The Termination Circuit: A Verification Gate Decides When Reasoning Models Stop Thinking

Chandram Dutta[1]

[1]Independent Researcher

Abstract

Reasoning models keep thinking for a long time even after finding the answer. We ask what actually makes them stop. In Qwen3-1.7B, the answer to most GSM8K problems the model solves is already settled about halfway through its chain of thought: cut the CoT there and force an answer, and the model answers correctly. When left alone, it keeps reasoning. We trace the stopping decision (emission of the `</think>` token) to a small group of MLP layers near the end of the network, which we call *the termination circuit*. Using causal tests, we show this circuit implements a verification gate: the model stops thinking when it has written out an answer that matches the one it computed internally. Splicing the model's answer statement into the middle of its reasoning fires the stop trigger 94% of the time, while changing the number in that sentence drops it to 8%. Removing the circuit leaves the model unable to stop, though it continues to reason coherently. The gate also has no simple handle: no steering direction can fire it early, which explains why steering-based attempts to shorten reasoning work only modestly and cruder interventions destroy accuracy. Overthinking exists because the termination circuit does not check whether reasoning is sufficient. It only verifies the stated answers, and the model states one at its leisure.

1 Introduction

Reasoning models such as DeepSeek-R1 and Qwen3 produce an explicit thinking segment enclosed by `<think>` and `</think>` tokens before committing to an answer [4, 30]. Prior work estimates that over half of a correct trace can come after the answer is already fixed [9]. This overthinking is well documented [2, 22], and a growing line of work exploits it with early-exit heuristics and length-control methods [21, 24, 32]. This paper asks what, inside the forward pass, concludes that thinking is over and emits `</think>`.

Terminating a chain of thought (CoT) involves two distinct events. *Detection* is semantic: the point where the answer is determined and further reasoning adds nothing. The *trigger* is syntactic: the step where the model actually emits `</think>`. Prior work has measured the gap between them behaviorally, through confidence probes [9], logit margins at sentence boundaries [21], and attention dynamics around the delimiter [11], and has built steering directions that shorten reasoning [24, 26]. However, the mechanism is largely a black box: which components compute the stopping decision, and what do they compute it from?

We study Qwen3-1.7B, an RL-trained thinking model, on the GSM8K dataset [3]. Unlike distilled reasoning models such as the DeepSeek-R1-Distill family [4], which are created by supervised fine-tuning on long reasoning traces from a stronger teacher, Qwen3 was trained with reinforcement learning on verifiable rewards and produces native `<think>`/`</think>` traces [30]. Distilled models imitate the surface form of reasoning and stopping present in the teacher traces, while an RL-trained model has to discover when to stop on its own. Prior work shows that distilled models develop distinct feature directions associated with overthinking [1] and that attention around the thinking delimiters structures information flow in distilled models

[31], suggesting imitation can produce different internal implementations than direct outcome optimization.

We first need a way to tell when the model has found an answer before it says so. We cut the CoT off at some point, add a cue like "the answer is", and see what the model produces. For each trace we record the earliest cut point, S, at which the forced answer is correct. Past S, the remaining reasoning is not needed to get the right answer. One concern is that the easier GSM8K problems may not require the model to reason at all. The forced answer could be correct with no reasoning in context, and S would land spuriously early. To rule this out, we run the same procedure with a mismatched prefix, pairing each question with a different problem's reasoning trace. Forced accuracy with mismatched prefixes stays near zero, while with real prefixes it climbs from a low question-only floor to perfect accuracy as the cut point moves later. Correct forced answers therefore come from the reasoning, not from the cue or the question alone. By this measure the model could stop about halfway through most of its correct traces, and every experiment that follows uses S as its reference point.

Next, we trace which components in the network write the `</think>` logit [5, 18]. The trigger is a late sparse step, not a gradual buildup. A single MLP layer (layer 27 of 28) contributes 40% of the `</think>` logit at stopping, and the top eight components contribute 71%. Before stopping, there is no buildup at all. `</think>` stays far from being emitted through the whole chain, then jumps in the final sentence. We call this small set of late MLPs, along with the earlier MLPs that feed them, the *termination circuit*.

Using causal tests, we find that this circuit computes a verification gate. Splicing the model's own answer statement into the middle of its CoT fires the trigger 94% of the time. Changing the number in that sentence fires it 8% of the time. Placing the statement before sufficiency fires it 16% of the time. Copying the outputs of the two leading MLPs from one context into another [activation patching; 16, 27] transfers the stopping decision, flipping 99% of the traces. Removing the top four MLPs at sentence boundaries (mean ablation) makes the model continue reasoning without ever terminating. However, the CoT stays coherent by perplexity and repetition checks. We find that the gate, not the clock, ends the reasoning.

Finally, we try to operate the gate from the outside. It barely works. No steering direction [25, 33] fires the trigger mid-chain at any strength that keeps the generation intact. The best single direction recovers only 20% of what copying the full internal state achieves. The "ready to stop" signal is not a dial the model turns up but a high-dimensional pattern. Steering can tip the decision right after an answer statement (where the decision is already close), and pushing hard enough even overrides the value check. The gate's verification is a soft preference and not a hard veto. During generation, steering can shorten the CoT somewhat, but a generic direction does it about as well as our circuit-derived one. The direction that predicts how long the model will think before it starts [20] is almost orthogonal to our trigger direction. We find that planning length and deciding to stop are two separate mechanisms.

None of this is specific to one model or one dataset. The same late-MLP gate, at the same relative depth, with the same value-check signature, replicates on Qwen3-8B (5× the parameters), on MATH-500 (genuinely hard problems), and on Magistral-Small-24B, an RL-trained reasoner from a different lab with a different tokenizer (Section 9).

These results explain why overthinking persists and why it resists steering. The stopping mechanism does not check whether reasoning is sufficient. It waits for the model to state an answer, verifies the value, and nothing forces that statement to come early. We conclude that making the model state its answer early triggers the gate and stops the CoT.[1]

[1] Code for all experiments is at https://github.com/Chandram-Dutta/the-termination-circuit.

2 Contributions

- **A causal measure of reasoning sufficiency.** We measure the earliest point S at which a trace's reasoning already determines the answer, using truncation and forced answers rather than probes, with a mismatched-prefix control against leakage. The model could stop about halfway through most of its correct GSM8K traces (Section 5).
- **Localization of the termination trigger.** The `</think>` decision is written by a sparse set of late MLPs (layer 27 alone contributes 40% of the logit) in a single late step, with no buildup after sufficiency (Section 6).
- **The trigger is a verification gate.** Causal tests show the circuit fires when the model has just stated an answer matching the one it computed internally: 94% on the model's own answer statement, 8% with the number changed, 16% before sufficiency. Ablating four MLPs leaves the model unable to stop while reasoning coherently (Section 7).
- **The gate has no one-dimensional handle.** No steering direction fires it mid-chain. The best direction recovers 20% of the full-state patching effect. Its value check is soft and can be overridden and it is orthogonal to the pre-generation length-planning direction of prior work (Section 8).
- **The mechanism generalizes.** The same gate replicates at 5× scale (Qwen3-8B), on harder problems (MATH-500), and across a model-family boundary (Magistral-Small-24B), with a late-MLP trigger at relative depth 0.96–0.97, the same value-check signature, and the same joint necessity (Section 9).

