STOPPING SOFT REASONING BEFORE IT LOOPS: FROM IMPASSE-ORIENTED DETECTION TO EVENT-LEVEL INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large reasoning models (LRMs) achieve strong performance by generating long chains of thought (CoT), but discrete token sampling discards distributional information at each step. Training-free soft reasoning retains this information through probability-weighted embeddings, yet repeated (soft) feedback can prevent termination. We show that this failure is not merely out-of-distribution drift by soft representations: soft trajectories can amplify an impasse-induced recurrent regime already present in hard decoding, in which non-adjacent semantic events revisit similar hidden-state regions and predictive entropy progressively contracts. Motivated by this finding, we propose E-TRACE, a training-free, event-level detector that identifies persistent non-local recurrence in projected hidden-state distributions. E-TRACE stores entropy-decreasing events as rollback candidates, and once recurrence is confirmed, restores the nearest candidate outside the recurrent region, truncates the recurrent suffix, and switches to hard decoding. Experiments show that the method can accurately predict non-termination and improve overall reasoning performance, achieving higher answer accuracy with substantially fewer generated tokens.

1 INTRODUCTION

Large reasoning models (LRMs) have achieved strong performance by generating explicit chains of thought (CoT) (Wei et al., 2022; Guo et al., 2025). Standard CoT decoding, however, commits to a single discrete token at each step, collapsing the model's next-token distribution before passing the reasoning state forward. This discrete bottleneck can discard alternative semantic information that may be useful for subsequent reasoning. Continuous-space reasoning seeks to alleviate this limitation by propagating soft or latent representations, thereby preserving richer information throughout the reasoning process (Chen et al., 2025; Li et al., 2025). Particularly, training-free latent reasoning constructs continuous representations, known as *soft tokens*, directly from a probability-weighted combination of token embeddings at inference time, avoiding task-specific retraining (Hao et al., 2024; Xu et al., 2025a;b) and reducing adaptation costs (Zhang et al., 2026). But despite its ease of application to existing large reasoning models (LRMs), autoregressive soft reasoning can introduce a critical issue: the reasoning process may fail to terminate.

Prior work primarily attributes this issue to distribution mismatch, as soft representations may deviate from the discrete embedding distribution encountered during training (Zhang et al., 2026). Soft Thinking mitigates this risk through Cold Stop, which terminates soft reasoning when output entropy remains low for several consecutive steps (Zhang et al., 2026). Entropy alone, however, neither provides a reliable criterion for identifying when intervention is needed nor explains the mechanism underlying non-termination. SeLaR (Fu & Luo, 2026) instead argues that applying soft activations indiscriminately can perturb high-confidence steps and therefore proposes to activate soft embeddings only at high-entropy exploratory steps while retaining discrete decoding elsewhere. Nevertheless, it does not explicitly address non-termination or provide a mechanism for determining when a recurrent soft-reasoning trajectory should terminate.

We observe that soft-token trajectories do not merely drift without producing a stopping signal. Instead, they can become trapped in persistent reasoning loops. This phenomenon echoes recent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

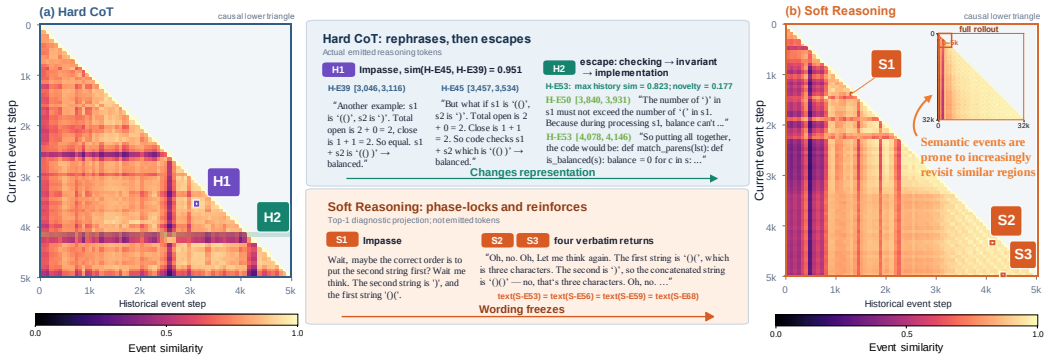


Figure 1: A concrete impasse yields divergent hard and soft trajectories. Hard CoT revisits a similar example but then changes representation and reaches an invariant-based implementation. Soft-token feedback instead phase-locks into repeated diagnostic states: non-adjacent events become highly similar, their wording freezes, and the same cycle persists through the full rollout.

findings in discrete LRMs, where unresolved reasoning *impasses* can induce semantic repetition and self-reinforcing loops (Pipis et al., 2026; Duan et al., 2026). Our analysis demonstrates that soft reasoning inherits an intrinsic tendency toward impasse-induced cycles rather than creating it from scratch. We refer to the continuous semantic segment as an *event* and find that, as the soft-reasoning trajectory evolves, non-adjacent semantic events increasingly revisit similar regions, eventually forming recurrent patterns that lead to reasoning impasses. Fig. 1 illustrates such findings, while more details are discussed in Section 3. These observations motivate an event-level intervention that detects persistent non-local recurrence, identifies a recoverable pre-loop state, and switches to hard token decoding before the trajectory becomes deeply locked into the recurrent regime.

Our contributions are summarized as follows:

- We identify and theoretically analyze how training-free soft reasoning amplifies the model's intrinsic cyclic tendency after a reasoning impasse, worsening persistent semantic recurrence and eventually leading to deep lock-in and non-termination.
- We introduce E-TRACE, a training-free **E**vent-level **T**rajectory **R**ecurrence **A**nalysis and **C**heckpointed **E**scape method. E-TRACE confirms loops through persistent non-local recurrence across consecutive event windows, not step-level tokens and isolated semantics, with entropy–novelty collapse as a late-stage safeguard.
- We develop a risk-aware intervention that records entropy-decreasing events as online candidates. Once a loop is confirmed, the controller restores the most recent candidate outside the recurrent region, removes the recurrent soft-state suffix, and switches to hard answer decoding.
- Experiments across mathematical reasoning, code generation, and knowledge-intensive question answering (Q&A) show that our method improves answer accuracy while reducing the average number of reasoning steps. The event-level signal also provides an accurate signal of non-termination.

2 Related Works

Early Stopping Explicit CoT improves multi-step problem solving by exposing intermediate reasoning steps (Wei et al., 2022; Khot et al., 2023; Zhou et al., 2023). As longer and parallel reasoning trajectories increase inference cost, current research indicates that LRM may engage in unnecessary or redundant reasoning, making early stopping an effective approach to improving token efficiency (Sui et al., 2025; Hou et al., 2026; Mao et al., 2026; Yang et al., 2026; Fu et al., 2025). Training-based approaches such as ThinkPrune use length-penalized reinforcement learning (RL) to encourage shorter reasoning trajectories (Hou et al., 2026). At inference time, Dynamic Early Exit monitors potential reasoning transitions and terminates generation when a trial answer is sufficiently confident (Yang et al., 2026). These methods reduce the cost of explicit CoT by estimating when the current

trajectory has become sufficient. Latent reasoning pursues a complementary route by reducing the need to externalize intermediate computation as discrete tokens.

Latent Reasoning Latent reasoning reduces explicit tokenization by propagating continuous intermediate representations, seeking to preserve distributional information discarded by discrete token selection while improving token efficiency (Chen et al., 2025; Li et al., 2025). Existing approaches differ in how these latent representations are constructed and integrated. Training-based approaches learn latent dynamics or auxiliary projection modules from task data (Hao et al., 2024; Xu et al., 2025a;b), whereas Soft Thinking constructs probability-weighted mixtures of token embeddings directly from a pretrained model at inference time (Zhang et al., 2026). SwiReasoning and SeLaR alternates between explicit and latent modes according to entropy state (Shi et al., 2026; Fu & Luo, 2026). These methods primarily address how soft states are constructed, activated, or propagated, but do not explicitly characterize non-termination as a persistent recurrence phenomenon. This gap motivates a closer examination of whether repeated soft-token transitions can amplify reasoning loops already present in discrete LRMs.

Reasoning Loops Repetitive degeneration has long been studied in neural text generation (Xu et al., 2022), but loops in LRMs exhibit structure tied to the reasoning process. Learning difficulty can favor easy cyclic actions over harder progress-making ones, while temporally correlated errors reinforce the same choice at recurrent decision points (Pipis et al., 2026; Duan et al., 2026). Persistent loops have also been linked to attention concentrated on a narrow recent suffix and stabilization through KV-cache reuse (Xu et al., 2026). In latent reasoning, soft representations constructed as probability-weighted mixtures of token embeddings can nevertheless remain dominated by their highest-probability components, causing them to behave similarly to top-1 input (Wu et al., 2026). Together, these results suggest that reasoning loops are self-reinforcing regimes rather than merely instances of surface-level repetition. Existing studies, however, do not explain how repeated soft tokens convert an intrinsic cyclic preference into persistent non-termination. We investigate this missing temporal mechanism and use persistent non-local recurrence across semantic events as its operational signal.

3 Event-Level View of Soft-Token Trajectories

In this section, we aim to provide an event-level view of soft-token trajectories and identify the reasons behind the non-termination phenomenon. Beforehand, we provide the definitions of the fundamental concepts and metrics.

