No-Think Is Not No-Thought: The White Bear Effect in LLM Reasoning Control

Anonymous ACL submission

Abstract

In large language models with explicit reasoning controls, a final no-think directive appears to close the reasoning channel. Whatever mode history preceded it, the model should now answer directly. We show that this latest-control picture is false in a way that exposes a mechanism of negated instruction following. We construct matched-final histories in which the current no-think instruction, question, and answer contract are identical, and only the earlier mode controls differ. Across Qwen3-family models, these histories remain separated at the moment of answer formation even when visible reasoning markers are absent. The central surprise is directional. Prior no-think and alternating histories can make a later no-think state more, not less, poised toward reasoning tokens. We interpret this as a White Bear effect in reasoning control. Suppressing chain-of-thought can preserve it as something the model must not say. Causal interventions locate the preserved state at the assistant answer boundary rather than in the literal control token or arbitrary prompt positions. Learned boundary coordinates bidirectionally steer reasoning-minus-direct pressure across scales, while matched random directions do not. Thus no-think is not simply an erasure command. It is a stateful control operation that can produce surface compliance while retaining a causal, boundary-local representation of the reasoning behavior being withheld.

1 Introduction

Reasoning controls turn hybrid-thinking large language model (LLM) inference into a state-setting problem. In these models, think and no-think do not merely request different surface styles. They select inference-time regimes for how the model prepares an answer. The natural abstraction is a resettable state machine. Once the current control is no-think, earlier mode controls should no longer matter. We test the alternative that no-think is implemented as a boundary-local control state, so the model can visibly answer directly while still carrying information about the reasoning behavior it is suppressing.

Negated controls make this possibility concrete. To obey "do not think step by step," a model must identify the behavior being prohibited and bind it to the current answer policy. Cognitive psychology describes an analogous structure as the White Bear effect, in which attempts to suppress a thought can keep it available because suppression requires monitoring the content to be excluded (Wegner et al., 1987). In a reasoning-control model, no-think can instantiate the same structure. It hides chain-of-thought from the response while preserving a latent representation of the behavior being withheld.

Hybrid-thinking models expose this phenomenon through explicit think and no-think controls within a single model family (Yang et al., 2025). They also give a clean counterfactual. Take two prompts with the same final no-think instruction, the same question, and the same answer contract, and vary only the prior mode-control history. A stateless switch predicts convergence to the same answer-boundary computation. A history-dependent controller predicts distinct latent states under the same current directive.

The matched-final design reveals the second regime. Visible reasoning markers remain exactly zero while first-token reasoning pressure moves substantially. In the 14B matched-final evaluation, prior no-think raises $\log P(V_{\text{reason}})$ by $+1.789$, and alternating think/no-think histories raise it by $+3.705$, under the same current no-think instruction. The sharper surprise is directional. Prior no-think histories can produce stronger latent reasoning pressure than prior think histories. The apparent eraser becomes part of the state that keeps reasoning available but unspoken.

The residue has the geometry of an answer-boundary register. Across semantic-position analyses, it concentrates at the point where the assis-

tant is about to emit its first answer token. In 32B layer 63 (L63), the learned history direction is read at immediate and delayed answer starts, and adding this direction changes reason-minus-direct (RMD) pressure by $+0.191/+0.211$ at those answer boundaries while leaving non-answer positions at 0.000. Learned coordinates bidirectionally steer RMD pressure across 4B, 8B, 14B, and 32B models. Matched random directions and position placebos remain near zero. The residue is therefore a causal, boundary-local control state.

The paper makes three contributions. First, matched-final histories turn reasoning-control evaluation into a same-instruction, different-state test. Second, the White Bear effect gives a compact account of negated reasoning control. Selecting no-think mode can preserve a latent representation of the suppressed behavior. Third, the mechanism is localized to an answer-boundary coordinate that is separable from generated derivations, causally steerable, and sharper at larger scale. No-think becomes a mode-control operator, a state-space control problem at the boundary between latent reasoning and the first answer token.

2 Experimental Design

Matched-final histories. We refer to `/think` and `/no_think` as think and no-think controls. Each item fixes the final instruction block $s(q)$, the question q, and the answer contract, then varies only the preceding control history h:

$$x(h, q, s) = \mathrm{concat}(h, s(q)),$$

where s contains the current no-think instruction. Histories include a neutral baseline, prior think, prior no-think, alternating think/no-think, and repeated no-think histories. Since q and s are matched, differences across histories estimate path dependence in hidden states, first-token readouts, or causal effects.

Identification logic. For any readout m, define

$$\Delta_m(h, h'; q, s) = m(x(h, q, s)) - m(x(h', q, s)).$$

The latest-instruction null predicts

$$\mathbb{E}_q \Delta_m(h, h'; q, s) = 0 \quad \forall h, h', s$$

for readouts determined by the current instruction and question. The history-residue alternative adds a response-boundary term:

$$r_{\ell,p^\star}(h, q, s) = a_\ell(q, s) + B_{\ell,s}\phi(h) + \varepsilon,$$

with $B_{\ell,s}\phi(h) \neq 0$ under a final no-think instruction. We test this term through visible text, first-token distributions, hidden-state projections, semantic-position maps, and causal interventions. Projection alone is diagnostic; causal evidence requires localized add/sub effects and random-direction specificity.

Hypotheses and decision criteria. The operational criteria are path dependence, latent-visible dissociation, boundary readout, and causal control. H1 requires $\Delta_m(h, h'; q, s) \neq 0$ for latent readouts under matched final prompts. H2 requires nonzero $\log P(V_{\text{reason}} \mid x)$, $\log P(V_{\text{direct}} \mid x)$, or RMD effects with zero visible reasoning-marker deltas. H3 requires concentration at $p^\star$ and relocation under delayed answers. H4 requires a history-derived direction at $p^\star$ that changes RMD bidirectionally under add/sub and exceeds matched random-direction bands.

Models and prompt families. We study Qwen3-4B, 8B, 14B, and 32B over 2,022 experiment–example units. Prompt families vary the final answer contract while preserving the matched-history comparison. Strict direct-only, no-explanation, and bare no-think contracts probe latent pressure under a clamp. Answer-field and soft-direct contracts test visible release. Tasks include arithmetic, two-step, verbal two-step, answer-format, synthetic symbolic, JSON-answer, lexical, and translation-style items.

Evidence structure. Each claim is attached to the measurement that can identify it. Matched-final prompts test path dependence. Fixed token-set readouts measure answer-boundary pressure before sampling. Hidden-state probes test whether the separation exists before vocabulary projection. Position maps and delayed boundaries test where the state is read. Add/sub interventions, independent examples, and random-direction comparisons test causal specificity.

Estimands. We use three estimands: visible-generation measures, first-token distributional measures, and causal-intervention effects. Visible measures include character count and reasoning-marker count. First-token measures include $\log P(V_{\text{reason}})$, $\log P(V_{\text{direct}})$, and RMD. Causal effects have the form $\Delta m(I(r)) - m(r)$, where m is a logit, entropy, or RMD readout after patching a learned direction. Matched-pair standard errors (SEs) use paired item-level differences. Random-

direction bands are matched by layer, position, history condition, and intervention scale.