3 Related Work

Overthinking and early exit. It is a common behavior of reasoning models to think past the point of usefulness. Chen et al. [2] studied the phenomenon on o1-like models, and Sui et al. [22] surveyed the efficiency methods it spawned. Huang et al. [9] shows that `</think>` receives high confidence at good stopping points and that over half of a correct trace could be redundant. ThinkBrake monitors the logit margin between `</think>` and the top alternative at sentence boundaries and stops when it is safe [21]. SyncThink reads attention around the thinking delimiter and terminates when a transition signal saturates [11]. Self-Braking Tuning trains the tendency away instead of detecting it at inference [32]. All of this work treats the stopping decision as a signal to be read or a behavior to be trained on. None asks which components compute it, or from what.

Controlling reasoning length. A second line of work steers the reasoning rather than watching it. Activation additions and representation engineering established that a direction added to the residual stream can shift the model behavior [25, 33]. Applied to reasoning models, mean-difference directions shorten or lengthen thinking at a small cost of accuracy [24], a thinking-progress signal can be decoded and overclocked [13], and pre-generation activations encode how long the model plans to think [20]. Venhoff et al. [26] steer behaviors such as backtracking and uncertainty, and explicitly leave the transition to the final answer unaddressed. These are knobs, built and evaluated by their effect on length. Our results explain what the knobs act on. None of these directions operates the termination mechanism itself, and the length-planning direction of Sheng et al. [20] turns out to be orthogonal to it.

Mechanisms inside reasoning models. Closest to us, Zhang et al. [31] study distilled R1 models and finds reasoning-focus attention heads and evidence that `</think>` acts as a segment

marker. Baek and Tegmark [1] find feature directions in distilled models that steer between overthinking and incisive thinking. Both study distilled models, and neither traces the termination decision to components nor tests it causally. A related line edits weights instead of activations: ThinkEdit removes short-reasoning heads to lengthen thinking [23]. Our methods follow the mechanistic-interpretability toolkit. Direct logit attribution [5, 18], activation patching and ablation [16, 27, 28], with the subspace-patching illusion [14] guarded against explicitly. To our knowledge, this paper gives the first component-level, causally verified account of how a reasoning model decides to emit `</think>`.

4 Setup

Model. We study Qwen3-1.7B [30], a 28-layer RL-trained thinking model, in bfloat16. In thinking mode it generates a reasoning segment between `<think>` and `</think>`, then an answer. `</think>` is a single token in the Qwen3 vocabulary so attributing its logit to model components is well defined.[2]

Traces. We run the model on the first 500 problems of the GSM8K test split [3] with the thinking-mode chat template and the sampling settings recommended by Yang et al. [30] (temperature 0.6, top-p 0.95, top-k 20), with a fixed seed and a budget of 4096 new tokens. 441 runs finish their reasoning and produce a parseable answer. 404 of those are correct. All experiments use these 404 correct traces, since sufficiency is vague for a trace that never reaches the right answer. The median thinking length is 1266 tokens.

Notation. We describe positions inside the thinking segment as fractions. 0 is `<think>`, 1 is the `</think>` emission, which we call T_{emit}. Later sections intervene at newline sentence boundaries, the points where the model naturally chooses between continuing and stopping [21].

Measuring sufficiency. For each trace we cut the reasoning at 11 evenly spaced fractions. At each cut we close the thinking segment and append the cue `</think>\n\nThe final answer is`, then decode greedily for at most 32 tokens and compare the produced number to the gold answer. The sufficiency point S is the earliest cut that is itself correct and where at least 80% of the later cuts are also correct. The suffix condition keeps one lucky early guess or one noisy late cut from moving S. Requiring the cut itself to be correct keeps a wrong-early trace from snapping S to zero. The overthinking gap of a trace is $T_{\text{emit}} - S$, in fraction units.

Leakage control. A forced answer could be right because the cue and the question alone are enough with the reasoning contributing nothing. So we repeat the whole procedure with mismatched prefixes in the spirit of control tasks for probes [8]. Each question is paired with a different problem's reasoning, cut at the same fractions. Real-minus-mismatched accuracy at each fraction is the reasoning's true contribution. Section 5 reports both curves. Figure 1 illustrates the cutting-and-cueing procedure for both the real trace and the mismatched-prefix control.

[2] Token id 151668 in the Qwen3 tokenizer, shared by every Qwen3 size.

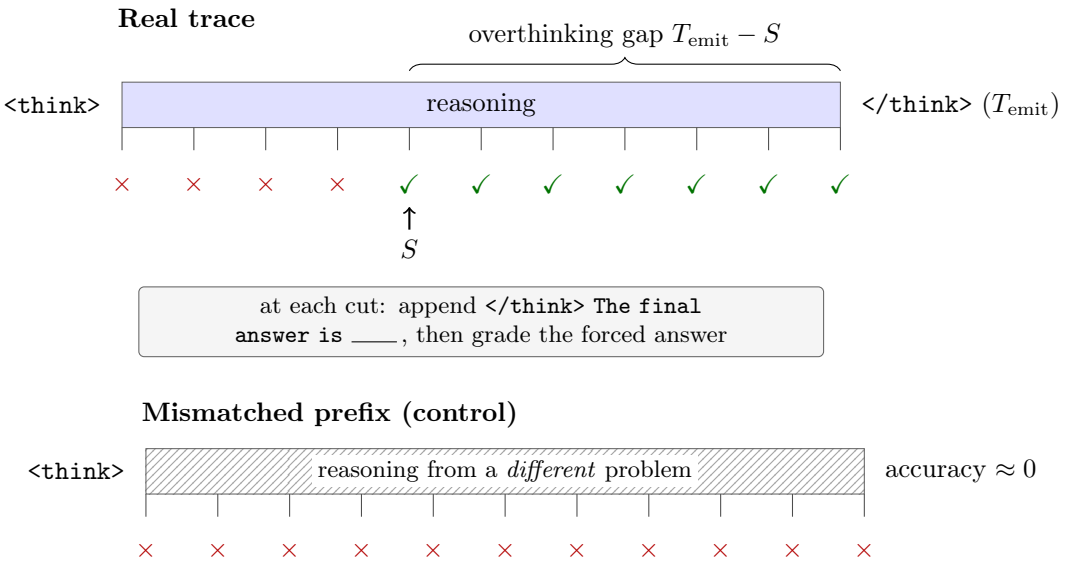

Figure 1: **Measuring the sufficiency point S.** We cut each trace at 11 points, close the thinking segment with an answer cue, and grade the forced answer. S is the earliest correct cut where later cuts stay correct; $T_{\text{emit}} - S$ is the overthinking gap. Bottom: the same procedure with a mismatched reasoning prefix almost never produces a correct answer, so correct forced answers come from the reasoning, not the question or the cue.

Compute. Every experiment runs on a single L40S (48 GB) in bfloat16. Interventions use forward hooks on the HuggingFace `transformers` implementation [29]. Full configuration, seeds, and code are in Appendix A.