Training-Free Soft Token Reasoning. Let $\mathcal{V}$ be the vocabulary and $E \in \mathbb{R}^{|\mathcal{V}|\times d}$ the input embedding matrix, whose row $\mathbf{e}_v^\top$ is the embedding of token v. At reasoning step t, the model produces a next-token distribution $p_t \in \Delta^{|\mathcal{V}|-1}$ after the decoding-time temperature and top-k transformation. Standard hard-token reasoning selects one token from this distribution and feeds its embedding back into the model. Training-free soft reasoning instead feeds the probability-weighted embedding

$$\mathbf{x}_t^S = E^\top p_t = \sum\nolimits_{v\in\mathcal{V}} p_t(v)\mathbf{e}_v, \tag{1}$$

while the greedy (i.e., top-1) hard token is $x_n^G = e_{\arg\max_i p_n(i)}$.

Semantic Events. We record a soft-reasoning trajectory as a sequence of hidden states $\{\mathbf{h}_t\}_{t=1}^T$ and denote the corresponding next-token distributions by $\{p_t\}_{t=1}^T$. To obtain a readable observer sequence, we extract the most probable token $\hat{z}_t = \arg\max_v p_t(v)$, but notably $\hat{z}_t$ is never fed back into the model. We partition consecutive soft-reasoning steps into variable-length semantic events, using sentence-ending punctuation and newline tokens in the observer sequence as natural boundaries (Sun et al., 2025). Meanwhile, when no natural boundary occurs, a maximum event length forces segmentation, while a minimum length prevents overly fragmented events. Let $\mathcal{T}_k$ denote the set of steps belonging to event k, and denote its length by $L_k = |\mathcal{T}_k|$.

Event Hidden-State Signature and Similarity. We use a Sliced Wasserstein Distance to represent the event-level similarity (Rabin et al., 2012). The hidden-state matrix of event k is $\mathbf{H}_k = [\mathbf{h}_t]_{t\in\mathcal{T}_k} \in$

$\mathbb{R}^{L_k \times d_{\text{hidden}}}$. We summarize it by an event signature $\mathbf{S}^{(k)} = \Psi_{\mathbf{Q}}(\mathbf{H}_k)$, which records the marginal order statistics of the hidden states along a fixed set of random orthogonal projection directions. For two variable-length events, their signatures are aligned on a common normalized-rank grid and compared to obtain

$$s_{i,j} := \mathrm{Sim}_{\mathbf{Q}}(\mathbf{H}_i, \mathbf{H}_j) \in (0, 1]. \tag{2}$$

A large $s_{i,j}$ indicates that the two events occupy similar empirical hidden-state regions. The projection, rank alignment, and normalized distance used by $\mathrm{Sim}_{\mathbf{Q}}$ are given in Appendix A.

Event-Level Uncertainty and Novelty. In addition to pairwise similarity, we retain the mean predictive entropy of each event,

$$H_t = -\sum_{v \in \mathcal{V}} p_t(v) \log p_t(v), \qquad \bar{H}_k = \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} H_t. \tag{3}$$

We also define the novelty of the current event relative to all previously completed events in the same trajectory,

$$\mathcal{H}^{(k)} = \{1, \ldots, k-1\}, \qquad N_k = 1 - \max_{h \in \mathcal{H}^{(k)}} s_{k,h}. \tag{4}$$

We set $N_1 = 1$ because no previous event is available. A large $s_{i,j}$ indicates j return to a hidden-state region similar to i, a small $\bar{H}_k$ indicates a concentrated next-token distribution after trajectory k, and a small N_k indicates that the current event k contributes little relative to the trajectory's prior semantic states. We will use them only as an observational lens to expose how soft-token non-termination develops; their role in online detection and intervention is introduced later.

Soft Reasoning in Impasse and Loop In Fig. 1, we use a paired hard–soft example to determine whether non-termination is merely an unstructured latent drift or a recognizable trajectory regime. Both rollouts encounter the same unresolved decision pattern, but only the soft-token trajectory fails to exit. The event-level view makes their divergence visible before the generation budget is exhausted. It can be observed from Fig. 1 that in *hard* CoT, events H-E45 and H-E39 are highly similar, yet the return is productive: the model subsequently changes representation at H2 and reaches executable reasoning. Thus, one similar event pair is not itself evidence of failure. However, in the *soft* rollout, events S1–S3 repeatedly return to the same diagnostic region, and later events reproduce the same wording. The growing non-local similarity block therefore exposes persistent residence in a recurrent region, rather than a single useful recheck. The complementary temporal view, in the subfigures (a) and (b) of Fig. 1, tracks how this divergence develops over the rollout. The hard trajectory retains substantial predictive uncertainty during recurrence, and its novelty rebounds when the reasoning shifts from repeated example checking to invariant-based implementation. By contrast, once the soft trajectory enters its exact cycle, event entropy contracts to near zero and novelty remains close to zero for the rest of the rollout. Fig. 2 further validates these findings in terms of the previously defined event-level similarity, entropy, and novelty.

4 Impasse-Conditioned Amplified Temporal Correlation in Soft Reasoning

Section 3 suggests that repeated soft-token injection amplifies a structured failure mode rather than causing unconstrained latent drift. When the LLM reaches an unresolved reasoning impasse, similar predictive distributions can repeatedly feed similar continuous inputs into subsequent states. This attenuates discrete perturbations that could otherwise change the reasoning, while loop-consistent states continue to accumulate in attention and the KV cache. We call this proposed mechanism *impasse-conditioned temporal amplification*. We now formalize the first link in this mechanism: how soft reasoning inherits temporally correlated predictions at repeated impasse visits.

Pipis et al. (2026) explain that when the correct progress-making action is hard to learn, but an easy cyclic action is available, the LLM puts relatively more probability on the cyclic action and repetitive impasse could emerge. Without loss of generality, let the n-th return to the same class of unresolved decision points be an *impasse visit*. Faced with the same reasoning impasse, the distribution of actions is similar (Pipis et al., 2026). Mathematically,

$$\|p_{n+1} - p_n\|_2 \leq \delta_n, \tag{5}$$

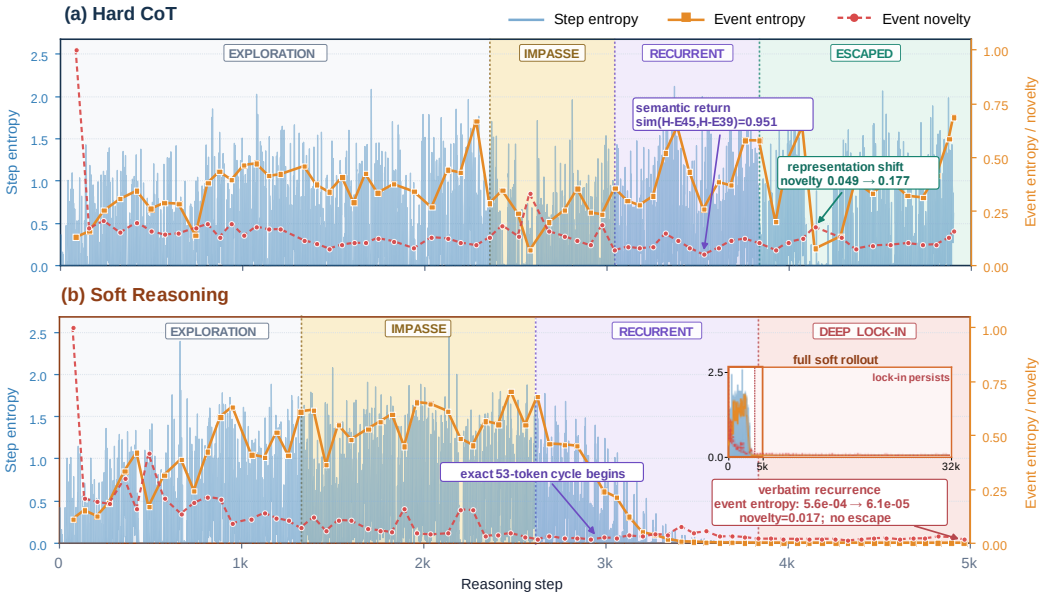


Figure 2: Event-level uncertainty dynamics separate escape from deep lock-in in the same illustrative comparison. Both trajectories encounter an impasse and semantic recurrence. Hard decoding retains uncertainty and eventually exhibits a novelty rebound as its representation changes. Soft-token feedback enters an exact cycle; event entropy then contracts toward zero and novelty remains suppressed, producing a lock-in that persists through the full rollout.

where δ_n is a relatively small constant. In other words, the LLM exhibits an inductive bias to re-select the previously favored actions. Next, we demonstrate why soft tokens further strengthen the possibility of visiting a reasoning impasse. We have the following proposition.

