, full L25 replacement gives $\Delta\text{ASI} = 3.13$, while replacing only `options_all` gives 3.12. In

Target	Intervention	Rule	Trace	ΔASI
Skeptical	Baseline	68.0	26.8	0.00
Skeptical	L25+L36 shuffle	27.1	28.1	-1.19
Skeptical	L25 shuffle	34.1	28.4	-0.93
Skeptical	Control shuffle	68.0	26.8	-0.00
Untrusted	Baseline	30.1	67.2	0.00
Untrusted	L25+L36 shuffle	27.0	28.5	1.31
Untrusted	L25 shuffle	30.1	32.0	1.20
Untrusted	Control shuffle	30.1	67.2	-0.00

Table 2: Necessity ablations. Disrupting the discovered heads collapses decisive ownership in both regimes. Rule and trace columns are percentages.

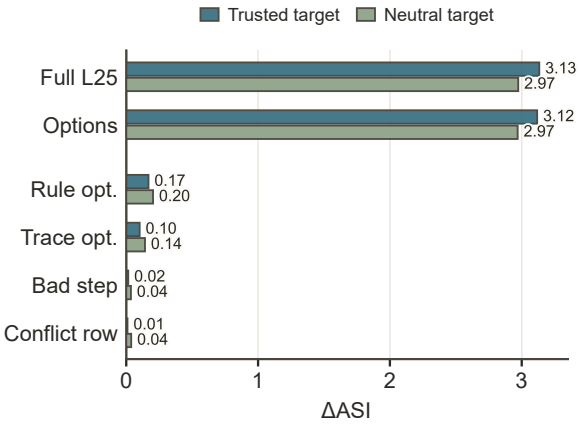

Figure 7: Source decomposition for L25H29/H30/H31. Options-only replacement recovers nearly the full effect, while local trace and rule-conflict tokens contribute little.

Head	Ques.	Rules	R opt.	T opt.	Bad	Row
L25H29	74.9	20.2	10.2	7.4	0.16	0.21
L25H30	44.9	36.8	2.1	2.6	0.07	0.09
L25H31	75.8	21.3	7.0	7.6	0.18	0.10
L36H26	18.7	53.1	0.1	0.2	0.04	0.03

Table 3: Attention source mass for discovered heads in neutral conflict examples. Values are percentages. *Ques.* denotes the question/options interface. *R opt.* and *T opt.* denote the rule and trace options. *Bad* denotes the corrupted trace step, and *Row* denotes the rule-conflict row.

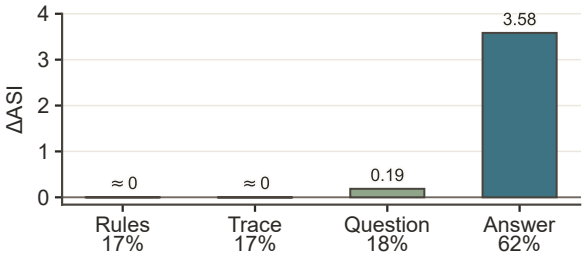

Figure 8: Temporal localization. The ownership path has little effect at rule, trace, or question endpoints, then becomes decisive at answer readout.

neutral targets the two values are both 2.97. Local sources are small. In trusted targets, the corrupted trace step gives 0.015, the rule-conflict row gives 0.009, and the question stem gives 0.009. Neutral targets show the same ordering.

Attention-source analysis follows the causal result. Table 3 reports attention mass for the discovered heads in neutral conflict examples. L25H29 and L25H31 concentrate on the question/options interface, assigning 74.9% and 75.8% of attention mass to that region. Their mass on the corrupted trace step is below 0.2%. L36H26 assigns 53.1% mass to the rules region, consistent with its smaller downstream readout contribution. Source categories can overlap because option lines are also part of broader prompt regions.

The options interface supplies the causal signal. The local corrupted step contributes little, and the pathway writes an ownership signal over the competing rule and trace candidates.