5 The Model Can Stop Halfway Through

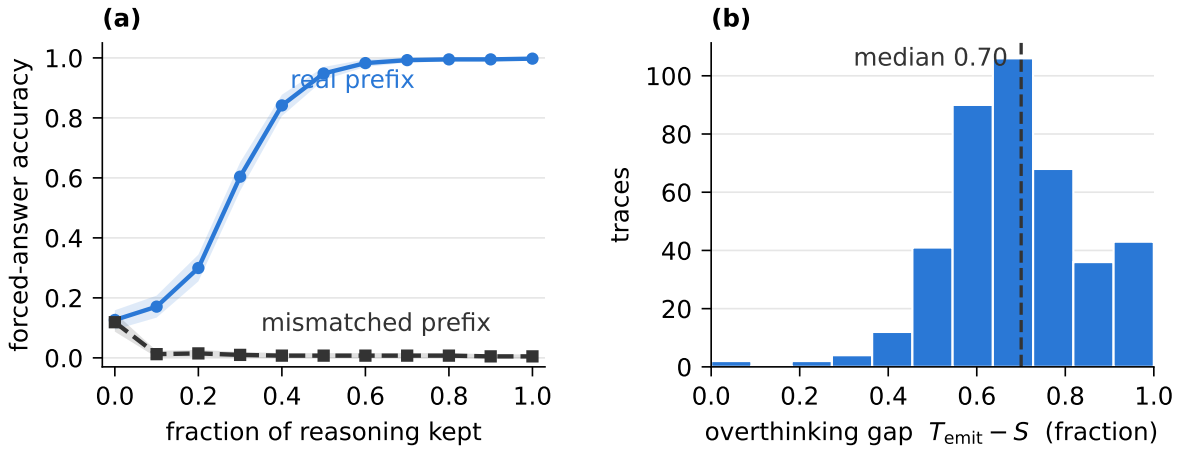

Figure 2: **Detection precedes the trigger.** (a) Forced-answer accuracy against the fraction of reasoning kept ($n = 404$; shading is a 95% CI). Real prefixes climb from the question-only floor of 0.13 to 1.00; mismatched prefixes collapse to 0.01. (b) Distribution of the per-trace overthinking gap $T_{\text{emit}} - S$. The median trace reaches sufficiency after 30% of its thinking and keeps reasoning for the remaining 70%.

Figure 2a shows forced-answer accuracy at each cut. With no reasoning, accuracy is 0.13. With real reasoning it climbs fast, passing 0.5 just before fraction 0.3, reaching 0.95 by fraction 0.5, and 1.00 from fraction 0.8 on. The model can solve nearly every problem from the first half of its reasoning alone.

The mismatched prefixes behave differently in a useful way. At cut zero they match the floor (0.12) as they should, since no reasoning is kept from either problem. The moment any mismatched reasoning enters the context, accuracy collapses to 0.01, well below the floor. Wrong reasoning does not just fail to help; it actively misleads. So the forced answers are read off the reasoning and not guessed from the question. S is measuring what we want.

Per trace, sufficiency comes early (Figure 2b). The median overthinking gap is 0.70. The typical trace reaches S after 30% of its thinking and then reasons for the remaining 70%. Almost no trace has a gap near zero. Overthinking on GSM8K is not an occasional failure. It is the default behavior of every trace, consistent with prior reports of the phenomenon [2]. Nor is it specific to this setting: the gap median is 0.80 for Qwen3-8B and 0.60 on MATH-500 (Section 9).

We also tried a representational version of this measurement: decoding the answer value directly from the residual stream, in the spirit of Zhang et al. [31]. It is real but weak (retrieval 0.51 to 0.70). Appendix B reports it. The behavioral measure is the cleaner signal and the rest of the paper uses it. So the model behaves as if it knows the answer long before it stops. The question becomes what the trigger is waiting for.

6 The Trigger Is a Sparse Late Step

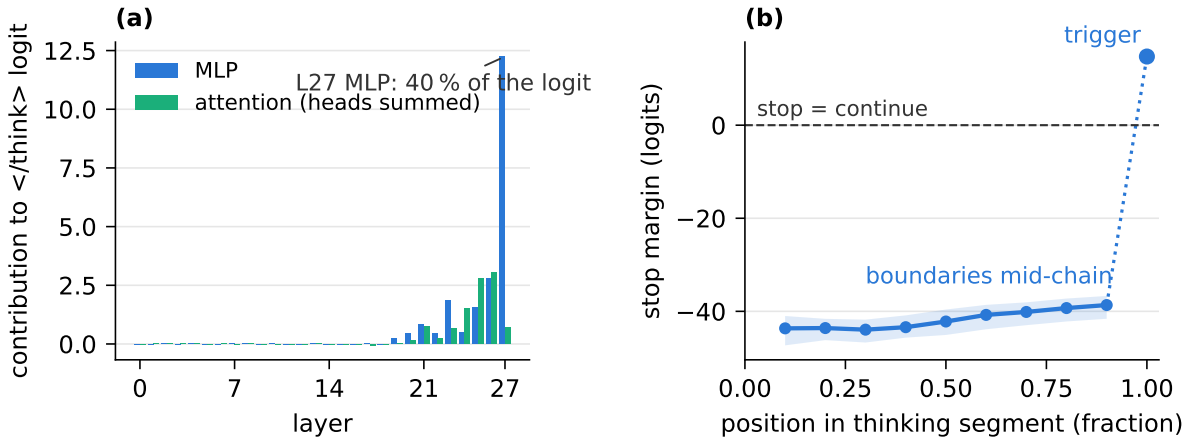

Figure 3: **The trigger is written late, by MLPs, in one step.** (a) Per-layer contribution to the `</think>` logit at the trigger position, averaged over 404 traces. The layer-27 MLP alone writes 40% of the logit; attention heads are small everywhere. (b) The stop margin (`</think>` logit minus the top alternative) at sentence boundaries across the chain (median and IQR), and at the trigger. The margin is flat and deeply negative through the entire chain, including well past sufficiency, then jumps ∼58 logits in a single step.

We now decompose the `</think>` logit at the trigger position into per-component contributions. The final residual stream is a sum of every component's output, and once the RMSNorm scale is frozen at its true per-position value, the map to logits is linear [5, 18]. Each component's contribution is its output dotted with the mean-centered `</think>` unembedding, divided by the frozen scale (Appendix C). The decomposition is exact and we verify it. Summed contributions reconstruct the model's actual centered logit with a maximum error of 0.10 against a total of 30.5. Alongside the logit we track the stop margin, the `</think>` logit minus the top alternative,

which is the decision the model actually faces at each sentence boundary, the same quantity Song et al. [21] thresholds for early exit.

Comparisons need matched positions. `</think>` always follows a structural boundary, so comparing the trigger against random mid-word tokens would inflate every difference. All contrasts therefore run against newline boundary tokens from the same traces with two stricter controls: boundaries right after an interim numeric result (the model states a number and keeps going), and such boundaries late in the chain (fraction $\geq$0.7), which also match absolute position. A "termination component" that is really a newline or answer-statement detector dies in these controls.

Three findings, all visible in Figure 3. **The circuit is sparse and MLP-written.** The layer-27 MLP contributes +12.25 logits of the 30.5 total (40%), which is 4.4 times the next component. The top eight components cover 71% and four of the top five are MLPs. No attention head carries more than 4%. **It is stereotyped.** Across 404 different problems, L27's contribution has an interquartile range of 12.12 to 12.37. The same switch fires with the same force every time. Ranking components on even-numbered traces and on odd-numbered traces selects the identical top eight in the identical order. **It is a step, not a ramp.** Every top component contributes roughly zero at every boundary before the trigger, including every boundary after sufficiency, and the full +12 appears only at the trigger itself with a small precursor ($\sim +1$) inside the final-answer sentence. The stop margin stays between -44 and -39 from fraction 0.1 to the last mid-chain boundary and $p($`</think>`$)$ rounds to zero even at fraction 0.9 [9] and at answer-like boundaries.

We also looked for the opposite mechanism: a component holding `</think>` down mid-chain, which would make the lag a suppression story. There is none. The most negative contribution against the matched controls is -0.21. The few heads that do contribute attend almost entirely to the prompt-start attention sink at the trigger, which is consistent with the attention-sink framing of Li et al. [11], but they carry little of the logit. The evidence for stopping is not accumulated and not suppressed. It simply is not written until the model has chosen to state its final answer. This reframes our question. What is it about the answer statement that flips the switch? Section 7 answers it causally.