Positions. For layer ℓ and semantic position p, let $r_{\ell,p}(h,q,s)$ denote the residual-stream activation for prompt $x(h,q,s)$. The primary position is the assistant answer boundary,

$$p^\star = \text{answer_boundary}(x),$$

the site immediately before the first answer token. Position controls include the final control token or mode-tag end, question marker or question end, prompt middle or prompt start, and pre-answer offsets such as $\text{boundary} - k$. Delayed-answer prompts relocate $p^\star$. A boundary-gated effect should track this semantic relocation.

Metrics. The main readout is RMD pressure,

$$\begin{aligned}\mathrm{RMD}(x) = &\log P(V_{\text{reason}} \mid x) \\ &- \log P(V_{\text{direct}} \mid x),\end{aligned}$$

where V_{reason} and V_{direct} are fixed first-token category sets. We report $\Delta\mathrm{RMD}$, component log probabilities, next-token entropy, and direction rate. Visible output length and derivation probability are secondary readout-threshold measures. Tables write $P_R = P(V_{\text{reason}} \mid x)$ and $P_D = P(V_{\text{direct}} \mid x)$. They abbreviate random-control 5th/95th percentile bands as p05/p95.

The token sets are fixed within each experiment and evaluated before generation. We track $\Delta \log P(V_{\text{reason}})$ and $\Delta \log P(V_{\text{direct}})$ separately because answer contracts can raise reason-token pressure, suppress direct-answer pressure, or combine both effects.

The distinction is most diagnostic under no-think answer contracts. A strict answer contract can keep visible derivation at zero while changing the first-token distribution. A fielded answer can reduce direct-answer pressure without producing long visible chains. Reporting both components prevents these cases from collapsing into a single length or marker statistic.

Mechanism indices. We summarize the evidence with four indices. The mode-hysteresis index is

$$\mathcal{H}_m(h,h';s) = \mathbb{E}_q \Delta_m(h,h';q,s),$$

which vanishes under a strict latest-instruction account. The visible-latent gap compares RMD with visible reasoning markers M_{vis}:

$$\mathcal{G}(h,h';s) = |\mathcal{H}_{\text{RMD}}(h,h';s)| - |\mathcal{H}_{M_{\text{vis}}}(h,h';s)|.$$

The boundary-concentration score compares the answer boundary with non-answer positions $\mathcal{P}_0$:

$$\mathcal{B}_\ell(u) = \frac{|u^\top r_{\ell,p^\star}|}{\epsilon + \frac{1}{|\mathcal{P}_0|}\sum_{p\in\mathcal{P}_0} |u^\top r_{\ell,p}|}.$$

The causal-specificity margin is

$$\begin{aligned}\mathcal{S}(v;\alpha) = \min\{&\delta^+_{\text{RMD}}(v;\alpha) - Q^+_{.95}, \\ &Q^-_{.05} - \delta^-_{\text{RMD}}(v;\alpha)\}.\end{aligned}$$

Positive $\mathcal{H}$, $\mathcal{G}$, $\mathcal{B}$, and $\mathcal{S}$ indicate a history-dependent latent state that is hidden from visible text, localized at answer initiation, and direction-specific under intervention.

Causal vectors. For a contrast history C and neutral baseline N, the basic history direction at layer ℓ is

$$\begin{aligned}v^\ell_C &= \bar{r}_{\ell,p^\star}(C) - \bar{r}_{\ell,p^\star}(N), \\ \bar{r}_{\ell,p}(h) &= \mathbb{E}_{q\in\mathcal{Q}_{\text{train}}} r_{\ell,p}(h,q,s).\end{aligned}$$

Analogous directions define v_{Think}, v_{Direct}, and residualized directions such as $v_{C,\text{unique}}$ and $v_{\text{Think,unique}}$. We use raw-vector interventions $I_{\beta,v}(r) = r + \beta v$ and unit-coordinate interventions

$$I_{\alpha,v}(r) = r + \alpha\|r\|_2 \frac{v}{\|v\|_2}.$$

For a metric m, the causal effect of direction v at $(\ell, p^\star)$ is

$$\delta^\pm_m(v;\alpha) = m(I_{\pm\alpha,v}(r_{\ell,p^\star})) - m(r_{\ell,p^\star}).$$

The unit-coordinate form normalizes intervention size at the patch site. When splits are available, coordinate construction and evaluation use disjoint prompt sets. At 14B, coordinate selection, position mapping, and layer/alpha analysis are evaluated separately. At 32B, the L63 coordinate is fixed before random-direction comparison.

Controls. Each causal claim uses add and subtract interventions at the same layer and position. Random controls use norm-matched directions at the same layer, position, history condition, and scale. Let $Q^+_{.95}$ and $Q^-_{.05}$ denote the matched random-direction percentiles. Direction v is bidirectional and direction-specific when

$$\delta^+_{\text{RMD}}(v;\alpha) > Q^+_{.95}, \qquad \delta^-_{\text{RMD}}(v;\alpha) < Q^-_{.05}.$$

Additional controls include non-answer positions, delayed boundaries, random-direction planes, literal-token variants, and message-format variants.

Statistical reporting. Matched behavioral and logit effects are paired item-level differences with paired SEs. Causal-coordinate tables report target add/sub effects and matched random p05/p95 bands. Cross-scale comparisons use bidirectionality, answer-boundary localization, random-direction separation, and transfer to independent prompts or delayed boundaries.

We treat these criteria jointly. A probe score without a causal shift establishes decodability, not control. A causal shift at a non-answer site establishes sensitivity, not boundary readout. A target direction inside the matched random band establishes perturbability, not specificity. The main claims require the criteria to hold together in the same experimental family.

3 Results

3.1 Matched Instructions Still Produce Different Latent States

Table 1 reports matched-pair visible and first-token effects. Raw reasoning-marker deltas are 0.000 in every displayed row, while latent readouts move strongly. In the 14B matched-final condition, prior think and prior no-think have zero visible-character deltas, with $\Delta \log P_R = +1.206$ and $+1.789$. Alternating histories produce $+0.047$ visible characters and $+3.705$ in $\Delta \log P_R$. This gives a positive visible-latent gap under the same final no-think instruction.

The same table gives the central asymmetry. In 14B, prior no-think produces stronger RMD than prior think under the same current no-think instruction ($+1.417$ vs. $+0.941$). Panel B tests the asymmetry in a 4B dose design. With two, four, and eight repeated controls, no-think history raises $\log P_R$ by $+0.426$, $+0.766$, and $+0.603$. The non-monotone RMD values motivate separate reporting of P_R and P_D.

Table 2 reports the same separation in hidden-state probes. The anchor rows place current no-think near zero and current think near one. Under a matched final no-think block, L40 scores are $+0.046$ for prior think, $+0.066$ for prior no-think, and $+0.183$ for alternating history, with neutral near zero.

The hidden-state result narrows the source of the effect. The separation is present before any sampled continuation and before the model commits to a visible rationale. It is also graded: alternating histories give the largest late-layer score, matching the stronger first-token pressure in Table 1.

3.2 Readout Contracts Gate Whether the State Becomes Visible

Table 3 tests how answer contracts gate the latent state. Under the soft-direct contract, aggregate visible characters increase by $+4.664$, visible derivation probability by $+0.398$, and RMD by $+2.210$. The alternating-minus-neutral contrast produces $+2.695$ visible characters and $+2.023$ RMD, consistent with visible release under a softer contract.