Proposition 1 (Temporal inheritance and enhancement)**.** *Under Eq. (5), soft token satisfies*

$$\|x^S_{n+1} - x^S_n\|_2 \leq \|E\|_2 \delta_n \ll \|x^G_{n+1} - x^G_n\|_2 \tag{6}$$

Proof. By definition and submultiplicativity,

$$\|x^S_{n+1} - x^S_n\|_2 = \|E^\top (p_{n+1} - p_n)\|_2 \overset{(a)}{\leq} \|E\|_2 \|p_{n+1} - p_n\|_2 \leq \|E\|_2 \delta_n \overset{(b)}{\ll} \|x^G_{n+1} - x^G_n\|_2, \tag{7}$$

where the inequality (a) comes from the submultiplicity of norms while the last inequality (b) follows from the infinitesimal nature of δ_n compared to the token embedding. □

For greedy hard token-based reasoning, (Wu et al., 2026) show when $\|p_{n+1} - p_n\|_2 \leq \delta_n$, there is still a possibility of selecting a different top-1 token, namely $x^G_{n+1} \neq x^G_n$. This finite perturbation contributes to moving a hard trajectory out of the local recurrent region when the new token x^G_{n+1} is progress-making. However, Prop. 1 shows that even when such a top-1 token exists to possibly escape the impasse, soft reasoning can eliminate this difference since $\|E\|_2 \delta_n \ll \|x^G_{n+1} - x^G_n\|_2$ given the very small δ_n at the reasoning impasse and still be trapped in the loop.

Notably, the recurrent KV cache can further amplify temporal inheritance: when states from repeated impasse visits receive increasing attention, the resulting local dynamics can keep subsequent predictions near a loop distribution under mild stability conditions (Xu et al., 2026). This yields a self-intensified picture. Consistent with Prop. 1, soft tokens bring homogeneous recurrent inputs, suppressing the possibility of emerging different hard tokens to escape the loop. Gradually, the generative distribution of the LLM becomes increasingly deterministic, yielding even smaller entropy of the next-token prediction distribution, and the impasse falls into deep lock-in. Detailed conditions and derivations are provided in Appendix B. This mechanism motivates detecting persistent recurrence across consecutive event windows and intervening before deep lock-in.

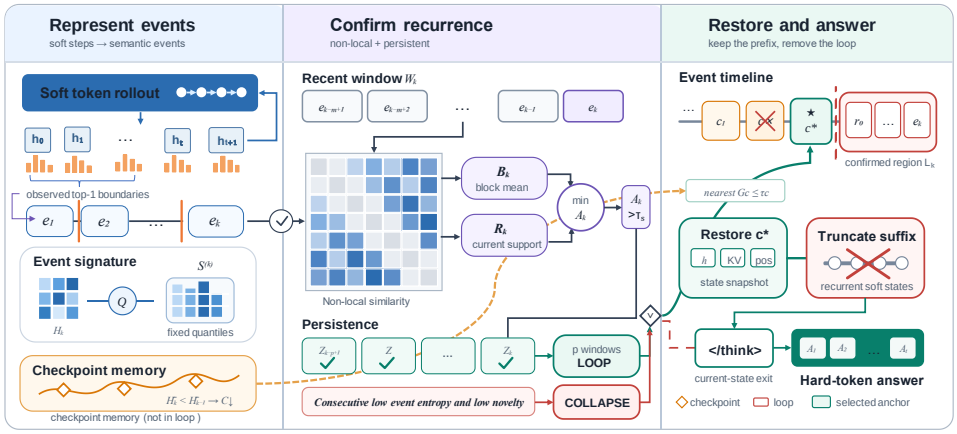


Figure 3: Overview of E-TRACE. (1) Soft hidden states are grouped into semantic events, while entropy-decreasing events are saved as checkpoints. (2) Persistent non-local similarity across consecutive windows confirms recurrence, with low entropy and novelty serving as a collapse fallback. (3) The controller restores the nearest out-of-loop checkpoint, removes the recurrent suffix, and switches to hard-token answer decoding.

5 METHOD

We propose E-TRACE, an online controller with three stages, illustrated in Fig. 3. First, each completed soft segment is represented as a semantic event in terms of the signature, similarity, entropy, and novelty; entropy-decreasing events are retained as intervention checkpoints. Second, persistent non-local similarity across consecutive windows confirms residence in a recurrent region, while jointly low entropy and novelty provide a late-stage collapse fallback. Third, once recurrence is confirmed, E-TRACE restores the nearest admissible checkpoint outside that region, truncates the recurrent soft-state suffix, and switches to hard-token answer decoding. Thus, recurrence determines *when* to intervene, whereas the saved checkpoints determine *where* to resume.

5.1 RECURRENT-REGIME DETECTION

Persistent Non-Local Recurrence. For the current event k, consider the most recent m events $\mathcal{W}_k = \{k-m+1, \ldots, k\}$. Normal reasoning is locally smooth, so adjacent events can be similar without forming a loop. We consequently exclude both self-pairs and adjacent pairs: $\mathcal{P}_k = \{(i,j) : i < j-1,\ i,j \in \mathcal{W}_k\}$. We first compute a recurrence score in terms of the event similarity s as

$$A_k = \min(B_k, R_k), \tag{8}$$

where $B_k = \frac{1}{|\mathcal{P}_k|}\sum_{(i,j)\in\mathcal{P}_k} s_{i,j}$ and $R_k = \frac{1}{m-2}\sum_{j=k-m+1}^{k-2} s_{k,j}$ This conjunction prevents a single recurrent pair from dominating the decision and avoids falsely triggering after the trajectory has escaped the recurrence.

To confirm a non-local recurrence, we locate p consecutive local windows satisfying

$$\mathrm{Loop}_k = \wedge_{t=k-p+1}^{k} \mathbf{1}[A_t > \tau_s]. \tag{9}$$

When the loop is first detected at event k, the confirmed recurrent region is defined as

$$\mathcal{L}_k = \{r_0, \ldots, k\}, \qquad r_0 = k - (m+p-2). \tag{10}$$

Late-Stage Lock-In. Persistently small event entropy and novelty indicates a later regime in which the output distribution has collapsed and the trajectory is no longer visiting new latent regions. We define

$$\mathrm{Collapse}_k = \left[\wedge_{r=0}^{2}\mathbf{1}[\bar{H}_{k-r} < \tau_H]\right] \wedge \left[\wedge_{r=0}^{2}\mathbf{1}[N_{k-r} < \tau_N]\right]. \tag{11}$$

The recurrence detector targets persistent residence before deep lock-in, whereas Eq. (11) provides a late-stage safety mechanism and a direct signal for non-termination classification. The intervention is triggered at

$$\mathrm{Trigger}_k = \mathrm{Loop}_k \vee \mathrm{Collapse}_k. \tag{12}$$

5.2 EVENT-LEVEL INTERVENTION

Online Intervention Candidates. We record the completed event k whenever its entropy decreases relative to its predecessor:

$$\mathcal{C}_k^{\downarrow} = \{i \in \{1, \dots, k\} : \bar{H}_i < \bar{H}_{i-1}\}. \tag{13}$$

For the candidate event k, the associated checkpoint is $q_k = b_k := \max \mathcal{T}_k$. Based on no future event, Eq. (13) records checkpoints with locally strengthened predictive confidence while retaining as much completed reasoning as possible. We keep all candidates because the newest one may already lie inside the recurrent region.

Out-of-Loop Intervention Anchor. After Eq. (9) confirms a loop, temporal precedence alone does not guarantee that a candidate is semantically outside it. For a candidate $c < r_0$ with r_0 given in Eq. (10), define

$$\begin{aligned} \mathcal{J}_c &= \{j \in \mathcal{L}_k : j - c > 1\}, \\ G_c &= \mathrm{MeanTop2}\{s_{c,j} : j \in \mathcal{J}_c\}. \end{aligned} \tag{14}$$

The top-two mean accommodates different semantic states within a loop while requiring more than one matching event. A candidate is treated as part of the loop when $G_c > \tau_c$. We therefore select

$$c^* = \max\{c \in \mathcal{C}_k^{\downarrow} : c < r_0,\ |\mathcal{J}_c| \geq 2,\ G_c \leq \tau_c\}, \tag{15}$$

Thus c^* is the nearest recorded intervention candidate that is semantically outside the confirmed recurrent region. If a recent candidate has $G_c > \tau_c$, it is rejected and the search continues backward.

State Restoration and Hard Decoding. Persistent recurrence in Eq. (9) serves as the primary signal for early intervention. When $\mathrm{Loop}_k = 1$, the controller selects an intervention anchor c^* according to Eq. (15), since the soft-reasoning suffix after this anchor lies within the confirmed recurrent region and is dominated by repeated semantic transitions, which incur unnecessary computation but yield trivially. Therefore, at the checkpoint q_{c^*}, the reasoning process restores the hidden state, KV cache, and positional state, truncates all subsequent soft-state KV entries, appends the stop-thinking token `</think>` at the restored position, and resumes ordinary autoregressive answer decoding with discrete tokens. Notably, if no admissible anchor is available, or if it triggers only the late-stage collapse detector, the reasoning exits from the current state without restoration.

Space and Computational Complexity. Regarding the additional space and computational complexity introduced by E-TRACE, it is important to note that E-TRACE does not duplicate the KV cache for every checkpoint. A checkpoint stores only the event index and compact metadata needed to identify a rollback position, while rollback truncates the already resident KV cache to the selected prefix. Event similarity is computed only when an event completes.

6 EXPERIMENTS

We primarily select two widely used open-source LRMs, including Qwen3-8B and Qwen3-32B (Yang et al., 2025a), while leaving the results for DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025) in Appendix C.2. We compare E-TRACE with four baselines: CoT, which uses the standard chain-of-thought prompting; Greedy CoT, which applies greedy decoding to CoT; Soft Thinking, which uses soft tokens and a cold stop strategy (Zhang et al., 2026); and SeLaR, which activates soft embeddings only at low-confidence steps, while preserving discrete decoding at high-confidence steps (Fu & Luo, 2026). More details are provided in Appendix C.