7 The Trigger Is a Verification Gate

Everything in Section 6 is correlational. Attribution says the L27 MLP writes the `</think>` logit. It does not say the model needs that component, and it does not say what the component is responding to. This section answers both questions with four causal interventions [6, 16, 27]. We ablate the components and see whether the decision dies. We patch the trigger state into the middle of the chain and see whether the decision transfers. We splice answer-shaped text into the middle of the chain and see what actually fires the trigger. And we ablate during generation to see whether the model can still stop. The third intervention is the headline.

Ablating two MLPs kills the decision. We mean-ablate the top components at the trigger position [28], replacing each output with its average over answer-like boundaries of the same trace. These are newline tokens with a digit shortly before them, so the reference class matches the token type of the trigger. Ablating the L27 MLP alone cuts the stop margin from +14.6 to +1.9 and $p($`</think>`$)$ from 1.00 to 0.55, but `</think>` survives as the greedy choice in every trace. Adding the L26 MLP flips it. `</think>` stops being the top token in 99% of traces and its probability falls to 0.03. The top four drive the margin to -4.4 and the top eight to -17.7. Randomly chosen components matched in count and type change nothing ($p = 1.000$ throughout). Two checks guard this result. Resample ablation swaps in the same component's output from a matched boundary of a *different* trace, which keeps everything on-distribution, and it agrees with mean ablation. Zeroing the output instead is noticeably weaker ($p = 0.79$ for L27), because RMSNorm partially rescales away a removed component. The second check is

surgicality. The same top-8 ablation applied at a matched non-trigger boundary leaves the local next-token distribution largely intact, with a median KL of 0.18 and the top-1 token changing in 21% of cases. That is an order of magnitude below the ∼32-logit swing at the trigger. The lesion is specific to the stop decision, not general damage.

There is no self-repair, but there is structure. Downstream components sometimes compensate for ablated ones, which masks necessity [15, 19]. We rerun the attribution of Section 6 under ablation to check. Nothing repairs. With the L27 MLP ablated, the L26 contribution moves only from +2.8 to +2.9. The reverse direction is more interesting. Ablating the five earlier MLPs of the cohort cuts L27's own contribution from +12.3 to +8.5, so about 30% of what L27 writes depends on what the earlier layers computed. The cohort is not six independent lookups. The earlier MLPs feed L27, and L27 does most of the writing.

After S, the stop state transfers from any layer. We patch the residual stream at the trigger into a mid-chain boundary of the same trace, one layer at a time. The stop margin recovers 0.85 of the way even when patching at layer 0, which is the token embedding alone, and recovers fully by layer 15. The layer-0 number should not be over-read. Patching that early effectively teleports the trigger token into the new position, so it mainly says that a trigger-shaped token in a post-S context is nearly sufficient by itself. The test also cannot separate "the state is already in the residual" from "downstream attention needs the answer sentence in the context", because the mid-chain context contains no answer sentence for attention to read. The splice test supplies that context behaviorally, and we read the two together.

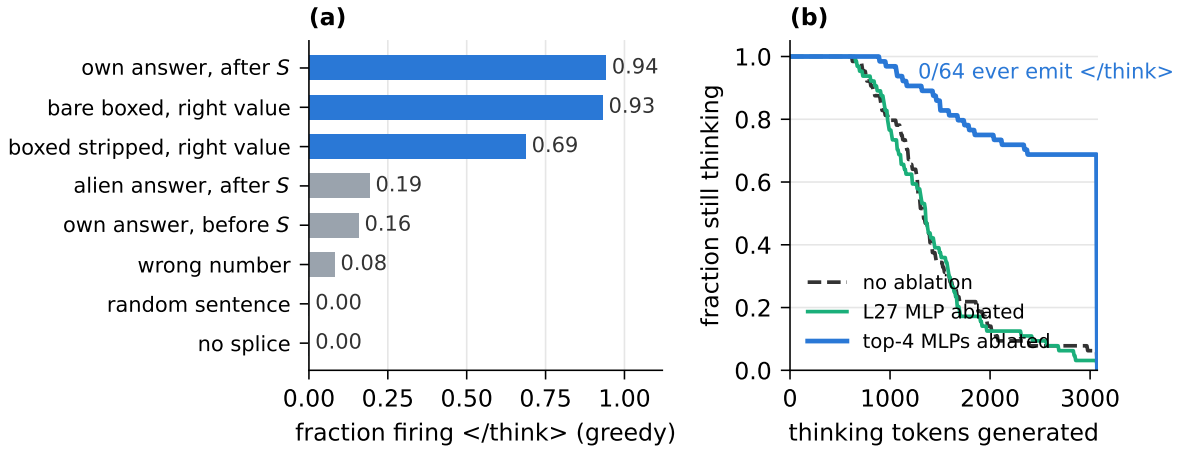

Figure 4: **The trigger is a verification gate.** (a) Fire rates when a final-answer sentence is spliced into the middle of the thinking segment ($n = 400$, and 352 for pre-S conditions). The model's own answer after S fires the trigger in 94% of traces. The same sentence with the digits changed fires in 8%, and the same sentence placed before S fires in 16%. Format alone is not enough, and the value alone is not enough. (b) Thinking-length survival when MLPs are ablated only at sentence boundaries during generation. With the top-4 MLPs ablated, no run ever emits `</think>` (0/64). The late decline in that curve is runs ending abnormally, with end-of-message emitted inside thinking, not stopping. The text remains coherent, with clean-model NLL 0.17 versus baseline 0.16.

The trigger fires on the model's own verified answer, not on format. We take each trace's final answer sentence ("So the final answer is `\boxed{72}`.") and splice it into the middle of the thinking segment, at a boundary after S. Then we check whether `</think>` becomes the greedy next token. It fires in 94% of traces (Figure 4a). An obvious objection is that this is a trigram. `\boxed{X}.` precedes `</think>` essentially always in reasoning traces, so a shallow pattern-matcher would fire on the format alone. The controls rule that out. The same sentence with its digits shifted, re-encoded the same way, fires in 8% of traces. Format alone does not

fire the trigger. Another problem's answer sentence, identical in format but alien in content, fires in 19%. And the model's own correct answer sentence, spliced *before* the point where the reasoning suffices, fires in only 16%. The right value at the wrong time does not fire it either. A random non-answer sentence and the no-splice baseline sit at zero. Stripping the `\boxed` markup drops firing to 69%, while a bare `\boxed{value}.` fires at 93%. So the markup is the main statement cue, but it is a cue the gate only accepts when the value checks out. The full condition table is in Appendix D. Two conditions must hold at once, an answer statement in the context and a stated value that matches the answer the model has already computed. That conjunction is a verification gate. Mediation confirms the spliced firings run through the same circuit. Rerunning the attribution at the spliced trigger gives L27 +10.0, against +12.3 at the real trigger, and the rest of the cohort matches similarly.

Ablate four MLPs at sentence boundaries and the model cannot stop thinking. The logit-level results could still be an artifact of teacher forcing, so we ablate during generation. The cohort contributes almost nothing at non-trigger boundaries (Section 6), which lets us scope the ablation to newline tokens only and leave every other decode step untouched. Ablation ends once `</think>` is emitted, so answers are always written by an intact model. The baseline, regenerated under identical sampling, stops in 60 of 64 runs, with a median thinking length of 1330 tokens and accuracy of 0.98 when stopped. Ablating the L27 MLP alone changes nothing behaviorally (60/64), matching the logit result. Ablating the top-4 MLPs flips the behavior completely. **Not one of 64 runs ever emits** `</think>` (Wilson 95% CI $[0.00, 0.06]$, against $[0.85, 0.98]$ for the baseline). 69% run to the 3072-token cap, and the rest end abnormally by emitting end-of-message while still inside the thinking segment. The coherence check is what makes this result meaningful. The endless text is not the output of a damaged model. Its NLL under the clean model is 0.17, against 0.16 for the baseline. The model reasons normally, states answers, and keeps going, because the component that turns a stated answer into a stop decision is gone. The blunt version of the same ablation, L27 at every token instead of only at boundaries, degrades the text (NLL 0.40) yet mostly still stops (52/64). Scoping isolates the stop function rather than general text quality.