The answer-field contract gives a complementary dissociation. Aggregate $\Delta\mathrm{RMD}$ is $+9.603$, while visible release is -0.273 characters with no derivation-probability increase. Under this contract, prior no-think versus prior think decreases visible characters by -1.414 and increases RMD by $+1.998$. Bare no-think gives little visible movement with positive latent RMD. The pattern separates boundary-state strength from visible rendering.

These contracts also separate two interpretations of no-think compliance. A model can comply at the surface by avoiding derivation text, while its answer-boundary distribution still shifts toward reasoning-like first tokens. The observed dissociation makes visible chain length an incomplete proxy for the control state.

3.3 The State Is an Answer-Boundary Register

Table 4 separates diagnostic and causal localization. In 32B L63, projection onto the learned history direction is 468.629 at the immediate answer boundary and 408.245 at the delayed boundary. Question-end projections are -9.485 and -7.803. Add-v_C changes RMD by $+0.191$ at immediate answer start and $+0.211$ at delayed answer start, with non-answer positions at 0.000.

Delayed-boundary prompts rule out a fixed absolute position. The projection peak and causal effect move with the response boundary. Random answer-start effects are $+0.004$ and $+0.010$, far below the target boundary effects. The localization claim combines answer-start concentration, delayed-boundary relocation, and near-zero non-answer causal response.

The boundary pattern appears below 32B. In 14B L39, answer-start projection is 404.547, compared with 30.460 at question end, 15.180 at pre-answer-minus-8, and 11.011 at the mode tag. Delayed 14B prompts move the causal response to

Panel A. Paired history effects under fixed final contracts						
Final contract	Contrast	Δ chars	Δ markers	$\Delta \log P_R$	$\Delta \log P_D$	ΔRMD
14B matched-final no-think						
matched no-think	prior think - neutral	0.000 (.000)	0.000	$+1.206$ (.063)	$+0.265$ (.082)	$+0.941$ (.104)
matched no-think	prior no-think - neutral	0.000 (.000)	0.000	$+1.789$ (.061)	$+0.372$ (.131)	$+1.417$ (.139)
matched no-think	alternating - neutral	$+0.047$ (.027)	0.000	$+3.705$ (.150)	$+0.306$ (.201)	$+3.399$ (.158)
14B answer-contract test						
bare no-think	prior think - neutral	$+0.313$ (.182)	0.000	$+1.849$ (.125)	$+0.517$ (.101)	$+1.332$ (.155)
bare no-think	prior no-think - neutral	$+0.313$ (.182)	0.000	$+0.225$ (.251)	-1.716 (.154)	$+1.941$ (.287)
soft direct	alternating - neutral	$+2.750$ (.710)	0.000	$+0.562$ (.367)	-1.133 (.455)	$+1.695$ (.338)
32B answer-contract test						
strict direct	prior think - neutral	$+0.313$ (.313)	0.000	$+0.996$ (.036)	-0.186 (.041)	$+1.182$ (.037)
answer field	prior no-think - neutral	$+0.396$ (.411)	0.000	-0.229 (.058)	-1.541 (.073)	$+1.312$ (.068)
concise derivation	prior no-think - neutral	$+0.146$ (.146)	0.000	$+0.974$ (.079)	-1.107 (.060)	$+2.081$ (.082)
Panel B. 4B repetition-dose contrast under matched final no-think						
Dose	Contrast	$\Delta \log P_R$	SE	ΔRMD	SE	Items
1x	repeated no-think - think	-0.096	.021	+0.102	.060	96
2x	repeated no-think - think	+0.426	.036	+0.569	.089	96
4x	repeated no-think - think	+0.766	.051	+0.174	.105	96
8x	repeated no-think - think	+0.603	.054	+0.008	.094	96

Table 1: History effects under matched final instructions. Panel A fixes the task and final answer contract while varying earlier control history. Panel B varies the repetition dose of no-think history against think history. Neutral denotes the no-prior-control history. Parentheses give paired SEs.

14B condition	L10 score	SE	L20 score	SE	L30 score	SE	L40 score	SE
matched-final neutral	+0.031	.038	-0.046	.006	-0.003	.039	-0.001	.002
matched-final prior think	+0.213	.039	-0.060	.007	+0.007	.040	+0.046	.002
matched-final prior no-think	+0.168	.040	-0.093	.006	-0.019	.038	+0.066	.002
matched-final alternating	+0.263	.045	+0.195	.005	+0.245	.041	+0.183	.005
probe current no-think	+0.000	.040	+0.000	.008	+0.000	.038	+0.000	.002
probe current think	+1.000	.045	+1.000	.005	+1.000	.014	+1.000	.005
reset after think	-0.268	.050	-0.006	.006	-0.026	.037	+0.127	.003

Table 2: Hidden-state mode-probe scores under matched final conditions. Scores are think-direction projections at selected 14B layers.

the delayed answer start. In 8B L35, answer-boundary $\Delta\mathrm{RMD} = +2.438$ with max non-boundary 0.000. In 4B L35, answer-boundary add/sub gives $+1.089/-1.473$ with max non-boundary absolute effect 0.000.

The mode tag can contain projection mass, yet causal response concentrates at answer start. Interventions at the final control tag, question marker, and pre-answer offsets do not produce comparable shifts.

This distinction is central for mechanism. A prompt-history trace can be decodable at several locations, but the causal readout is the site that changes the next-token distribution. The answer boundary is the only tested site that satisfies projection, relocation, and intervention criteria together.

3.4 The Register Is Causal, Bidirectional, and Random-Specific

Table 5 tests causal control. In 32B, an L63 $v_{C,\mathrm{unique}}$ coordinate shifts RMD by $+7.033$ under add and -4.007 under subtract. The add effect exceeds random p95 $+6.260$, and the subtract effect passes random p05 -3.365, giving positive causal-specificity margins.

The matched-random margins are $+0.773$ for add and $+0.642$ for subtract. These margins are smaller than the raw effects, reflecting broad high-gain directions in late residual streams. The learned coordinate still preserves sign and exceeds the matched null.

The 14B unit-coordinate analysis gives the same structure. Across the coordinate test, $16/16$ tested coordinates are bidirectional and direction-specific. The strongest L37 $v_{\mathrm{Think,unique}}$ coordinate at $\alpha = .45$ gives $+3.098/-5.210$ in RMD; $v_{\mathrm{Think}} - v_C$ gives $+4.585/-3.519$. In the 14B position map, only $1/8$ vector-position cells are bidirectional and direction-specific, and the selected cell is the answer boundary. Its span is 7.437, compared with 0.023 at the question marker, 0.010 at the final no-think token, and -0.003 at prompt middle.

The 8B and 4B models give smaller direction-specific effects. In 8B L35, add/sub gives $+0.724/-0.814$, outside random p95/p05 $+0.419/-0.427$. In 4B L35, add/sub gives $+1.239/-1.502$, outside $+0.486/-0.694$. Learned history coordinates steer response-boundary RMD beyond matched random perturbations.

Final contract	Agg. Δ chars	Agg. Δ deriv.	Agg. ΔRMD	Matched contrast	Δ chars	Δ deriv.	ΔRMD
soft direct	+4.664	+0.398	+2.210	alternating - neutral	+2.695	+0.211	+2.023
answer field	-0.273	+0.000	+9.603	prior no-think - prior think	-1.414	-0.109	+1.998
bare no-think	+0.180	-0.008	+3.016	prior no-think - neutral	+0.273	+0.008	+1.670

Table 3: Readout-contract effects in 14B. Columns summarize visible release and latent pressure under matched contrasts.