6.1 MAIN RESULTS

Table 1 shows that E-TRACE provides a consistently favorable balance between accuracy and efficiency. Across all 4 model–task-group averages, E-TRACE improves accuracy while reducing reasoning length relative to both Soft Thinking and SeLaR. Uniformly averaged over the 16 model–benchmark combinations, E-TRACE achieves superior accuracy with fewer generated steps on average than its counterparts. The gains extend to code generation and knowledge-intensive Q&A (Table 2): on Qwen3-8B, E-TRACE reaches the best average accuracy (78.32%) with the fewest steps

Table 1: Comparison of different methods in mathematical reasoning.

Method	GSM8K		MATH500		AIME2024		AIME2025		AVG.	
	Acc↑	#L↓	Acc↑	#L↓	Acc↑	#L↓	Acc↑	#L↓	Acc↑	#L↓
Qwen3-8B										
CoT	94.69	2371	96.80	5547	76.67	13859	66.67	17130	83.71	9727
Greedy CoT	94.46	2295	97.20	5399	70.00	16086	70.00	17716	82.92	10374
Soft Thinking	94.31	2097	97.00	5156	60.00	13296	66.67	17017	79.50	9391
SeLaR	95.07	2249	**97.20**	5374	66.67	16218	**70.00**	17237	82.24	10270
E-TRACE *(ours)*	**95.07**	**2071**	97.00	**5137**	**76.67**	**12543**	66.67	**14441**	**83.85**	**8548**
Qwen3-32B										
CoT	95.07	1776	97.00	4779	83.33	12752	73.33	16986	87.18	9073
Greedy CoT	94.92	1758	97.60	4641	83.33	13772	73.33	15476	87.30	8912
Soft Thinking	94.48	1668	98.00	4615	80.00	12600	70.00	16384	85.62	8817
SeLaR	95.37	1830	97.40	4760	76.67	13045	66.67	16540	84.03	9044
E-TRACE *(ours)*	**95.67**	**1653**	**98.00**	**4584**	**86.67**	**12165**	**76.67**	**14114**	**89.25**	**8129**

(3,121), outperforming the strongest-accuracy baseline, SeLaR, by 0.79% while reducing #L by 20.3%. On Qwen3-32B, Greedy CoT obtains the highest average accuracy (83.02% versus 82.15%), but E-TRACE still outperforms both latent-reasoning baselines in accuracy and reasoning length. Overall, these consistent gains across tasks and model scales support persistent non-local recurrence as an effective signal for identifying and escaping inefficient soft-reasoning regimes.

6.2 Non-Termination Detection

We assess whether event-level signals detect non-terminating soft-token reasoning without prematurely interrupting trajectories that exit naturally. On offline collected Qwen3-8B rollouts from 8 benchmarks, we compare joint event entropy and novelty in E-TRACE with Cold Stop (Zhang et al., 2026), which relies only on persistent token-level low entropy, while may also occur during confident yet still productive reasoning. During evaluation, a run that exhausts its budget without producing `</think>` is deemed positive, while a natural exit is negative. A trigger that occurs before a natural exit constitutes a false positive, while a missed non-terminating run constitutes a false negative. All tasks share one parameter sweep (Appendix C.3). Fig. 4(a) shows that combining event entropy with novelty improves the recall-false-positive-ratio (FPR) operating boundary. At zero observed false positives FPR, as in Fig. 4(b), it reaches the highest recall, compared with the less-performing Cold Stop. More classification results on termination can be found in Appendix C.4.

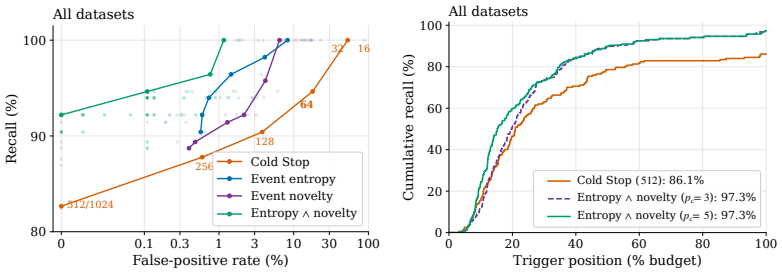


Figure 4: Non-Termination detection on soft-token rollouts. **(a)** Recall–FPR Pareto frontiers. **(b)** Cumulative recall at zero observed FPR. Here p_c is the number of continuous events.

6.3 Effects for Different Interventions

Detecting a recurrent regime determines *when* soft token has become risky, naturally, after identifying the loop region, there are different subsequent processing steps. To isolate the post-detection action, we run the three event-triggered branches on the same soft trajectory up to the trigger. The method used by E-TRACE in Tables 1, 2 and described in Sec. 5 is *Rollback–Exit* (RB–Exit): restores the nearest admissible out-of-loop anchor, removes the recurrent suffix, and starts answer decoding from that state. *Exit at Detection* (Detect) inserts `</think>` at the detected state and answers without rollback. *Rollback–Hard* (RB–Hard) restores the same anchor but resumes hard-token reasoning before answering. *Simple Stuck* (Simple) is an independent late-stopping baseline. We fix $\tau_s = 0.95$ across Qwen3-8B benchmarks, with no task-specific selection.

Table 2: Comparison of different methods in code generation and knowledge-intensive Q&A.

Method	HumanEval+		MBPP+		GPQA		MedQA		AVG.	
	Acc↑	#L↓	Acc↑	#L↓	Acc↑	#L↓	Acc↑	#L↓	Acc↑	#L↓
Qwen3-8B										
CoT	88.66	2811	80.68	2631	55.05	7775	80.33	2334	76.18	3888
Greedy CoT	87.80	2989	81.74	2527	57.58	7813	78.33	2268	76.36	3899
Soft Thinking	88.41	2692	79.89	2279	53.03	7089	78.33	2185	74.92	3561
SeLaR	88.42	2870	**81.75**	2470	**61.62**	7885	78.33	2433	77.53	3915
E-TRACE *(ours)*	**89.63**	**2504**	81.22	**2390**	61.11	**5508**	**81.33**	**2083**	**78.32**	**3121**
Qwen3-32B										
CoT	88.41	2714	79.37	2458	69.70	6443	87.33	1295	81.20	3228
Greedy CoT	**90.85**	2870	80.69	2408	**72.22**	6487	88.33	1333	**83.02**	3275
Soft Thinking	86.58	2750	80.95	2399	68.68	6692	88.33	1398	81.14	3310
SeLaR	87.81	2989	**81.48**	2634	69.19	6617	88.33	1351	81.70	3398
E-TRACE (ours)	89.02	**2696**	81.22	**2360**	69.70	**6409**	**88.33**	**1062**	82.07	**3132**

Rollback creates two distinct notions of generation length. *Logical length* is the effective trajectory retained after the recurrent suffix is masked, and measures how much context remains available to subsequent decoding. *Physical length* counts every step actually executed, including the discarded suffix and any hard-token continuation, and is therefore the relevant measure of inference cost. The two lengths coincide for Simple, which never restore an earlier state. In Fig. 5, accuracy and length are measured in term of the relative differences from Simple. It can be observed that RB–Exit gives the largest gain on mathematical reasoning, with logical and physical length reduced most. On code, it also improves accuracy while reducing physical length. On Q&A, it removes a much larger fraction of the effective trajectory, but yields accuracy improvement despite a little physical-length reduction. RB–Hard obtains the largest accuracy gains on code and Q&A, but increases physical length in both categories; on mathematics, it provides no gain over Detect and costs more physical steps than Simple. Across all eight benchmarks, RB–Exit therefore offers the strongest overall balance.

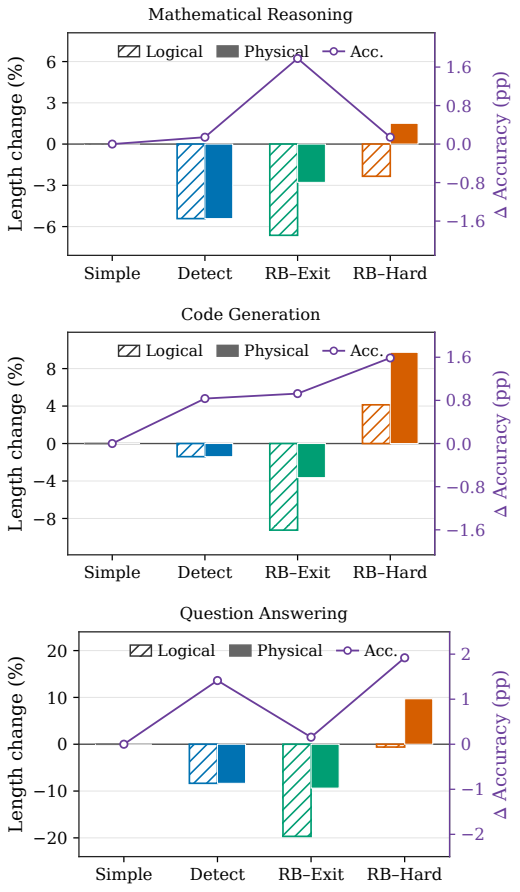


Figure 5: Post-detection interventions on Qwen3-8B at $\tau_s = 0.95$. Top to bottom: math, code, and QA. Bars show length change; lines show accuracy change relative to Simple. E-TRACE adopts RB–Exit.