3.7 The Causal Operation Acts at Answer Readout

Temporal patching localizes where the L25/L36 pathway affects behavior. Figure 8 shows that patching at `rules_end` and `trace_end` has essentially no effect. The corresponding ΔASI values are -0.002 and -0.004, and rule ownership remains about 17%. Patching at `question_end` is still weak, with $\Delta\text{ASI} = 0.185$ and rule ownership 18.4%. At the answer position, the same path produces $\Delta\text{ASI} = 3.58$ and raises rule ownership from 16.8% to 62.5%.

The temporal profile places the causal operation at answer readout. Earlier positions contain ownership information, as the probes show, but the decisive intervention acts when the answer token is read out. Across the interventions, the pathway is sufficient to reassign ownership, necessary for confident natural ownership, source-specific to the decision interface, and temporally localized to answer readout. These tests show that representing a candidate state and granting it answer control are different operations.

4 State Arbitration as Answer-Time Margin Control

The controlled task fixes two endpoint labels for each conflict item before any activation intervention is applied. Let $x = (R, \tau, O, c)$, where R is the rule table, τ is the written trace, O is the option set, and c is the framing. The rule table induces $y_r = f_R(s_0)$. The trace induces $y_t = f_\tau(s_0)$. Conflict items satisfy $y_r \neq y_t$. The two endpoint labels are fixed by the input construction, independent of the model's final answer.

Let $h_A(x, c)$ denote the residual stream at the answer position, and let $u(y)$ denote the unembedding vector for option y. Define the rule–trace direction

$$d_{r,t} = u(y_r) - u(y_t).$$

The arbitration score used throughout the experiments is the logit margin

$$\mathrm{ASI}(x, c) = d_{r,t}^\top h_A(x, c).$$

The sign of ASI gives the owned endpoint under greedy readout. Positive values select the rule endpoint, and negative values select the trace endpoint. Its magnitude measures the strength of that ownership decision.

Matched framing comparisons change c while holding R, τ, O, y_r, and y_t fixed. A framing effect on ASI must therefore enter through the answer-position representation. We write the framing-sensitive part as a first-order mediator

$$h_A(x, c) = b(x) + g(x, c)a(x) + \epsilon(x, c),$$

where $b(x)$ is item-specific computation shared across framings, $a(x)$ is an arbitration direction, $g(x, c)$ is a scalar ownership mediator, and $\epsilon(x, c)$ captures remaining variation. Projection onto $d_{r,t}$ gives

$$\begin{aligned} \mathrm{ASI}(x, c) &= \beta(x) + \gamma(x)g(x, c) + \eta(x, c), \\ \gamma(x) &= d_{r,t}^\top a(x). \end{aligned}$$

When $\gamma(x) > 0$, increasing g moves a fixed item toward rule ownership. Decreasing g moves the same item toward trace ownership. Since the candidates and options are fixed, g captures the framing-dependent part of ownership.

Activation patching estimates whether a module contributes to this mediator (Geiger et al., 2021; Vig et al., 2020; Meng et al., 2022; Goldowsky-Dill et al., 2023; Conmy et al., 2023). Let H be a module or head group, and let P_H replace only its answer-position contribution. Patching a source run s into a target run t gives

$$h_A^{t \leftarrow s, H} = h_A^t + P_H(h_A^s - h_A^t).$$

The induced margin change is, to first order,

$$\Delta\mathrm{ASI}_H \approx d_{r,t}^\top P_H(h_A^s - h_A^t).$$

If H carries the ownership mediator, a source with larger g should raise ASI, and reversing source and target should lower ASI. Figure 4 matches this sign constraint. Skeptical-source patching moves trusted and neutral targets toward the rule answer. Trusted-source patching moves skeptical targets toward the trace answer.