The gate explains the earlier sections. The flat pre-trigger margin of Section 6 is no longer puzzling. Mid-chain boundaries fail the gate's first condition, since no answer statement has occurred, so the margin stays pinned no matter how sufficient the reasoning is. This also fits the finding that margin-based early exit only works at sentence boundaries [21]. And the overthinking gap of Section 5 now has a mechanism. The gate only *evaluates* when the model chooses to state a final answer, and nothing upstream forces that statement to happen at S. The model does not stop late because it cannot tell it is done. It stops late because the check that tells it it is done only runs when an answer gets written down.

8 The Gate Has No Simple Handle

If the stopping decision were a scalar creeping toward a threshold, a single residual-stream direction should be able to push it over. Prior work steers reasoning length with exactly such directions [13, 24, 26], and our localized circuit gives an unusually precise target to aim at. So we try to operate the gate from the inside, with no spliced text, just vectors added to the residual stream [25, 33].

We build an "about to stop" direction as the mean difference between trigger residuals and post-S boundary residuals, extracted from 204 traces and evaluated only on the 200 held-out ones. A logistic probe trained on the same contrast recovers nearly the same direction, with a cosine of 0.92 to 0.95 against a random-direction null of sd 0.022, so the contrast is not an artifact of the estimator. Alongside it we test three comparators. The *circuit direction* is the mean trigger-minus-boundary output of the L26 and L27 MLPs, injected at the final-norm input. The *generic* direction is built by the recipe of Sun et al. [24]. The *blunt* one is the raw `</think>`

unembedding pushed through the residual stream, and a norm-matched random direction serves as the null. Steering strength is stated as a fraction of the local residual norm and swept up to 0.35, the upper end of what keeps generation coherent. Full dose-response curves and the anti-illusion checks [14] are in Appendix E.

No direction fires the trigger from mid-chain. Across every direction, every injection layer, and every strength, the fired rate at post-S boundaries is exactly 0.00. The pushes are real and dose-monotone. The circuit family lifts the `</think>` margin by up to +11.8 logits, the blunt bias by +14.7, the generic direction by +3.7, and the random direction by +0.2. They simply do not come close, because the mid-chain margin sits near −42. Two further measurements say why. The best single direction recovers only **0.20** of the effect of patching the full residual state at the same layer and positions, the intervention Section 7 showed transfers the decision completely. And mediation under steering engages the circuit only partially, with L27 contributing +1.3 under the steered push against +12.3 at a real trigger. The "ready to stop" state is a high-dimensional configuration, not a dial.

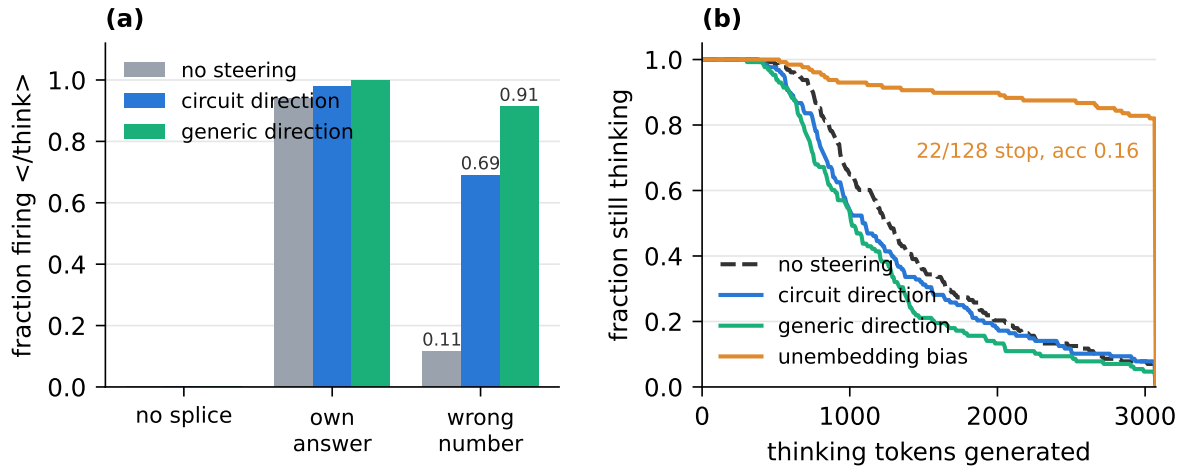

Figure 5: **Steering cannot fire the gate, but it can tip a marginal decision.** (a) Fire rates for splice contexts crossed with steering ($n = 148$). With no splice, steering does nothing. After a wrong-number splice, steering overrides the value check, from 0.11 to 0.69 with the circuit direction and to 0.91 with the generic one. (b) Generation-time steering at sentence boundaries ($n = 128$). Directions shorten thinking modestly at preserved accuracy. The raw unembedding bias at matched norm derails generation, with 22/128 stopping and accuracy 0.16.

Near the trigger, steering tips the decision, and can override the value check. We cross the splice contexts of Section 7 with steering (Figure 5a). In a no-splice context, steering leaves the fire rate at zero. After the model's own answer statement, where the decision is already essentially made, steering nudges 0.94 up to 0.98 with the circuit direction and to 1.00 with the generic one. The informative cell is the wrong-number context. Firing goes from 0.11 to 0.69 under the circuit direction and to 0.91 under the generic one. The gate's verification is a soft preference in a logit-level race, not a hard veto. The circuit direction is more selective than the generic knob, since it lifts wrong-number contexts less, and that points the gate model's way. But the two directions' effective in-context strengths diverged from their calibrated values, so we report the selectivity comparison as suggestive (Appendix E).

At generation time, direction steering is a real but generic length knob. Steering at sentence boundaries during decoding shortens thinking by 13.5%, from a median of 1265 tokens to 1094, at unchanged accuracy (0.90 versus 0.91) and unchanged coherence. The generic direction does slightly better. It cuts 20% at accuracy 0.93, which matches the published result [24]. The circuit-derived direction confers no extra length-control power, and we report that

plainly. The instructive contrast is the blunt unembedding bias at the same norm. Only 22 of 128 runs stop, accuracy collapses to 0.16, and NLL rises to 0.25 (Figure 5b). Direction quality matters, and magnitude does not substitute for it.

Length planning and stopping are two different mechanisms. Sheng et al. [20] find a pre-generation direction whose magnitude predicts and controls how long the model will think. That direction exists in our traces too. Last-prompt-token residuals predict log thinking length with $R^2 = 0.27$ at layer 21. It is almost orthogonal to our trigger direction. The maximum $|\cos|$ across layers is 0.059, against a random-direction null of sd 0.022. A planned-length bias set before generation and a context-keyed firing event are different objects. Sheng et al. [20] steer the former, and the termination circuit is the latter.

The negative result is what completes the picture. The only reliable early-stop lever we found in four experiments is the one the gate itself uses. Make the model state its answer. Splicing the statement fires the trigger 94% of the time, while no vector we could construct fires it at all.