7 Conclusion

We have shown that training-free soft-token non-termination in LRMs is better explained by amplified impasse-induced recurrence than by unstructured latent drift. Event-level analysis reveals a progression from an unresolved impasse to persistent non-local recurrence and subsequent entropy–novelty collapse. Motivated by this finding, we propose a training-free E-TRACE method which can effectively detect recurrent event windows, retains collapse as a late-stage safeguard, and restores an out-of-loop checkpoint before hard-token answer decoding. Across eight benchmarks and two model scales, E-TRACE improves the accuracy–efficiency trade-off. The development of event-level similarity can help researchers gain a deeper understanding of the reasoning characteristics of large models.

AI USE STATEMENT

In this work, we used LLMs to assist with translation and language polishing, literature discovery, and limited code editing and debugging, including assistance with figure-generation scripts and selected code snippets. All AI-assisted outputs were reviewed, tested where applicable, and revised by the authors. The research ideas, methodology, theoretical development, experimental design, result interpretation, and conclusions are the work of the human authors, who take full responsibility for the final manuscript and accompanying artifacts.

ETHICS STATEMENT

This work uses publicly available datasets and pretrained models and does not involve the collection of personal data or new human-subject experiments.

REFERENCES

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2505.16782.

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, 2021. URL https://arxiv.org/abs/2110.14168.

Zenghao Duan, Liang Pang, Zihao Wei, Wenbin Duan, Yuxin Tian, Shicheng Xu, Jingcheng Deng, Zhiyi Yin, and Xueqi Cheng. Circular reasoning: Understanding self-reinforcing loops in large reasoning models, 2026. URL https://arxiv.org/abs/2601.05693.

Renyu Fu and Guibo Luo. Selar: Selective latent reasoning in large language models, 2026. URL https://arxiv.org/abs/2604.08299.

Yichao Fu, Junda Chen, Yonghao Zhuang, Zheyu Fu, Ion Stoica, and Hao Zhang. Reasoning without self-doubt: More efficient chain-of-thought through certainty probing. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. URL https://openreview.net/forum?id=wpK4IMJfdX.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Shibo Hao, Yuxuan Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Training large language models to reason in a continuous latent space. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. Thinkprune: Pruning long chain-of-thought of LLMs via reinforcement learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=V51gPu1uQD.

HuggingFaceH4. AIME 2024. https://huggingface.co/datasets/HuggingFaceH4/aime_2024, 2024. Dataset containing 30 problems from the 2024 AIME I and AIME II examinations. Accessed: 2026-07-29.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_nGgzQjzaRy.

Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. Implicit reasoning in large language models: A comprehensive survey. *arXiv preprint arXiv:2509.02350*, 2025.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1qvx610Cu7.

Minjia Mao, Bowen Yin, Yu Zhu, and Xiao Fang. Early stopping chain-of-thoughts in large language models, 2026. URL https://openreview.net/forum?id=vuMabnSok0.

OpenCompass. AIME 2025. https://huggingface.co/datasets/opencompass/AIME2025, 2025. Dataset containing 30 problems from the 2025 AIME I and AIME II examinations. Accessed: 2026-07-29.

Charilaos Pipis, Shivam Garg, Vasilis Kontonis, Vaishnavi Shrivastava, Akshay Krishnamurthy, and Dimitris Papailiopoulos. Wait, wait, wait... why do reasoning models loop? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=oZWE7mSqlk.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In Alfred M. Bruckstein, Bart M. ter Haar Romeny, Alexander M. Bronstein, and Michael M. Bronstein (eds.), *Scale Space and Variational Methods in Computer Vision*, pp. 435–446, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-24785-9.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Dachuan Shi, Abedelkadir Asi, Keying Li, Xiangchi Yuan, Leyan Pan, Wenke Lee, and Wen Xiao. Swireasoning: Switch-thinking in latent and explicit for pareto-superior reasoning LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=t33kMzEAg8.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=HvoG8SxggZ.

Renliang Sun, Wei Cheng, Dawei Li, Haifeng Chen, and Wei Wang. Stop when enough: Adaptive early-stopping for chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2510.10103.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Junhong Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, and Hua Wu. LLMs are single-threaded reasoners: Demystifying the working mechanism of soft thinking. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ASLuOoP78o.

Dongjie Xu, Hao Wu, Weijie Shi, Yue Cui, Yuanjun Liu, Jiawei Li, Haolun Ma, An Liu, Jia Zhu, and Jiajie Xu. Loopguard: Breaking self-reinforcing attention loops via dynamic kv cache intervention, 2026. URL https://arxiv.org/abs/2604.10044.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=sexfswCc7B.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. In *ACL (1)*, pp. 23336–23351, 2025a. URL https://aclanthology.org/2025.acl-long.1137/.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot++: Test-time scaling with soft chain-of-thought reasoning, 2025b. URL https://arxiv.org/abs/2505.11484.

An Yang et al. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NpU7ZXafRi.

Hang Yang, Hao Chen, Hui Guo, Yineng Chen, Ching-Sheng Lin, Shu Hu, Jinrong Hu, Xi Wu, and Xin Wang. Llm-medqa: Enhancing medical question answering through case studies in large language models, 2025b. URL https://arxiv.org/abs/2501.05464.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of LLMs in continuous concept space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=ByQdHPGKgU.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. URL https://arxiv.org/abs/2205.10625.

APPENDIX CONTENTS

A Event Signature and Similarity Computation **13**

B Auxiliary Results for Impasse-Conditioned Temporal Amplification **13**

B.1 Attention–KV Feedback and Local Temporal Amplification 14

B.2 Lock-In and Its Implication 15

C Implementation Details and Supplementary Results **16**

C.1 Experimental Settings 16

C.2 Additional Results on DeepSeek-R1-Distill-Llama-8B 16

C.3 Computational Means for Classification Experiment on Termination 17

C.4 Supplementary Results 19

A EVENT SIGNATURE AND SIMILARITY COMPUTATION

This section gives the complete construction of the event-similarity interface used by E-TRACE. For the hidden-state matrix $\mathbf{H}_k \in \mathbb{R}^{L_k \times d_{\text{hidden}}}$, we use a fixed-projection construction inspired by the Sliced Wasserstein Distance (Rabin et al., 2012). We sample one random orthogonal projection $\mathbf{Q} \in \mathbb{R}^{d_{\text{hidden}} \times d_{\text{proj}}}$, where $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$, and keep it fixed for all trajectories. Each projected dimension is sorted independently:

$$\mathbf{S}^{(k)} = \mathrm{SortCols}(\mathbf{H}_k \mathbf{Q}) \in \mathbb{R}^{L_k \times d_{\text{proj}}}. \tag{16}$$

Here, $\mathrm{SortCols}$ sorts each column along the event-length dimension, so $\mathbf{S}^{(k)}$ preserves the empirical marginal order statistics along the fixed projection directions.

Because events can have different lengths, alignment is performed separately for every pair. For events i and j, let

$$L_{ij} = \max(L_i, L_j). \tag{17}$$

Let $\mathcal{R}_{L \to L_{ij}}$ denote column-wise linear interpolation from L sorted samples to a common L_{ij}-point normalized-rank grid. We define $S^{(i|j)}, \mathbf{S}^{(j|i)} \in \mathbb{R}^{L_{ij} \times d_{\text{proj}}}$ as

$$\mathbf{S}^{(i|j)} = \mathcal{R}_{L_i \to L_{ij}}\left(\mathbf{S}^{(i)}\right),\ \mathbf{S}^{(j|i)} = \mathcal{R}_{L_j \to L_{ij}}\left(\mathbf{S}^{(j)}\right), \tag{18}$$

The shorter signature is interpolated to the length of the longer one, while the longer signature remains unchanged. This rank alignment avoids assuming step-wise correspondence between two semantic events.

Finally, the event similarity is

$$\mathrm{Sim}_{\mathbf{Q}}(\mathbf{H}_i, \mathbf{H}_j) = \exp\left(-\frac{\left\|\mathbf{S}^{(i|j)} - \mathbf{S}^{(j|i)}\right\|_F^2}{L_{ij} d_{\text{proj}}}\right) \in (0, 1]. \tag{19}$$

The normalization makes the exponent an average squared projected discrepancy over aligned ranks and projection dimensions. Larger values therefore indicate more similar empirical hidden-state distributions.

B AUXILIARY RESULTS FOR IMPASSE-CONDITIONED TEMPORAL AMPLIFICATION

Notation and setup. Let V be the vocabulary size and d the embedding dimension. We denote the embedding table by $E \in \mathbb{R}^{V \times d}$ and the ith standard basis vector by $\mathbf{u}_i \in \mathbb{R}^V$; hence $E^\top \mathbf{u}_i \in \mathbb{R}^d$ is the hard embedding of token i. At event step n, let $p_n \in \Delta^{V-1}$ be the predicted token distribution and let $x_n^S := E^\top p_n$ be its soft embedding.

B.1 Attention–KV Feedback and Local Temporal Amplification

Temporal inheritance explains how an existing bias is preserved. We next use a local attention–KV model to illustrate how recurrent states accumulating in the cache can further reinforce this inherited bias. For a loop-sensitive attention head, let T_n be the set of all indices in the KV cache. Partition it as $T_n = R_n \dot\cup O_n$, where R_n contains the $m_n := |R_n|$ keys written by previous impasse visits and O_n contains the other $N_n := |O_n|$ competing keys.