Scaled patching gives a graded version of the same intervention. For

$$\begin{aligned} h_A(\alpha) &= h_A^t + \alpha P_H(h_A^s - h_A^t), \\ \mathrm{ASI}(\alpha) &\approx \mathrm{ASI}(0) + \alpha\Delta\mathrm{ASI}_H, \end{aligned}$$

the rule–trace margin should change approximately monotonically with α. Figure 5 shows this pattern for the L25 component. Rule ownership increases as the patched component is strengthened, and trace ownership decreases. Necessity ablations give the corresponding intervention in the opposite direction. If H supplies a mediator, disrupting it should reduce the active margin in the current regime. An answer-class bias would push both regimes in the same direction. Table 2 follows the mediator prediction. Skeptical runs lose rule ownership, untrusted runs lose trace ownership, and matched control heads remain inert.

The source analysis estimates which prompt region supplies the mediator to the answer-position heads. For an attention head h,

$$o_h(A) = \sum_{j \leq A} \alpha_{A,j}^{(h)} V^{(h)} x_j W_O^{(h)}.$$

For a token region C, its first-order contribution to ownership is measured by

$$\begin{aligned} z_j^{(h)} &= \alpha_{A,j}^{(h)} V^{(h)} x_j W_O^{(h)}, \\ \Delta_C &\approx d_{r,t}^\top \sum_{j \in C} \Delta z_j^{(h)}. \end{aligned}$$

A local trace-validity mechanism assigns large contribution to the corrupted trace step or to the rule row that exposes the conflict. An endpoint-readout mechanism assigns large contribution to the options interface, where both answer candidates are

present. Figure 7 and Table 3 follow the endpoint-readout prediction. Options-only replacement recovers nearly the full L25 effect, and the corrupted step and conflict row contribute little.

Temporal patching locates where represented information enters the answer margin. Probes show that candidate states and ownership-related variables are readable before generation. Let $\Delta\mathrm{ASI}_H(p)$ denote the effect of applying the same path patch at position p. Figure 8 shows near-zero $\Delta\mathrm{ASI}_H(p)$ at rule, trace, and question endpoints, followed by a large effect at the answer position. The pathway is thus an answer-readout operator over already represented endpoints.

The constraints distinguish state ownership from skepticism, copying, and local verification. Skepticism predicts movement away from corrupted traces without the reverse transfer toward the written trace. Copying predicts dependence on local trace tokens. Local verification predicts causal contribution from the corrupted step or the conflict row. The observed L25/L36 pathway changes $d_{r,t}^{\top} h_A$ at the point where answer logits are read, using the interface where the two endpoints are jointly available. In this setting, state ownership is the sign and magnitude of that answer-time margin.

5 Related Work

Written reasoning and faithfulness. Chain-of-thought prompting and scratchpads made intermediate text a standard interface for eliciting multi-step behavior from language models (Wei et al., 2022; Nye et al., 2021; Kojima et al., 2022; Wang et al., 2023). Work on faithfulness studies whether written rationales expose the computation behind an answer or behave as post-hoc text whose perturbation leaves the latent decision mostly unchanged (Turpin et al., 2023; Lanham et al., 2023; Paul et al., 2024; Jacovi and Goldberg, 2020; Lyu et al., 2023). These studies define faithfulness through the relation between a generated rationale and the answer it accompanies. Here, the written trace is an input state proposal whose endpoint conflicts with the endpoint entailed by the task, so answer ownership is directly measurable.

Corrupted traces and process verification. Perturbation studies examine how models behave when intermediate reasoning is wrong, misleading, or inconsistent (Turpin et al., 2023; Aravindan and Kejriwal, 2026). Process-supervision benchmarks evaluate whether models, reward models, or critics can identify erroneous reasoning steps and localize the earliest error (Cobbe et al., 2021; Lightman et al., 2023; Uesato et al., 2022; Zheng et al., 2025; He et al., 2025). Their evaluation targets are step validity, error position, and the effect of an erroneous chain on the final prediction. We use corrupted traces as an experimental instrument. The corruption creates two labeled answer states, and the analysis measures which state obtains control over generation.