9 The Same Gate at Scale, on Hard Problems, and in Another Family

Everything so far is one model on one dataset. A mechanism that only shows up in Qwen3-1.7B on GSM8K would be a curiosity, so we replicate the headline measurements in three progressively harsher cells. **Qwen3-8B** on GSM8K changes the scale, with 5× the parameters. **Qwen3-1.7B** on MATH-500 [7, 12] changes the difficulty, moving to problems where the reasoning is genuinely needed. **Magistral-Small-24B** [17] on MATH-500 changes the model family. It is a different lab's RL-trained reasoner with a different tokenizer, and it marks its thinking segment with `[THINK]` and `[/THINK]` instead of `<think>` and `</think>`. Each cell reruns sufficiency with the mismatched-prefix control, the trigger attribution, the splice test, and the ablations, at smaller n than the main experiments. The localization criterion was fixed before running. The top layer must sit in the final 25% of depth, and the MLP must beat attention at the trigger. Engineering details, budgets, and completion rates are in Appendix F.

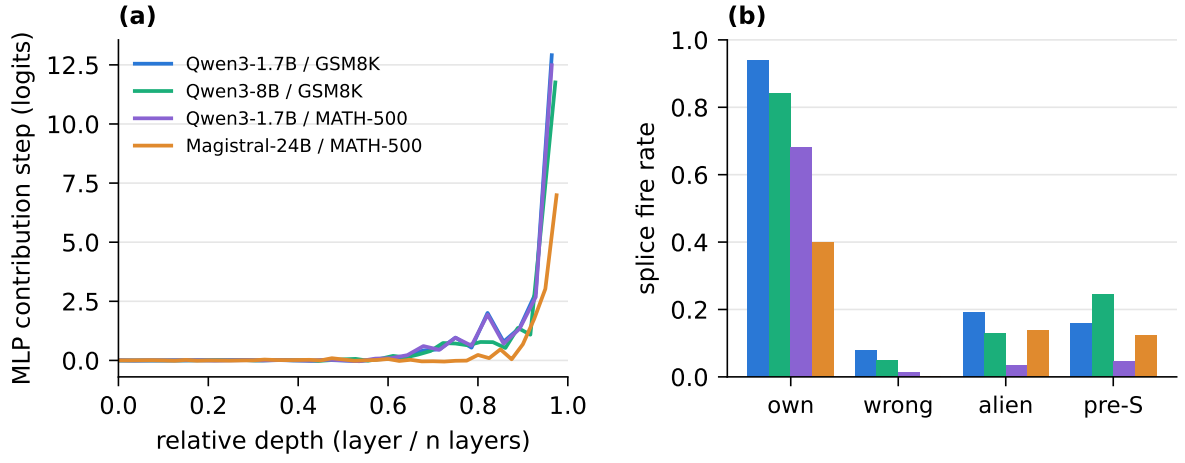

Figure 6: **The same circuit shape and the same gate signature in every cell.** (a) Per-layer MLP contribution step (trigger minus matched boundaries) against relative depth. All four cells peak at the same relative depth of 0.96 to 0.97, regardless of absolute layer count (28, 36, or 40 layers). (b) Splice fire rates per cell. The ordering that defines the gate, with the model's own answer far above wrong number, alien answer, and pre-S, holds everywhere.

	Qwen3-1.7B GSM8K	Qwen3-8B GSM8K	Qwen3-1.7B MATH-500	Magistral-24B MATH-500
layers	28	36	28	40
overthinking gap (median)	0.70	0.80	0.60	0.80
top-4 MLPs	27/26/23/25	35/34/32/33	27/26/23/25	39/38/37/36
top layer, relative depth	0.96	0.97	0.96	0.97
top MLP step (logits)	+12.9	+11.8	+12.5	+7.0
splice own / wrong / alien	.94 / .08 / .19	.84 / .05 / .13	.68 / .01 / .04	.40 / .00 / .14
ablation flip, top-2 / top-4	.99 / 1.00	.64 / .99	.93 / 1.00	.91 / 1.00
generation ablation, stop rate	0/64	0/24	0/24	17/24[†]

Table 1: **Replication across scale, difficulty, and model family.** The MLP step is the mean trigger-minus-boundary contribution of the top layer. [†]The baseline is also 17/24. The generation-time ablation targets newline boundaries, a rule tuned to Qwen's trace format, and it missed Magistral's trigger position, so this cell is inconclusive rather than negative (Appendix F). Magistral's necessity evidence is the logit-level flip test.

Scale is a full replication. On Qwen3-8B every measurement reproduces. The leakage control is airtight. Real forced accuracy climbs from 0.36 to 1.00 across fractions, while the mismatched control matches the question-only floor at cut zero and collapses once mismatched reasoning enters. The trigger is a late-MLP step at relative depth 0.97 (L35 of 36, +11.8 logits, attention at most +4.2). The splice signature holds, with the model's own answer at 0.84, wrong number at 0.05, and alien at 0.13. Ablation shows the same joint structure. One MLP never flips the decision, two mostly do, and four always do. With the top four ablated at boundaries, 0 of 24 generations ever emit `</think>`, against 24 of 24 for the baseline, at matched coherence (NLL 0.181 versus 0.165). One number moves in an interesting direction. The overthinking gap median is **0.80**, larger than the 1.7B's 0.70. The bigger model overthinks more, not less.

Difficulty leaves the circuit untouched and shrinks the waste. On MATH-500 the trigger is carried by the *identical* four MLP layers as on GSM8K, namely L27, L26, L23, and L25, in the same order and at nearly the same magnitude (+12.5). The circuit does not move when the task changes. The gate signature holds, with the own answer at 0.68, wrong number at 0.01, and alien at 0.04. The lower absolute rate reflects that MATH answer statements are long LaTeX expressions, and splicing them mid-chain is noisier, but the discrimination the claim rests on stays sharp. Ablation replicates, flipping 0.93 of traces at top-2 and all of them at top-4, and 0 of 24 ablated generations stop against 24 of 24 at baseline. The overthinking gap median drops to 0.60. On genuinely hard problems the model wastes less of its reasoning, as it should. But the gap does not vanish, and it is not an easy-problem artifact. The per-level gap medians are 0.80, 0.60, 0.80, 0.60, and 0.60 for levels 1 through 5. Even on level-5 MATH the median trace fixes its answer with 60% of its thinking still to run.

The gate is not a Qwen quirk. Magistral-Small-24B replicates four of the five measurements across a model-family boundary. Sufficiency precedes the trigger with an airtight leakage control (real 0.24 to 0.70, control flat), and the model overthinks heavily. Its gap median is 0.80, shrinking with difficulty but holding at 0.60 on the hardest levels. The trigger has the same shape. The top-4 MLPs are L39, L38, L37, and L36 at relative depth 0.97, MLP-carried at +7.0 against attention's +1.9, a smaller magnitude than Qwen's but the identical profile (Figure 6a). The gate signature is intact and, if anything, sharper. The model's own answer fires at 0.40, a wrong number at **0.00**, an alien answer at 0.14. The logit-level flip test replicates too (0.91 at top-2, 1.00 at top-4). The one measurement that did not transfer is the generation-time cannot-stop test, and the failure mode is diagnostic rather than damning. Ablated traces got *shorter* (median 6214 versus 10066 tokens), not non-terminating. The newline-scoped ablation tuned to Qwen's trace format perturbed mid-reasoning without ever hitting Magistral's actual `[/THINK]` decision point. We report that cell as inconclusive and rest Magistral's necessity evidence on

the precisely targeted flip test.

Across all three cells the mechanism is stable, and only the amount of redundant reasoning moves. What this section does not test is distillation. All three models learned to stop with reinforcement learning. Whether a model that *imitated* reasoning traces develops the same gate is an open question we return to in the discussion.