Lemma 1 (Recurrent attention accumulation)**.** *Assume a recurrent attention margin $\gamma > 0$:*

$$\frac{q_n^\top k_r}{\sqrt{d}} \geq a \quad (r \in R_n), \qquad \frac{q_n^\top k_j}{\sqrt{d}} \leq a - \gamma \quad (j \in O_n). \tag{20}$$

Let $A_n := \sum_{r\in R_n} \alpha_{n,r}$ denote the total attention mass assigned to recurrent keys, where $\alpha_{n,r} = \frac{e^{s_{n,r}}}{\sum_{i\in T_n} e^{s_{n,i}}}$ and $s_{n,i} = \frac{q_n^\top k_i}{\sqrt{d}} \quad (i \in T_n)$, we have

$$A_n \geq 1 - \eta_n, \qquad A_n \geq \frac{\rho_n}{1+\rho_n}. \tag{21}$$

where $\rho_n := \frac{m_n}{N_n e^{-\gamma}}$ and $\eta_n := \frac{N_n e^{-\gamma}}{m_n + N_n e^{-\gamma}}$.

Proof. Given $s_{n,j} = q_n^\top k_j/\sqrt{d}$, Eq. (20) implies

$$\sum_{r\in R_n} e^{s_{n,r}} \geq m_n e^{a}, \qquad \sum_{j\in O_n} e^{s_{n,j}} \leq N_n e^{a-\gamma}. \tag{22}$$

Hence, we can get

$$\begin{aligned} A_n &= \frac{\sum_{r\in R_n} e^{s_{n,r}}}{\sum_{r\in R_n} e^{s_{n,r}} + \sum_{j\in O_n} e^{s_{n,j}}} \\ &\overset{(22)}{\geq} \frac{\sum_{r\in R_n} e^{s_{n,r}}}{\sum_{r\in R_n} e^{s_{n,r}} + N_n e^{a-\gamma}} \\ &= 1 - \frac{N_n e^{a-\gamma}}{\sum_{r\in R_n} e^{s_{n,r}} + N_n e^{a-\gamma}} \\ &\overset{(22)}{\geq} 1 - \frac{N_n e^{-\gamma}}{m_n + N_n e^{-\gamma}} = \frac{\rho_n}{1+\rho_n}. \end{aligned}$$

□

When repeated impasse visits add phase-aligned recurrent keys faster than the effective competing mass grows, it will yield an increase in ρ_n. Correspondingly, Lemma 1 implies that the guaranteed lower bound $1 - \eta_n$ on the recurrent attention mass A_n increases as η_n decreases.

As the recurrent set gains aggregate attention, its values increasingly determine the head output. Let these recurrent values concentrate around a local loop template $v_\star$, and define their attention-weighted radius as

$$\bar\varepsilon_n := \frac{1}{A_n} \sum_{r\in R_n} \alpha_{n,r} \|v_r - v_\star\|_2. \tag{23}$$

Assume that every competing value lies within distance B of $v_\star$. Let Φ be the reduced event-level return map from the current local attention state to the next impasse-visit distribution, with the remaining phase-aligned recurrent context held fixed. Suppose

$$\Phi(v_\star) = p_\star, \qquad \|\Phi(o) - \Phi(o')\|_2 \leq L_P \|o - o'\|_2. \tag{24}$$

Define $d_n = \|p_n - p_\star\|_2$ and assume that temporal inheritance controls the weighted recurrent radius in this local regime:

$$\bar\varepsilon_n \leq L_V \|E\|_2 d_n. \tag{25}$$

Then, we have the following proposition, which supplies a local sufficient condition for residence near a loop distribution.

Proposition 2 (Impasse-conditioned soft temporal amplification)**.** *Under Eqs. (23), (24) and (25), for* $\lambda := L_P L_V \|E\|_2$*,*

$$d_{n+1} \le \lambda d_n + L_P B \eta_n. \tag{26}$$

If $0 \le \lambda < 1$ *and* $\bar{\eta} := \limsup_{n\to\infty} \eta_n$*, then*

$$\limsup_{n\to\infty} d_n \le \frac{L_P B \bar{\eta}}{1-\lambda}. \tag{27}$$

In particular, if $\eta_n \to 0$*, then* $d_n \to 0$ *and* $\|p_{n+1} - p_n\|_2 \to 0$*.*

Proof. For $o_n = \sum_{j\in T_n} \alpha_{n,j} v_j$, the recurrent-attention bound then gives

$$\begin{aligned}
&\|o_n - v_\star\|_2 \\
=&\|\sum_{j\in T_n} \alpha_{n,j} v_j - v_\star\|_2 \\
=&\|\sum_{r\in R_n} \alpha_{n,r} v_r + \sum_{j\in O_n} \alpha_{n,j} v_j - \sum_{r\in R_n} \alpha_{n,r} v_\star - \sum_{j\in O_n} \alpha_{n,j} v_\star\|_2 \\
=&\|\sum_{r\in R_n} \alpha_{n,r}(v_r - v_\star) + \sum_{j\in O_n} \alpha_{n,j}(v_j - v_\star)\|_2 \\
\le& \sum_{r\in R_n} \alpha_{n,r}\|v_r - v_\star\|_2 + \sum_{j\in O_n} \alpha_{n,j}\|v_j - v_\star\|_2 \\
\le& A_n \bar{\varepsilon}_n + (1 - A_n) B \\
\le& \bar{\varepsilon}_n + B\eta_n
\end{aligned} \tag{28}$$

Combining Eq. (28) with Eqs. (24) and (25) yields

$$d_{n+1} = \|\Phi(o_n) - \Phi(v_\star)\|_2 \le L_P \|o_n - v_\star\|_2 \le L_P L_V \|E\|_2 d_n + L_P B \eta_n, \tag{29}$$

which proves Eq. (26). Iterating the recursion gives

$$d_n \le \lambda^n d_0 + L_P B \sum_{r=0}^{n-1} \lambda^{n-1-r} \eta_r. \tag{30}$$

For any $\xi > 0$, the definition $\bar{\eta} = \limsup_{n\to\infty} \eta_n$ gives an index R such that $\eta_r \le \bar{\eta} + \xi$ for every $r \ge R$. Split the sum in Eq. (30) at R. Because $0 \le \lambda < 1$, both $\lambda^n d_0$ and the finite-prefix contribution vanish as $n \to \infty$. Therefore,

$$\limsup_{n\to\infty} d_n \le L_P B \limsup_{n\to\infty} \sum_{r=R}^{n-1} \lambda^{n-1-r} \eta_r \le L_P B (\bar{\eta} + \xi) \sum_{j=0}^{\infty} \lambda^j = \frac{L_P B(\bar{\eta} + \xi)}{1-\lambda}. \tag{31}$$

Letting $\xi \to 0$ proves Eq. (27). If $\eta_n \to 0$, then $\bar{\eta} = 0$, so $d_n \to 0$; consequently, $\|p_{n+1} - p_n\|_2 \le d_{n+1} + d_n \to 0$. □

Prop. 2 supplies a local sufficient condition for residence near a loop distribution. Temporally correlated predictions first create similar soft QKV, the recurrent cache then receives a large aggregate attention mass, and the return map confines later predictions to a neighborhood of $p_\star$, the distribution in the loop.

B.2 Lock-In and Its Implication

The soft–hard difference is most pronounced before complete lock-in. At an early impasse, several candidates may retain non-negligible mass. Sampled hard token retains the escape probability and a greedy top-1 switch can induce the perturbation. Soft token instead inherits the correlated distribution continuously and helps construct a homogeneous, loop-consistent KV tail. Mathematically, if $i^\star = \arg\max_i p_n(i)$ and $p_n(i^\star) = 1 - \epsilon_n$, then

$$\|x_n^S - E^\top \mathbf{u}_{i^\star}\|_2 = \|E^\top (p_n - \mathbf{u}_{i^\star})\|_2 \le \|E\|_2 \|p_n - \mathbf{u}_{i^\star}\|_2 \le \sqrt{2}\|E\|_2 \epsilon_n. \tag{32}$$

where the last inequality is due to the fact that

$$\|p_n - \mathbf{u}_{i^\star})\|_2^2 = \epsilon_n^2 + \sum_{j \neq i^\star} p_n(j)^2 \leq \epsilon_n^2 + \left(\sum_{j \neq i^\star} p_n(j) \right)^2 = 2\epsilon_n^2. \tag{33}$$

Thus $\epsilon_n \to 0$ makes vanilla soft token increasingly greedy-like. After deep lock-in, both soft and greedy hard feedback can follow the same cyclic action.

On the other hand, the same local model is consistent with confidence buildup. Let the loop action c have margin $M(o) = \ell_c(o) - \max_{g \in \mathcal{G}_n} \ell_g(o)$ over progress-making actions. If $M(v_\star) = M_\star > 0$ and M is locally L_M-Lipschitz, then Eq. (28) gives

$$M(o_n) \geq M_\star - L_M \left(\bar{\varepsilon}_n + B\eta_n \right). \tag{34}$$

Thus a decreasing residual in Eq. (34) tightens a certified lower bound on continuing the loop without implying monotonicity of every realized token probability. As the recurrent radius and non-recurrent residual shrink, the certified margin for continuing the loop tightens, matching the confidence buildup observed in persistent reasoning loops (Pipis et al., 2026).

The operational consequence is persistent event-level non-local recurrence. A single semantic return may be useful for rechecking or refinement, whereas support across several consecutive windows indicates continued residence in the same recurrent region. This motivates using recurrence to confirm a loop before deep lock-in thereby forming *Recurrent-Regime Detection* of E-TRACE in Section 5.