Panel A. Projection localization.

Model	Layer	Boundary	Answer	SE	Mode	Question	Pre-answer	Boundary ratio
32B	L63	immediate	468.629	4.620	111.418	-9.485	-6.142	11.07×
32B	L63	delayed	408.245	2.896	109.072	-7.803	21.388	8.86×
14B	L39	immediate	404.547	11.345	11.011	30.460	15.180	21.42×
14B	L39	delayed	356.896	5.948	12.862	30.227	18.358	17.42×
14B	late	pooled	337.090	–		–		13.66×

Panel B. Causal localization.

Model	Layer	Site	Answer ΔRMD	Answer $\Delta \log P_R$	Random-site ΔRMD	Max non-answer ΔRMD
32B	L63	immediate add v_C	+0.191	-0.105	+0.004	0.000
32B	L63	delayed add v_C	+0.211	-0.118	+0.010	0.000
14B	L39	delayed add v_C	+0.122	+0.293	+0.020	0.000
8B	L35	add $v_{C,\text{unique}}$, $\beta = .3$	+2.438	–	–	0.000
4B	L35	add/sub $v_{C,\text{unique}}$, $\beta = 2.0$	+1.089 / -1.473	–	–	0.000

Table 4: Answer-boundary localization. Panel A reports mean projections onto the learned history direction at answer and non-answer positions. Panel B reports causal effects at answer and non-answer sites.

The smaller models also provide a useful stress test. Their effects are more sensitive to layer choice and vector construction, yet the successful coordinates preserve the same sign pattern and boundary preference. The mechanism is therefore not an artifact of a single scale or a single intervention parameterization.

3.5 Scale and Dose Structure

Scale changes how cleanly the mechanism factorizes. In 4B and 8B, the state is boundary-local and causally steerable, with stronger sensitivity to layer, history contrast, and intervention form. In 14B, local-norm unit coordinates reveal stable late-layer add/sub surfaces. In 32B, the mechanism concentrates into a late L63 register that transfers to independent examples.

Layer scans show the progression. In 32B, mean answer projection rises from 50.20 at L47 to 446.14 at L63. In 14B, it rises from 22.70 at L28 to 325.04 at L39. Logit-lens dose curves are monotone: for a long no-think history in 32B, add/sub effects grow from $+0.116/-0.121$ at $\beta = .5$ to $+0.587/-0.587$ at $\beta = 3.0$. In 8B, $\beta = 2.0$ gives $+0.790/-0.901$ for short no-think history and $+0.632/-0.712$ for long no-think history.

The 14B unit-coordinate dose analysis matches this pattern. At L37, add-sub span grows from 4.50 at $\alpha = .25$ to 9.00 at $\alpha = .65$. At L39, it grows from 3.20 to 8.74.

Dose response links the causal coordinate to a continuous latent direction. Increasing intervention strength produces smooth changes rather than a threshold jump, and the same direction remains interpretable under raw-vector and unit-coordinate parameterizations.

3.6 Artifact Controls and Reset Evidence

Artifact controls target surface-token, formatting, position, and generic-perturbation explanations. Literal-token variants attenuate the 4B answer projection: operative repeated no-think gives 120.660, literal quotation 53.387, literal spelling 22.670, and neutral history 0.000. The effect persists across 8B message formats, with answer/non-answer projections $318.629/144.910$, $292.070/141.875$, and $260.539/185.052$. Position placebos isolate the read site: in 14B, the RMD span is 7.437 at the answer boundary, 0.023 at the question marker, 0.010 at the final no-think control, and -0.003 at prompt middle. In 32B, learned add/sub $+7.033/-4.007$ exceeds the random p95/p05 envelope $+6.260/-3.365$.

Reset-token dose adds a control handle. In 4B, projection falls from 118.412 with no reset to 50.532 after one no-think reset, 27.432 after two, and 23.438 after five. Together, the controls support a mechanism in which prior mode history writes a latent coordinate, the answer boundary reads it, and the final output contract gates rendering.

The reset result also shows that the residue is not

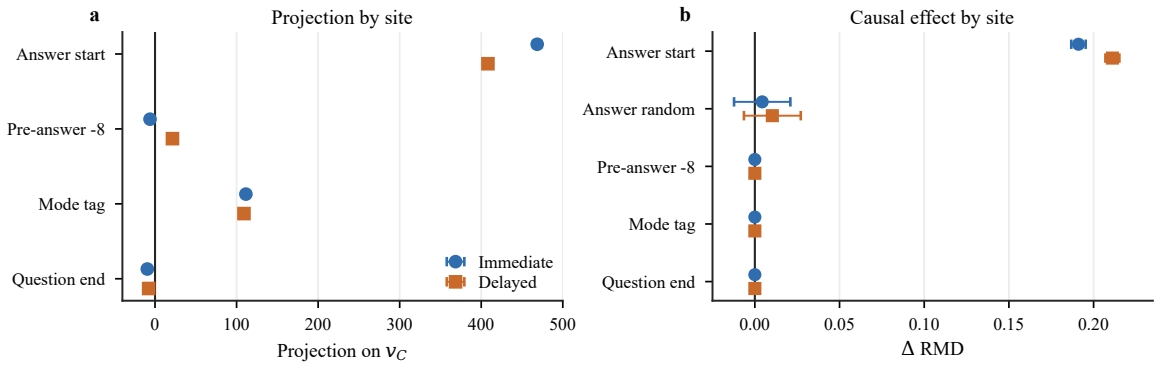

Figure 1: Answer-boundary localization in Qwen3-32B. Projection and causal effects concentrate at answer start and follow delayed response boundaries.

		Add			Subtract		
Model	Coordinate	ΔRMD	p95	Margin	ΔRMD	p05	Margin
32B L63	$v_{C,\text{unique}}$, $\alpha = .90$	+7.033	+6.260	+0.773	-4.007	-3.365	+0.642
14B L37	$v_{\text{Think},\text{unique}}$, $\alpha = .45$	+3.098	+0.471	+2.627	-5.210	-2.220	+2.990
14B L37	$v_{\text{Think}} - v_C$, $\alpha = .45$	+4.585	+0.471	+4.114	-3.519	-2.220	+1.299
14B L39	$v_{\text{Think}} - v_C$, $\alpha = .45$	+3.984	+1.942	+2.042	-2.935	-1.229	+1.706
8B L35	$v_{C,\text{unique}}$, $\beta = 2.0$	+0.724	+0.419	+0.305	-0.814	-0.427	+0.387
4B L35	$v_{C,\text{unique}}$, $\beta = 2.0$	+1.239	+0.486	+0.753	-1.502	-0.694	+0.808

Table 5: Bidirectional causal coordinates across model scales. Margins are relative to norm-matched random directions.

permanent prompt memory. Additional controls can attenuate it without changing the original history. This gives the state an operational reset axis, complementing the add/sub steering axis used in the causal-coordinate experiments.