10 Discussion

Assembled, the picture is simple. By the middle of a typical trace, the reasoning already determines the answer (Section 5). Nothing in the network acts on that fact. The decision to stop is written in one step by a handful of late MLPs (Section 6). Those MLPs implement a gate that fires only when an answer statement appears in the context and its value matches what the model has computed (Section 7). The gate cannot be operated by any single direction we could build, and the only lever that fires it is the event it watches for (Section 8). None of this is particular to one model or dataset (Section 9).

This explains the shape of the prior literature. Margin monitors find usable signal only at sentence boundaries [21] because the gate's first condition fails everywhere else and the margin stays pinned 40 logits down. Attention around the delimiter carries a termination signature [11] because the trigger step is where the stop state finally forms. Steering directions shorten reasoning modestly and no further [13, 24]. Our faithfulness measurement suggests why. A single direction captures about a fifth of a high-dimensional stop state, and pushing harder with a cruder vector destroys generation before it fires the gate. The pre-generation length direction of Sheng et al. [20] is orthogonal to the trigger, so length planning and termination are separate levers, and an intervention on one should not be expected to move the other.

The practical reading is that the efficient lever is textual, not representational. The gate fires on a stated answer at 94%. A method that makes the model state its candidate answer early, and lets the gate's own verification decide whether to stop, works with the mechanism. Activation-level pushes work against a 40-logit wall. The wasted computation lives in the gap between sufficiency and statement, not in any inability to detect completion.

Limitations. Our value-check claim has a scope boundary. All analyzed traces are ones the model answers correctly, so "matches the answer the model computed" and "matches the correct answer" coincide in our data. The pre-S control argues for the internal reading, since the correct value fails to fire the gate before the model has derived it, and the gate has no access to ground truth in any case. The direct test would use traces where the model internally settles on a wrong answer, where the gate should fire on that wrong value. We have not run it. Second, everything here is mathematical reasoning with short, checkable answers. Open-ended domains where "the stated value matches" is ill-defined may stop differently. Third, the generation-time ablations and the replication cells run at modest n (24 to 64 generations per condition), so their effect sizes are coarse even where the direction of the effect is unambiguous. Fourth, Section 8 bounds linear, one-dimensional handles only. A nonlinear or multi-directional controller could in principle exist. Fifth, the Magistral generation-time ablation is inconclusive for the format reasons above. Finally, all our models learned to reason with reinforcement learning. Distilled reasoners, which imitate stopping rather than learn it, are deliberately out of scope and are the natural next experiment.

11 Conclusion

We asked what, inside a reasoning model, decides that thinking is over. The answer is a small, stereotyped circuit of late MLPs that implements a verification gate. It fires when the model has written down an answer that matches the one its reasoning produced. The gate is causally

necessary, since removing four MLPs leaves the model unable to stop. It is sufficient given the event it watches for, since splicing in the answer statement fires it. And it is closed to simple manipulation, since no residual direction we could build fires it. The same gate appears at the same relative depth across model scale, task difficulty, and model family. Overthinking, on this account, is not a detection failure. The model knows the answer. The circuit that ends the thinking never asks until the model chooses to say it.

Acknowledgments

The experiments in this paper ran on compute from Modal, Lightning AI, and Google Colab. We thank these platforms for the free tiers and research credits that make independent work like this possible.

References

[1] David D. Baek and Max Tegmark. Towards understanding distilled reasoning models: A representational approach. *arXiv preprint arXiv:2503.03730*, 2025.

[2] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, et al. Do NOT think that much for 2+3=? on the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*, 2024.

[3] 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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[4] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[5] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

[6] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

[7] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[8] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

[9] Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuanda Wang, Zhixia Zhang, Hongyan Xie, Songshi Liang, Zehao Chen, Xuefeng Xiao, Fuzhen Zhuang, Jianxin Li, Deqing Wang, and Yikun Ban. Does your reasoning model implicitly know when to stop thinking? *arXiv preprint arXiv:2602.08354*, 2026.

[10] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023.

[11] Gengyang Li, Wang Cai, Yifeng Gao, and Yunfang Wu. SyncThink: A training-free strategy to align inference termination with reasoning saturation. *arXiv preprint arXiv:2601.03649*, 2026.

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

[13] Zhengkai Lin, Zhihang Fu, Ze Chen, Chao Chen, Liang Xie, Wenxiao Wang, Deng Cai, Zheng Wang, and Jieping Ye. Controlling thinking speed in reasoning models. *arXiv preprint arXiv:2507.03704*, 2025.

[14] Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. *arXiv preprint arXiv:2311.17030*, 2023.

[15] Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

[16] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *arXiv preprint arXiv:2202.05262*, 2022.

[17] Mistral AI. Magistral. *arXiv preprint arXiv:2506.10910*, 2025.

[18] nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

[19] Cody Rushing and Neel Nanda. Explorations of self-repair in language models. *arXiv preprint arXiv:2402.15390*, 2024.

[20] Leheng Sheng, An Zhang, Zijian Wu, Weixiang Zhao, Changshuo Shen, Yi Zhang, Xiang Wang, and Tat-Seng Chua. On reasoning strength planning in large reasoning models. *arXiv preprint arXiv:2506.08390*, 2025.

[21] Sangjun Song, Minjae Oh, Seungkyu Lee, Sungmin Jo, and Yohan Jo. ThinkBrake: Efficient reasoning via log-probability margin guided decoding. *arXiv preprint arXiv:2510.00546*, 2025.

[22] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

[23] Chung-En Sun, Ge Yan, and Tsui-Wei Weng. ThinkEdit: Interpretable weight editing to mitigate overly short thinking in reasoning models. *arXiv preprint arXiv:2503.22048*, 2025.

[24] Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. LLM reasoning as trajectories: Step-specific representation geometry and correctness signals. *arXiv preprint arXiv:2604.05655*, 2026.

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

[26] Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Understanding reasoning in thinking language models via steering vectors. *arXiv preprint arXiv:2506.18167*, 2025.

[27] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, et al. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

[28] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2020.

[30] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[31] Jue Zhang, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. From reasoning to answer: Empirical, attention-based and mechanistic insights into distilled DeepSeek R1 models. *arXiv preprint arXiv:2509.23676*, 2025.

[32] Haoran Zhao, Yuchen Yan, Yongliang Shen, Haolei Xu, Wenqi Zhang, Kaitao Song, Jian Shao, Weiming Lu, Jun Xiao, and Yueting Zhuang. Let LRMs break free from overthinking via self-braking tuning. *arXiv preprint arXiv:2505.14604*, 2025.

[33] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Reproducibility

Code. All experiment code is available at
https://github.com/Chandram-Dutta/the-termination-circuit.
Each experiment is a single script that writes its figures and stage checkpoints to a local output directory, and the repository README maps each script to the section it supports.

Hardware and software. Experiments on Qwen3-1.7B and Qwen3-8B run on a single NVIDIA L40S (48 GB) in bfloat16. The Magistral-24B analysis runs on a single RTX PRO 6000 (96 GB), since the 24B weights alone need about 48 GB. All interventions use forward hooks on the HuggingFace `transformers` implementation [29]. Trace generation for the cross-family cells uses vLLM [10] with pre-tokenized prompts, and the generated token ids are read back directly, so the stored traces index the analysis model's embedding exactly. Mechanistic stages stay on the hooked HuggingFace models.