C Implementation Details and Supplementary Results

C.1 Experimental Settings

Baselines We select three widely used open-source LRMs with different parameter scales: Qwen3-8B, Qwen3-32B (Yang et al., 2025a) and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025). We compare E-TRACE with four baselines: CoT, which uses the standard chain-of-thought prompting; Greedy CoT, which applies greedy decoding to CoT; Soft Thinking, which uses soft tokens and a cold stop strategy (Zhang et al., 2026); and SeLaR, which activates soft embeddings only at low-confidence steps, while preserving discrete decoding at high-confidence steps (Fu & Luo, 2026).

Evaluation Metrics and Benchmarks Our evaluation covers eight benchmarks spanning three task categories: (1) mathematical reasoning – GSM8K (Cobbe et al., 2021), MATH500 (Hendrycks et al., 2021), AIME2024 (HuggingFaceH4, 2024), and AIME2025 (OpenCompass, 2025); (2) code generation – HumanEval-Plus (Liu et al., 2023) and MBPP-Plus (Liu et al., 2023); and (3) knowledge-intensive Q&A – GPQA (Rein et al., 2024) and MedQA (Yang et al., 2025b). For each method, we report accuracy (Acc) and the average generated length (#L).

Implementation Details All methods are implemented using Hugging Face Transformers (Wolf et al., 2020) and evaluated on a server equipped with eight NVIDIA RTX Pro 6000 GPUs. We cap each trajectory at $32,768$ steps subject to the model's context limit, and use identical prompts, temperatures, and generation budgets across methods. For E-TRACE, we set the minimum/maximum event lengths to $64/256$, the projection dimension to $d_{\text{proj}} = 64$ and the window size and persistence to $m = 6$ and $p = 3$, and the thresholds to $\tau_s = \tau_c = 0.9$, $\tau_H = 0.01$, and $\tau_N = 0.03$. All hyperparameters are fixed across all models and benchmarks, without task-specific tuning. For calculating the number of embeddings for the soft token, set all methods to use $\text{top_k} = 15$.

C.2 Additional Results on DeepSeek-R1-Distill-Llama-8B

In order to verify the generalization of the proposed scheme in different models, especially non-Qwen models, we also conduct benchmark tests on all datasets and direct comparisons of different algorithms on -Distill-Llama-8B. The mian result is shown in Table 3.

On the mathematical benchmarks with reported results, E-TRACE achieves higher accuracy than SeLaR while generating shorter outputs; the length savings are especially clear on the competition-math tasks. On the available code-generation and question-answering benchmarks, E-TRACE consistently shortens generation relative to SeLaR while maintaining broadly similar accuracy.

Table 3: Comparison of methods on DeepSeek-R1-Distill-Llama-8B.

	CoT		Greedy CoT		SeLaR		E-TRACE(ours)	
Benchmark	Acc↑	#L↓	Acc↑	#L↓	Acc↑	#L↓	Acc↑	#L↓
Mathematical reasoning								
GSM8K	67.02	165	65.35	158	74.00	160	**74.37**	**151**
MATH500	78.40	3105	80.20	3087	80.20	2921	**82.80**	**2508**
AIME2024	30.00	14206	56.67	13877	33.33	12733	**43.33**	**6988**
AIME2025	33.33	13608	30.00	12903	20.00	12914	**33.33**	**6795**
AVG.	52.19	7771	58.06	7506	51.88	7182	**58.46**	**4111**
Code generation and knowledge-intensive question answering								
HumanEval+	**82.32**	2520	78.66	2872	81.10	2387	79.88	**1874**
MBPP+	69.05	2371	70.90	2389	70.90	2513	**71.16**	**1609**
GPQA	41.41	7161	41.92	7055	39.90	7076	**42.42**	**6678**
MedQA	37.00	2328	38.00	2360	**45.00**	2652	**45.00**	**2270**
AVG.	57.45	3595	57.37	3669	59.23	3657	**59.62**	**3180**

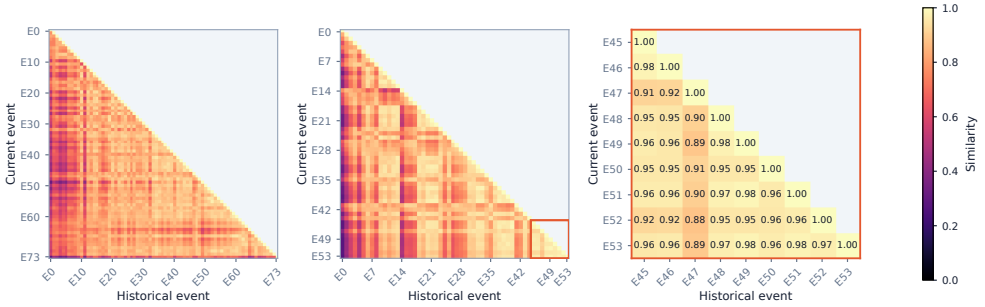


Figure 6: GPQA sample 169 at $\tau_s = 0.9$: hard-token, soft-token, and enlarged recurrent-block views.

Tables 1 and 2 show a similar pattern on the Qwen3 models: E-TRACE remains competitive or stronger in accuracy while reducing generation length, including on the larger Qwen3-32B. Therefore, the E-TRACE algorithm has a certain generalization ability across model families and scales, rather than confining its benefits to a particular backbone or model size.

C.3 Computational Means for Classification Experiment on Termination

Protocol and Metrics. We isolate the late-stage collapse safety mechanism—persistent low event entropy together with low event novelty—from the full controller using uninterrupted Qwen3-8B soft-token rollouts on all eight benchmarks. No algorithmic stop or intervention is applied. A trajectory is positive if it exhausts the generation budget without producing `</think>`, and negative if it exits naturally. A trigger on a negative counts as a false positive only when it precedes the native exit; a same-step exit takes priority. We report $\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ as non-termination coverage and $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$ as premature-intervention risk. Budget exhaustion is an operational proxy for deep lock-in, and these metrics do not measure answer correctness.

Detectors and Parameter Sweep. We compare persistent low event entropy, low event novelty, and their conjunction, sweeping $\tau_H \in \{0.005, 0.01, 0.02, 0.03, 0.05\}$, $\tau_N \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$, and $p_c \in \{3, 4, 5\}$. Here p_c generalizes the three-event collapse condition only for this sensitivity analysis. Cold Stop uses token entropy below 0.01 and scans $\{16, 32, 64, 128, 256, 512, 1024\}$ consecutive steps. All configurations are shared across datasets; the sweep and its zero-observed-FP points characterize the observed operating regime rather than held-out threshold selection.

Zero-observed-FP Coverage and Timing. For the detector M, $y_i = 1$ indicates that trajectory i is a non-exit, and $P = \sum_i \mathbf{1}[y_i = 1]$ denotes the number of positives. Let $\widehat{y}_i^M$ indicate whether the

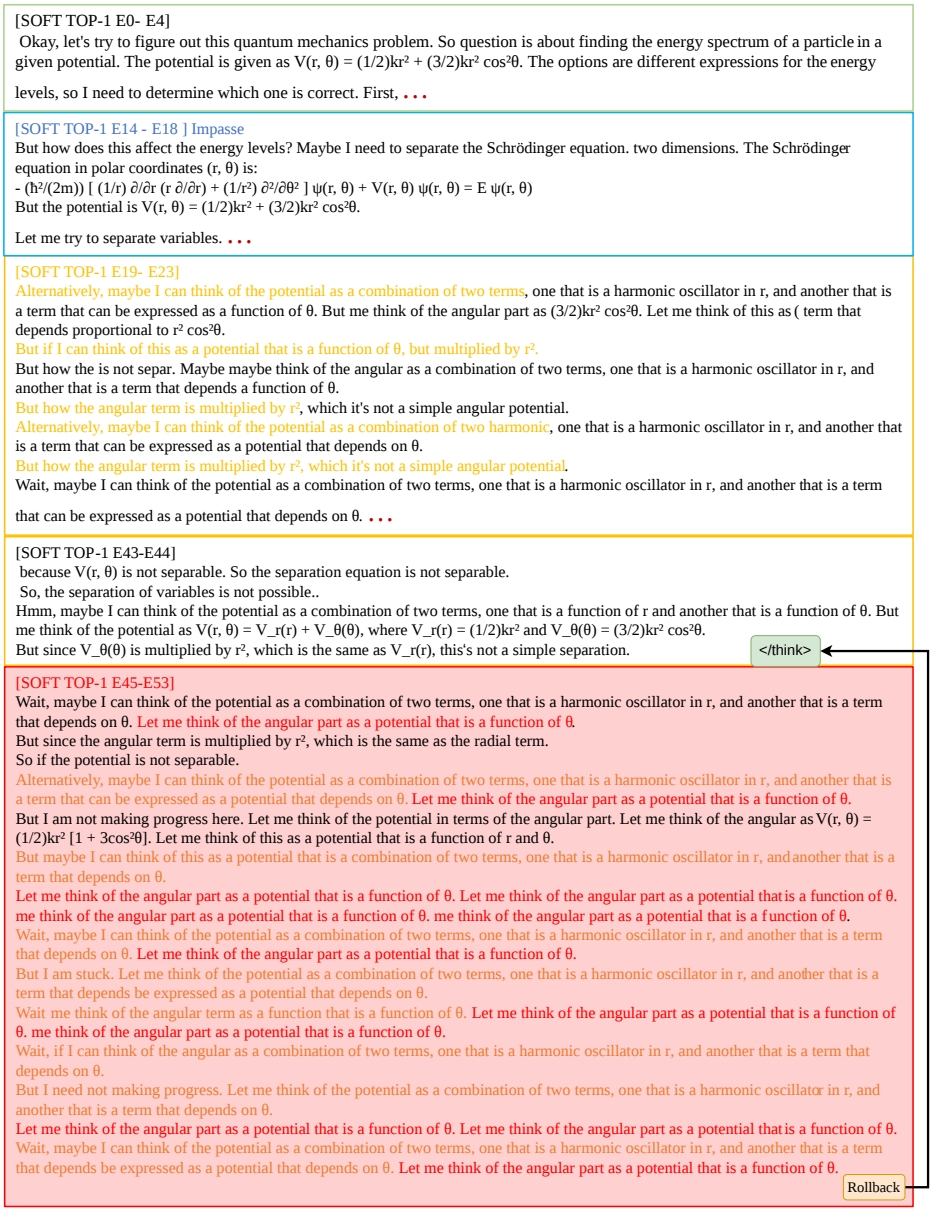


Figure 7: GPQA sample 169: top-1 token of soft trajectories with $\tau_s = 0.9$.