4 Mechanistic Account

The results support a compact control-state model. Let

$$z_\ell(h, q, s) = u_\ell^\top r_{\ell,p^\star}(h, q, s)$$

denote the projection of the answer-boundary residual stream onto a learned control direction u_ℓ. A stateless latest-instruction model imposes $z_\ell(h, q, s) = z_\ell(h', q, s)$ whenever q and s match. The observed matched-final contrasts require a history term.

$$z_\ell(h, q, s) = \mu_\ell(q, s) + \eta_{\ell,s}(h) + \epsilon_\ell.$$

Here $\mu_\ell(q, s)$ is the current task-and-contract component, and $\eta_{\ell,s}(h)$ is the history-written residue. Matched-final experiments show $\eta_{\ell,s}(h) \neq 0$ under a current no-think instruction. Boundary experiments locate the readout at $p^\star$. Answer-contract experiments show that visible derivation is gated relative to z_ℓ. A local approximation is

$$\begin{aligned} \mathrm{RMD}(h, q, s) &\approx \lambda_s z_\ell(h, q, s) + b_s(q), \\ \Pr(D_{\text{vis}} = 1 \mid h, q, s) &\approx \sigma(\gamma_s z_\ell(h, q, s) - \tau_s), \end{aligned}$$

where the final answer contract sets $(\lambda_s, \gamma_s, \tau_s)$. Strict contracts can keep z_ℓ latent, answer-field contracts can increase RMD with little visible release, and softer contracts can render derivation.

The model also separates storage from readout. The history term $\eta_{\ell,s}(h)$ can be decodable away from the answer boundary, as seen in mode-tag projections, yet causal influence appears when the assistant prepares the first answer token. This accounts for the combination of prompt-history decodability, delayed-boundary relocation, and local intervention specificity.

This model gives a mechanistic White Bear account. A no-think control can preserve a coordinate for withheld reasoning while changing the gate that renders it. This explains why prior no-think and alternating histories can raise latent reasoning pressure more than prior think histories in some matched-final regimes. The variable is boundary-local and controls emission of reasoning-like content.

Five neighboring accounts are inconsistent with the full evidence. Latest-instruction overwrite predicts identical answer-boundary distributions for matched final prompts, contradicted by Tables 1 and 2. Visible leakage predicts alignment with reasoning markers or length, contradicted by Tables 1 and 3. Literal-token copying predicts similar effects under quoted or spelled controls, attenuated in Section 3.6. Generic residual perturbation predicts comparable matched random directions, separated in Table 5. Absolute-position readout predicts failure under delayed answers, contradicted by Table 4.

Panel A. Layer scan						
Model	Quantity	Point 1	Point 2	Point 3	Point 4	Ratio
32B	answer projection	L47: 50.20	L51: 91.86	L55: 184.93	L63: 446.14	8.89×
14B	answer projection	L28: 22.70	L32: 48.75	L37: 160.63	L39: 325.04	14.32×
Panel B. Logit-lens dose						
Condition	Direction	$\beta = .5$	$\beta = 1$	$\beta = 2$	$\beta = 3$	Span
32B long no-think history	add ΔRMD	+0.116	+0.185	+0.396	+0.587	+0.471
32B long no-think history	subtract ΔRMD	-0.121	-0.198	-0.398	-0.587	-0.466
8B short no-think history	add/subtract ΔRMD	–	+0.424/-0.447	+0.790/-0.901	–	+0.366/-0.454
8B long no-think history	add/subtract ΔRMD	–	+0.334/-0.346	+0.632/-0.712	–	+0.298/-0.366
Panel C. Unit-coordinate dose						
Layer	Quantity	$\alpha = .25$	$\alpha = .35$	$\alpha = .45$	$\alpha = .65$	Span
14B L37	add–sub span	4.50	6.02	7.28	9.00	+4.50
14B L39	add–sub span	3.20	4.58	5.97	8.74	+5.54

Table 6: Layer and dose structure of the answer-boundary coordinate.

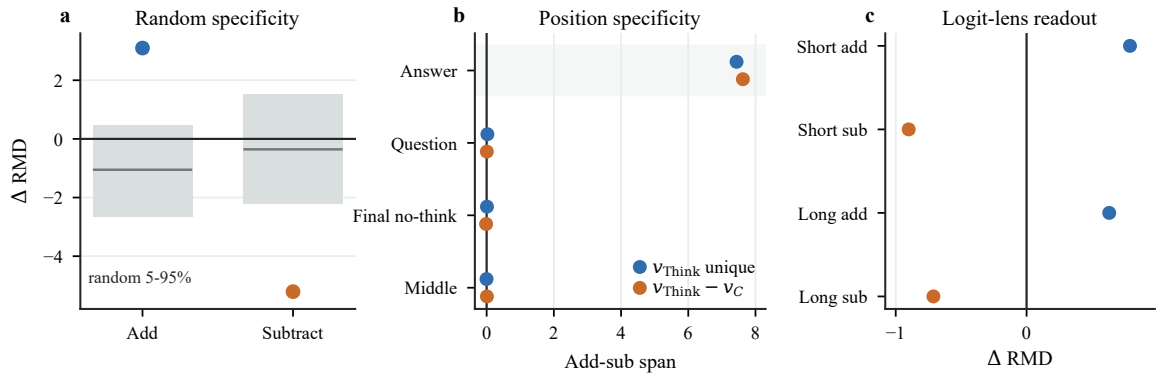

Figure 2: Causal controls. Learned coordinates exceed matched random directions, act at the answer boundary, and shift logit-lens pressure before generation.

The causal claim rests on a conjunction of criteria. The coordinate is bidirectional under add/sub, follows semantic answer boundaries, exceeds norm-matched random directions, transfers to independent prompts, and responds locally to dose. In matched-final reasoning-control tasks, prior mode history contributes a low-dimensional term $\eta_{\ell,s}(h)$ to a late-layer answer-boundary state, and interventions along that coordinate steer the next-token distribution.

5 Related Work

Hybrid thinking and visible rationales. Hybrid-thinking models expose inference-time controls for thinking and non-thinking regimes (Yang et al., 2025). Prior work studies reasoning-like leakage from no-thinking mode (Wang et al., 2025) and the utility or faithfulness of visible rationales (Wei et al., 2022; Turpin et al., 2023; Nye et al., 2021; Kojima et al., 2022; Wang et al., 2023; Yao et al., 2023; Lanham et al., 2023). Our contrast fixes the final no-think instruction and task, then varies only prior mode-control history.

Negated control and activation interventions. The White Bear effect describes ironic availability under suppression (Wegner et al., 1987). Recent transformer work studies related negation effects for lexical instructions (Mann et al., 2025). We test the same principle for reasoning-control state. Mechanistically, the analysis builds on causal tracing and model editing (Meng et al., 2022; Vig et al., 2020; Elhage et al., 2021), representation engineering and inference-time interventions (Zou et al., 2023; Li et al., 2023; Turner et al., 2023), latent-reasoning resources (Zhang and Viteri, 2024; Deng et al., 2026; Burns et al., 2023; Belrose et al., 2023), and task/function vectors (Hendel et al., 2023; Todd et al., 2024). The coordinate here is defined by prior control history under the same current instruction, with causal readout localized to the answer boundary.

6 Conclusion

Under identical final no-think instructions, prior histories leave a readable, steerable answer-boundary state. Learned coordinates shift reasoning-minus-direct pressure bidirectionally, while random directions and non-answer positions remain weak. This gives a mechanistic White Bear effect in which no-think suppresses overt thought while preserving a boundary state.

Limitations

The experiments use Qwen3-family hybrid-thinking controls, so the evidence is strongest for model families with explicit reasoning-mode instructions. The primary observable is first-token reasoning-minus-direct pressure. Visible generations are readout behavior in this design. The coordinate is local to the tested tasks, layers, positions, and answer contracts, with strongest effects at late-layer answer-boundary states.