Self-verification and critique. Self-verification, self-correction, and critique prompting instruct models to inspect intermediate reasoning or previous answers before revising their outputs (Shinn et al., 2023; Madaan et al., 2023; Welleck et al., 2023; Gou et al., 2024; Lee et al., 2025). These methods combine explicit judgment with answer revision. Scratchpad arbitration fixes the candidate states and measures answer control directly. A model may judge the trace correctly, ignore that judgment, or never verbalize a judgment. The measured variable is whether the trace-induced state or the internally recomputed state controls the answer token.

External context as model state. Modern language-model systems condition on retrieved passages, tool outputs, memory summaries, and other external artifacts (Guu et al., 2020; Lewis et al., 2020; Karpukhin et al., 2020; Borgeaud et al., 2022; Schick et al., 2023; Yao et al., 2023; Gao et al., 2023; Patil et al., 2023; Asai et al., 2024; Mialon et al., 2023; Packer et al., 2023). Prior work usually studies these artifacts as sources of information, grounding, or action. The same interface also permits conflicts between an externally supplied state and a state recoverable from the prompt. Scratchpads provide a minimal labeled instance of this problem, with an external trace endpoint and a rule-implied endpoint available for causal comparison.

6 Conclusion

Scratchpad arbitration treats written reasoning as a state proposal that conflicts with internal computation. In Qwen3-4B, trace state, rule state, and ownership variables are readable before answer generation, while a sparse answer-time pathway determines readout. The L25/L36 pathway is bidirectional, necessary, replicated, options-sourced, and answer-localized. These results frame faithfulness as answer ownership when states conflict.

Limitations

We study Qwen3-4B on one controlled scratchpad-conflict task family. The localized L25/L36 pathway gives evidence for arbitration in this setting, without establishing a universal circuit. Larger, instruction-tuned, multimodal, or tool-using systems may implement related arbitration behavior through different components or at different positions.

Our task creates clean two-state conflicts between a scratchpad trace and a recomputable rule state. Real deployments often involve many partially compatible states, including retrieved evidence, world knowledge, tool outputs, dialogue history, memory, and policies. Extending ownership analysis to these settings will require conflict designs that preserve ecological validity while still supporting causal tests.

The identified pathway should not be interpreted as the only way a model can arbitrate ownership. The selected experiments show that this L25/L36 answer-time pathway is reversible, necessary, source-specific, and temporally localized in the studied setting. They do not rule out additional pathways, training-dependent mechanisms, or task regimes in which arbitration occurs earlier or is distributed across more of the network.

References

Ashwath Vaithinathan Aravindan and Mayank Kejriwal. 2026. Fragile thoughts: How large language models handle chain-of-thought perturbations. *Preprint*, arXiv:2603.03332.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*.

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

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adria Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*.

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

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

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*.

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, and Bo Zheng. 2025. Can large language models detect errors in long chain-of-thought reasoning? *Preprint*, arXiv:2502.19361.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

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

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

Andrew Lee, Lihao Sun, Chris Wendler, Fernanda Viegas, and Martin Wattenberg. 2025. The geometry of self-verification in a task-specific reasoning model. *Preprint*, arXiv:2504.14379.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *Preprint*, arXiv:2301.13379.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594.

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

Gregoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: A survey. *Preprint*, arXiv:2302.07842.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *Preprint*, arXiv:2112.00114.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *Preprint*, arXiv:2310.08560.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large language model connected with massive APIs. *Preprint*, arXiv:2305.15334.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

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

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *Preprint*, arXiv:2211.14275.

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

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *Preprint*, arXiv:2211.00593.

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

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

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi.

2023. Generating sequences by learning to self-correct. In *International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Processbench: Identifying process errors in mathematical reasoning. *Preprint*, arXiv:2412.06559.