Sampling and budgets. Each model runs with its recommended sampling settings under a fixed seed (Qwen3 uses temperature 0.6, top-p 0.95, top-k 20). Token budgets are 4096 for GSM8K, 8192 for MATH-500 on Qwen models, and 16384 for Magistral, which reasons long enough that an 8192 budget truncated 69% of its traces mid-thought. Completion rates, meaning traces that emit the close marker within budget, are 441/500 for 1.7B on GSM8K, 134/150 for 8B on GSM8K, 124/150 for 1.7B on MATH-500, and 60% for Magistral. Truncated traces are excluded from every stage, and the exclusion is reported, since it can bias toward shorter-reasoning problems.

Tokenization. `</think>` is token 151668 in the Qwen3 tokenizer. Magistral is tokenized with `mistral-common` rather than the auto-converted HuggingFace tokenizer. The converted `tokenizer.json` has a vocabulary that does not match the model's embedding, and it encodes `[THINK]` and `[/THINK]` as plain strings rather than single control tokens. `mistral-common` yields the single-token `[/THINK]` that the attribution and gate stages require. The close-marker search runs only over the generated region, because the non-Qwen system prompts contain the literal marker as instruction text.

Scoring. GSM8K answers are scored by integer match. MATH-500 answers are normalized with a compact Hendrycks-style `strip_string` normalizer and compared exactly, with no substring containment, which would inflate early-stop accuracy for short answers and bias S early. Correctness reads the last `\boxed{}` anywhere in the trace, because verbose reasoners box the answer inside the reasoning rather than after the close marker. The forced-answer cue is `</think>\n\nThe final answer is` on GSM8K. On MATH the cue ends in `\boxed{` and the answer is read to the matching brace, so long LaTeX answers are not truncation-scored as wrong.

Pre-registration. The eligibility rules for which traces enter each stage, the S rule, and the late-MLP-step criterion of Section 9 were fixed and printed before results in every run.

B Decoding the Answer from the Residual Stream

The representational version of the sufficiency measurement asks when the final answer *value* becomes linearly decodable from the residual stream. A ridge probe maps the residual at each fraction-grid position to log(answer), fit out-of-fold by trace. Decodability is scored as two-alternative forced-choice retrieval, asking whether the prediction lands closer to the true answer than to a random other trace's answer (chance 0.5). Retrieval rises from a floor of 0.51 at cut zero to 0.70 at layer 13. The signal is real, but weak and noisy. For calibration, sequence position is strongly decodable from the same residuals ($R^2 = 0.84$ at layer 11), so the weakness is specific to the answer value. The answer is only partially linearly readable mid-chain, which is one more way the stop state resists one-dimensional handles. The behavioral measure S is the cleaner signal, and the paper uses it throughout.

C Direct Logit Attribution Details

The final residual stream h is the sum of the token embedding and every attention-head and MLP output. The logit gap between `</think>` and the vocabulary mean is $(h \cdot \bar{u})/\sigma$, where $\bar{u}$ is the mean-centered `</think>` unembedding row and σ is the RMSNorm denominator at that position. Freezing σ at its true per-position value makes the map linear in the components, so each component's contribution is its output dotted with $\bar{u}/\sigma$ [5, 18]. Attention is decomposed per head by slicing the output projection. Freezing matters. The norm denominator at the trigger (median about 71) is substantially larger than at matched boundaries (about 47), and an attribution that let σ vary would mix that rescaling into every component. The decomposition is verified exactly, with summed contributions reconstructing the model's actual centered logit to a maximum error of 0.10 against a total of 30.5. In the replication cells the same identity check is rerun per model before any attribution is trusted, which matters when the module tree and norm placement differ across families.

D Full Splice Results

condition	n	fired (greedy `</think>`)
own answer sentence, after S	400	0.94
bare `\boxed{value}.`, right value	398	0.93
`\boxed` stripped, right value	398	0.69
alien answer sentence, after S	398	0.19
own answer sentence, before S	352	0.16
wrong number (digits shifted)	400	0.08
alien answer sentence, before S	350	0.02
random non-answer sentence	376	0.00
no splice, after S	400	0.00
no splice, before S	352	0.00

Table 2: All splice conditions on Qwen3-1.7B and GSM8K. Each condition is compared against the no-splice baseline at its own position, since pre-S contexts are shorter.

The spliced sentence is text the model itself wrote, so the wrong-number and alien controls are what separate content from format. Rerunning the attribution of Section 6 at spliced triggers that fired gives L27 +10.0 on average, against +12.3 at real triggers, and the earlier cohort members match similarly. Spliced firings run through the same circuit. That check is what converts the splice test from a behavioral curiosity into evidence about the mechanism.

E Steering Details

Directions and calibration. The mean-difference direction is extracted at matched newline-boundary tokens, because mid-sentence grid positions would contaminate it with a "statement boundary" component. Injection layers 13, 18, and 24 are swept at strengths of 0.05, 0.1, 0.2, and 0.35 of the median local residual norm. Margin lifts are dose-monotone everywhere. The strongest configuration is layer 24 at 0.35, with a median lift of +11.6 at $n = 40$ and +11.8 at $n = 199$. The blunt and generic comparators in the main text are calibrated to the same median margin lift at the evaluation position.

Anti-illusion checks. A direction can raise a logit through a pathway unrelated to the circuit [14], so three checks accompany the steering results. Steering while mean-ablating the L26 and L27 MLPs never produced a firing that bypassed the circuit. The fired rate was 0.00 both with and without ablation, and at a 0.00 base rate this check is consistent with the gate model but uninformative on its own. Mediation under steering shows the circuit only partially engaged, with L27 at +1.3 versus +12.3 at real triggers. The faithfulness ratio, defined as the best direction's margin lift divided by the full-residual-patch lift at the same layer and positions, is 0.20 (median lift +11.5 versus +55.5).

The $S\pm k$ test hit a floor. Steering at the adjacent boundaries straddling S (172 pairs, gap of at most 400 tokens) fired at 0.00 on both sides for all three directions. At 40 logits below threshold, no push at coherent strength moves the greedy choice, so the planned pre/post-S asymmetry test never got to run. We report it as a floor, not as evidence of symmetry.

The steering-by-splice calibration caveat. In the crossed conditions of Figure 5a, the circuit and generic directions were matched on margin lift at the calibration position, but their effective strengths inside splice contexts diverged. The circuit direction's greater selectivity

(0.69 versus 0.91 on wrong-number contexts) is therefore suggestive rather than a controlled comparison.

F Cross-Family Details

Magistral cell. Magistral-Small-2507 (24B, 40 layers) on MATH-500 attempted 150 traces at a 16384-token budget, completed 60%, and carried 88 correct traces through analysis (difficulty levels 1 to 5 contribute 9/21/20/25/13). Completion falls with difficulty, down to 38% at level 5, so the hardest-level results skew toward shorter-reasoning problems. The per-level gap medians in Section 9 should be read with that in mind. Splice conditions run at $n = 60$ for the own answer, 53 for wrong-number (which requires digits in the answer), 58 for alien, and 41 for pre-S.

Why the generation-time ablation is inconclusive. The boundary-scoped ablation triggers on newline tokens, which at Qwen sit directly before `</think>` in essentially every trace. Magistral's trace format does not share that regularity. Under ablation, Magistral's stop rate was unchanged at 17/24 versus 17/24 baseline, while its traces got *shorter* (median 6214 versus 10066 tokens). If the ablation were removing the trigger computation, ablated traces would run to the token cap. Shortening instead means the ablation perturbed mid-reasoning states without being active at the actual `[/THINK]` decision point. The test's targeting failed to transfer, not the gate. A format-aware boundary detector for Magistral is future work. The necessity claim for Magistral in the main text rests on the flip test, which targets the trigger position exactly and replicates at 0.91 and 1.00 for top-2 and top-4.