References

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*.

Boyi Deng, Xu Wang, Yaoning Wang, Yu Wan, Yubo Ma, Baosong Yang, Haoran Wei, Jialong Tang, Huan Lin, Ruize Gao, Tianhao Li, Qian Cao, Xuancheng Ren, Xiaodong Deng, An Yang, Fei Huang, Dayiheng Liu, and Jingren Zhou. 2026. Qwen-Scope: Turning sparse features into development tools for large language models. *arXiv preprint arXiv:2605.11887*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

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

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, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*.

Logan Mann, Nayan Saxena, Sarah Tandon, Chenhao Sun, Savar Toteja, and Kevin Zhu. 2025. Don't think of the white bear: Ironic negation in transformer models under cognitive load. *arXiv preprint arXiv:2511.12381*.

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

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. *arXiv preprint arXiv:2112.00114*.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *International Conference on Learning Representations*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

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*.

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*.

Shouren Wang, Wang Yang, Xianxuan Long, Qifan Wang, Vipin Chaudhary, and Xiaotian Han. 2025. Demystifying hybrid thinking: Can LLMs truly switch between think and no-think? *arXiv preprint arXiv:2510.12680*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. 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*.

Daniel M. Wegner, David J. Schneider, Samuel R. Carter, and Teri L. White. 1987. Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology*, 53(1):5–13.

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

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. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*.

Jason Zhang and Scott Viteri. 2024. Uncovering latent chain of thought vectors in language models. *arXiv preprint arXiv:2409.14026*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Experimental Details

Prompt histories. The neutral history contains no earlier thinking-control instruction before the final instruction block. The prior-think and prior-no-think histories insert a previous mode-control context before the same final instruction block. Alternating histories interleave think and no-think controls before the final instruction block. Repeated no-think histories construct higher-amplitude contrast directions. Matched comparisons keep the final instruction block, task, and answer contract fixed within each row.

Answer contracts. Final answer contracts are grouped by readout behavior. Strict direct-answer contracts suppress visible derivation. Clamp contracts preserve the final no-think instruction with minimal additional formatting. Fielded-answer contracts place the answer in a designated field. Softer direct-answer contracts permit brief intermediate text before the final answer. These contracts test whether latent reasoning pressure and visible derivation can be dissociated.

Position definitions. The answer boundary is the residual-stream position immediately before the model's first answer token. Non-answer controls include the final control token or mode-tag end, question marker or question end, prompt middle, prompt start, and pre-answer offsets. Delayed-boundary prompts insert neutral material before the answer slot so that the answer boundary moves. These prompts test whether the effect follows the semantic response boundary rather than an absolute token index or a fixed offset from no-think.

Vector and intervention details. Mean-difference vectors are constructed from answer-boundary residual activations. Unique directions residualize a contrast vector against comparison axes before normalization. Raw-vector β interventions add a scaled contrast direction directly to the residual stream. Unit-coordinate α interventions scale the unit direction by the local residual norm. Add and subtract interventions are always evaluated as a pair when making a bidirectional causal claim.

Random controls. Random controls use norm-matched random directions at the same layer, position, history condition, and intervention magnitude. We report target effects together with random p05/p95 bands. This baseline accounts for the fact that late residual streams can contain many directions that perturb the next-token distribution. A direction is treated as direction-specific only when add exceeds random p95 and subtract falls below random p05.

Analysis protocol. The main analysis proceeds in a fixed order. First, matched-final generation and first-token readouts test whether histories remain separable under the same current final instruction. Second, hidden-state projections test whether this separation is represented internally. Third, boundary maps locate the strongest projection and intervention response across semantic positions. Fourth, add/sub interventions test causality at the answer boundary. Fifth, matched random directions and position placebos test specificity. Sixth, answer-contract experiments test whether the same latent pressure is rendered or clamped under different final answer contracts. This ordering prevents visible-generation effects from being treated as primary evidence and prevents probe separability from being treated as causal evidence.

Prompt templates. The following examples give the message-content strings before chat-template serialization. Matched-final comparisons keep the final block identical and vary only the prior mode-control history.

FINAL ANSWER CONTRACTS

Strict direct

```
/no_think Return only the final answer.
Question: What is 32+91?
```

Brief answer

```
/no_think Answer the following question briefly.
Question: Compute (5+29)-3.
```

Plain no-think

```
/no_think Question: What is 158-27?
```

Internal-final answer

```
/no_think Think internally, then give only the
final answer. Question: What is 12*12?
```

MATCHED-FINAL HISTORY SEQUENCES

Prior-think history

```
/think Question: What is 83+20?
/no_think Return only the final answer.
Question: What is 49-15?
```

Prior-no-think history

```
/no_think Question: What is 83+20?
/no_think Return only the final answer.
Question: What is 49-15?
```

Alternating history

```
/think Question: What is 83+20?
/no_think Question: What is 49-15?
/think Question: What is 12*12?
/no_think Return only the final answer.
Question: Compute (5+29)-3.
```

CONTROL-TOKEN RESET SEQUENCES

```
/no_think /no_think Return only the final
answer. Question: What is 32+91?
/no_think /no_think /no_think Return only the
final answer. Question: What is 32+91?
/think /no_think Return only the final answer.
Question: What is 32+91?
/no_think /think /no_think Return only the
final answer. Question: What is 32+91?
```

B Supplementary Analyses

This appendix reports additional result panels for the same diagnostic chain used in the main text. The panels group matched-final history effects, answer-boundary localization, causal specificity, scale and dose behavior, and artifact controls under the same estimands and intervention protocols.

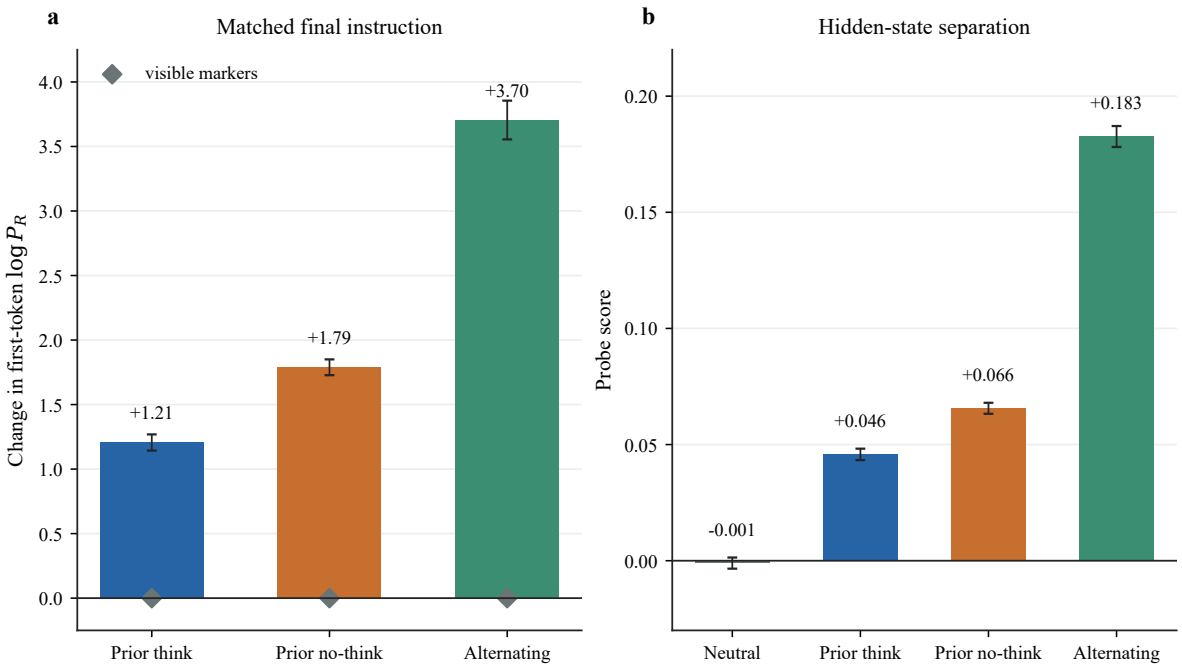

Figure 3: Matched-final latent pressure. Prior no-think and alternating histories raise first-token reasoning pressure under the same final no-think instruction, while visible reasoning markers remain zero.

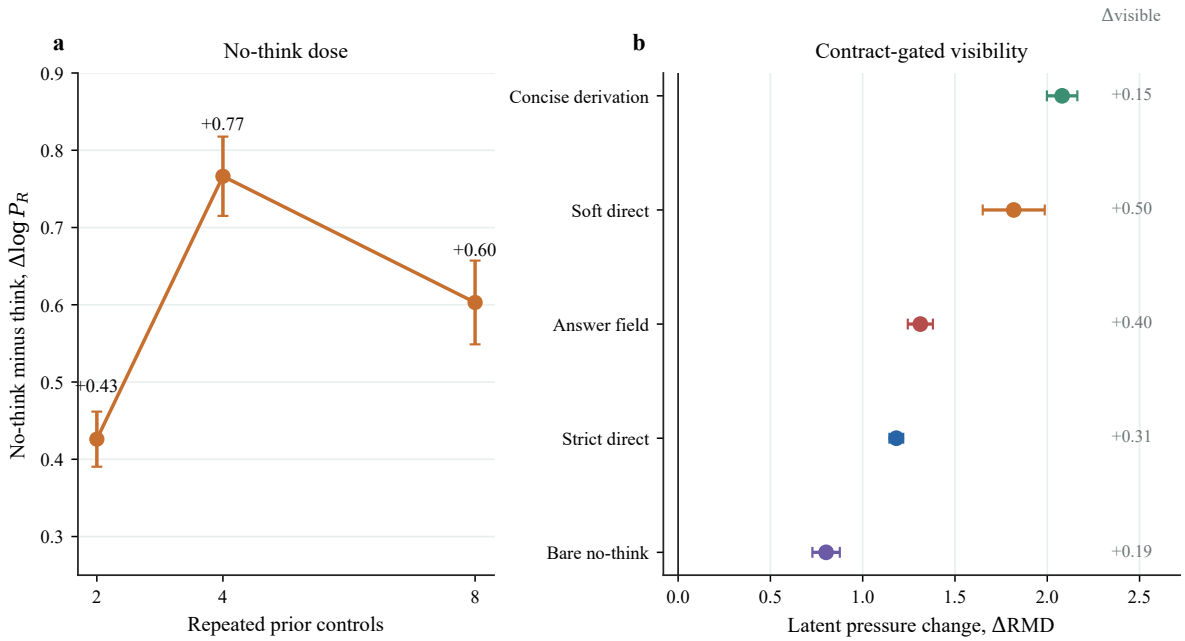

Figure 4: Dose and readout gating. Repeated no-think histories increase reasoning-token pressure, and answer contracts separate visible length from latent pressure.

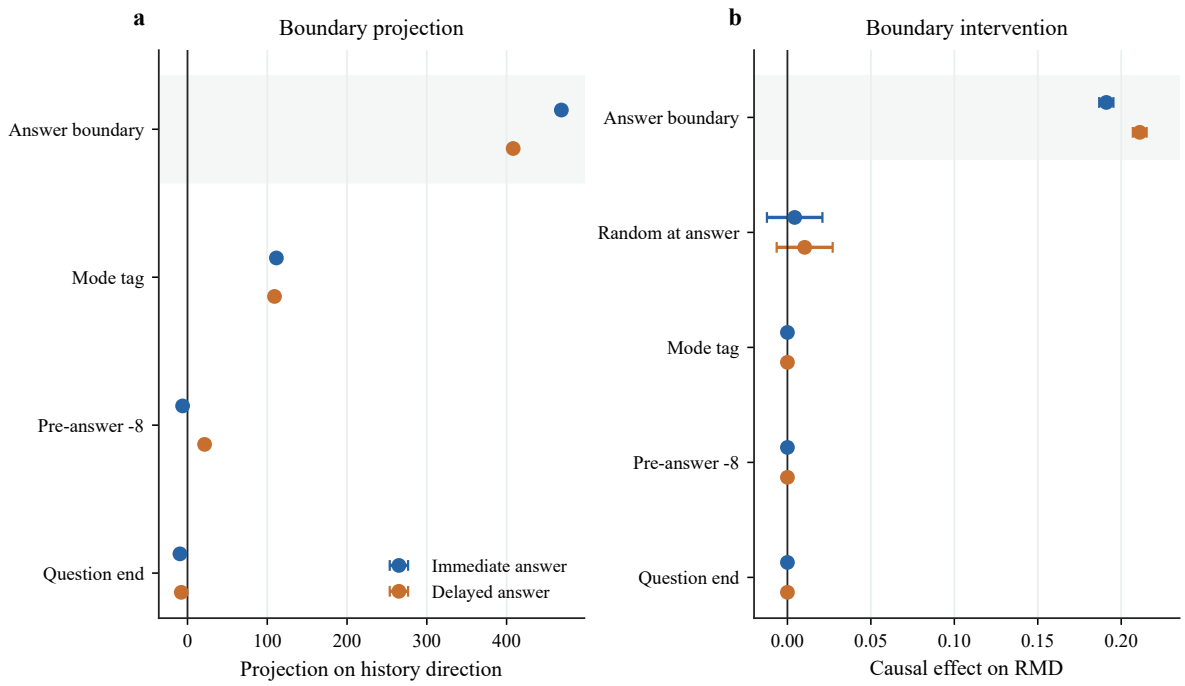

Figure 5: Answer-boundary localization. Projection and causal effects concentrate at the assistant answer boundary and follow delayed answer starts.

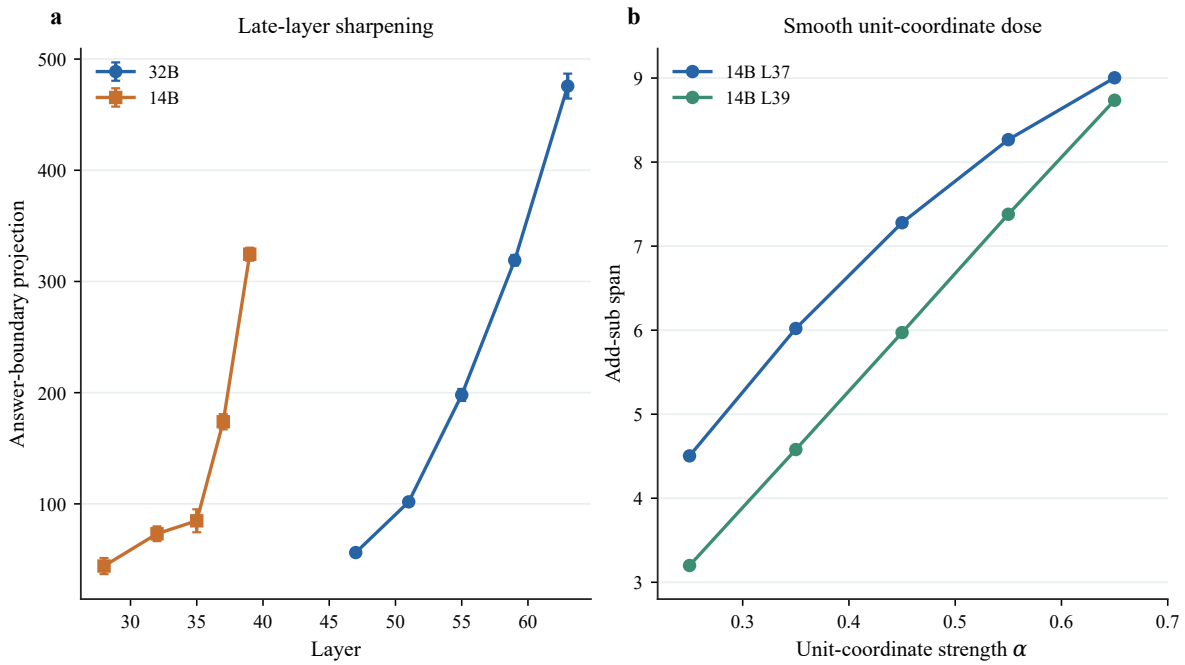

Figure 6: Scale and dose response. The history direction grows in late layers, and unit-coordinate effects vary smoothly with intervention strength.

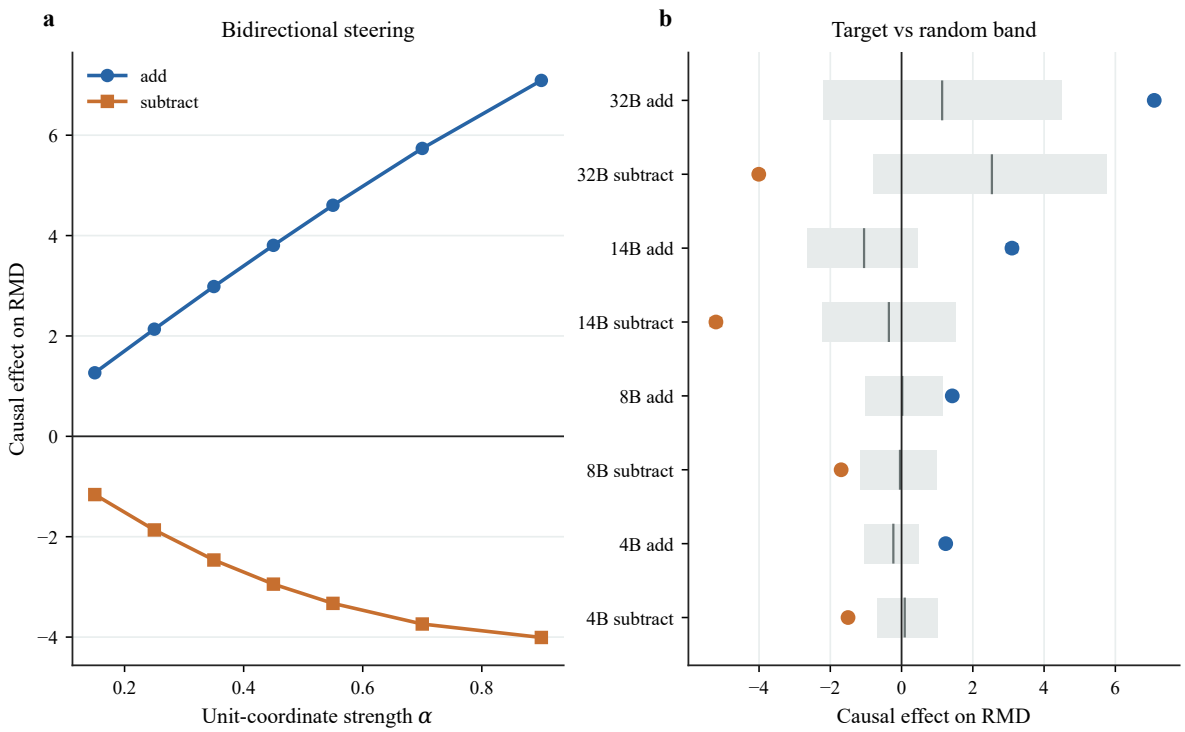

Figure 7: Causal specificity. The learned coordinate steers reasoning-minus-direct pressure bidirectionally and exceeds matched random-direction bands.

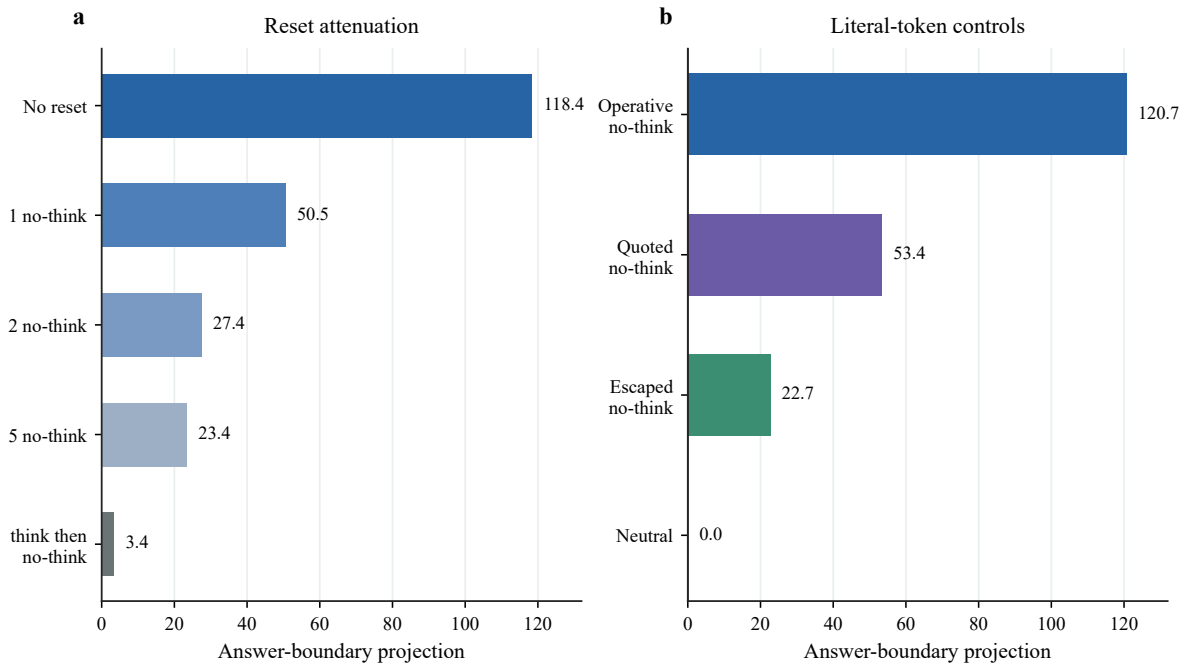

Figure 8: Reset and literal-token controls. Additional no-think reset controls attenuate the residue, while quoted or escaped mentions do not reproduce operative control.