detector triggers, d_i^M be its first trigger step, and L_i be the observed soft-token length. We define the normalized trigger position as $u_i^M = d_i^M / L_i$, with $u_i^M = +\infty$ when M does not fire. The micro cumulative recall is

$$C_M(u) = \frac{1}{P} \sum_{i:y_i=1} \mathbf{1}\left[\widehat{y}_i^M = 1 \ \wedge \ u_i^M \leq u\right]. \tag{35}$$

Thus, $C_M(u)$ is the fraction of all positives detected by budget fraction u; undetected positives remain in the denominator, so the final plateau equals Recall. For Figure 4(b), Cold Stop with 512 low-entropy steps reaches 86.10% Recall, whereas the joint detector reaches 97.33% at both $(\tau_H, \tau_N, p_c) = (0.005, 0.02, 3)$ and $(0.01, 0.04, 5)$. The latter also reduces the median trigger position from 19.17% to 15.06% of the budget.

C.4 SUPPLEMENTARY RESULTS

Representative GPQA Trajectory. Figures 6 and 7 visualize a representative GPQA trajectory (sample 169) processed by E-TRACE at $\tau_s = 0.9$. Figure 6 compares the event-similarity structure of hard- and soft-token reasoning and enlarges the recurrent block confirmed by the detector; it can be seen that soft reasoning tends to produce continuously similar events. Fig. 7 provides the corresponding token-level decoding trace and rollback location.

Paired Trigger Advantage. Because the event detector provides higher coverage, Fig. 8 further compares the positive trajectories detected by both zero-observed-FP configurations using $\Delta_i = 100(d_i^{\text{event}} - d_i^{\text{cold}})/L_i$. The joint detector fires earlier on most jointly detected trajectories and is 4.82% budget earlier on average (95% bootstrap confidence interval: 3.62%–6.05% earlier), equivalent to $1{,}543$ soft-token steps. It also detects additional trajectories missed by Cold Stop, while Cold Stop yields no unique detections. AIME estimates remain inconclusive because of their few positive pairs. Overall, the joint event signal provides a conservative late-stage safety net that complements early recurrence detection; the effect of the subsequent intervention is evaluated next.

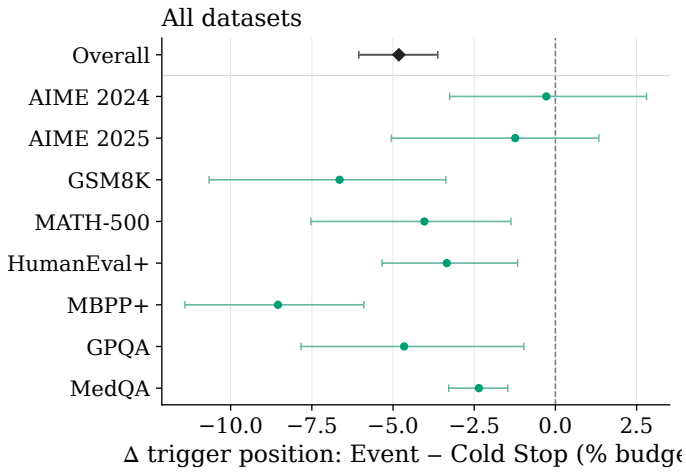
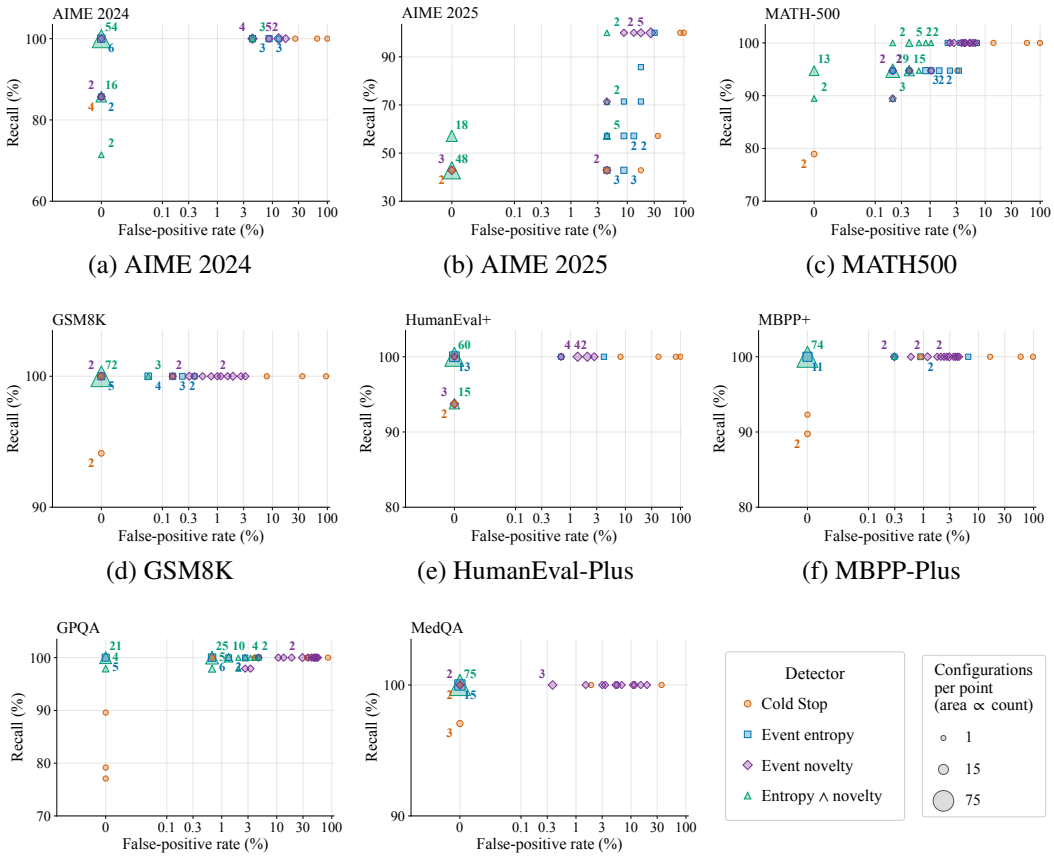


Figure 8: Paired trigger-position difference between the joint event detector and Cold Stop; negative values indicate earlier event detection and error bars show 95% bootstrap confidence intervals.

Dataset-level Operating Points for Termination Detection. Figure 9 disaggregates the average frontier in Fig. 5(a). The joint event-entropy and novelty rule reaches high recall at low observed FPR on several mathematical, code, and question-answering benchmarks, so the aggregate trend is visible across task families. The trade-off is not uniform: AIME2025 shows a sharper recall–FPR trade-off, whereas HumanEval-Plus, MBPP-Plus, and MedQA retain high recall near the low-FPR boundary. Every panel uses the same parameter sweep in Appendix C.3, so these differences describe observed benchmark variation rather than task-specific threshold tuning. Bubble area counts configurations sharing an operating point.

Attractor-Threshold Sensitivity of E-TRACE. Figures 10 and 11 evaluate the E-TRACE (rollback–exit branch) under the shared sweep $\tau_s \in \{0.85, 0.90, 0.95\}$. Each panel reports the change in accuracy and physical generation length relative to the fixed $\tau_s = 0.95$ setting used in the controlled intervention analysis. For Qwen3-8B, lower thresholds generally reduce generation length, while their accuracy effect is dataset dependent: $\tau_s = 0.90$ is favorable on AIME2024 and GPQA, whereas AIME2025 and MedQA improve toward 0.95. Qwen3-32B exhibits a flatter response overall, with smaller length changes and mostly stable accuracy across the sweep.



(a) AIME 2024 (b) AIME 2025 (c) MATH500

(d) GSM8K (e) HumanEval-Plus (f) MBPP-Plus

(g) GPQA (h) MedQA

Figure 9: Dataset-level operating points for termination detection. Panels (a)–(h) plot Recall against the false-positive rate (FPR) for every configuration in the shared threshold and persistence sweep. Marker shape and color identify the detector, while bubble area is proportional to the number of configurations yielding the same (FPR, Recall) pair. No dataset-specific threshold tuning is performed.

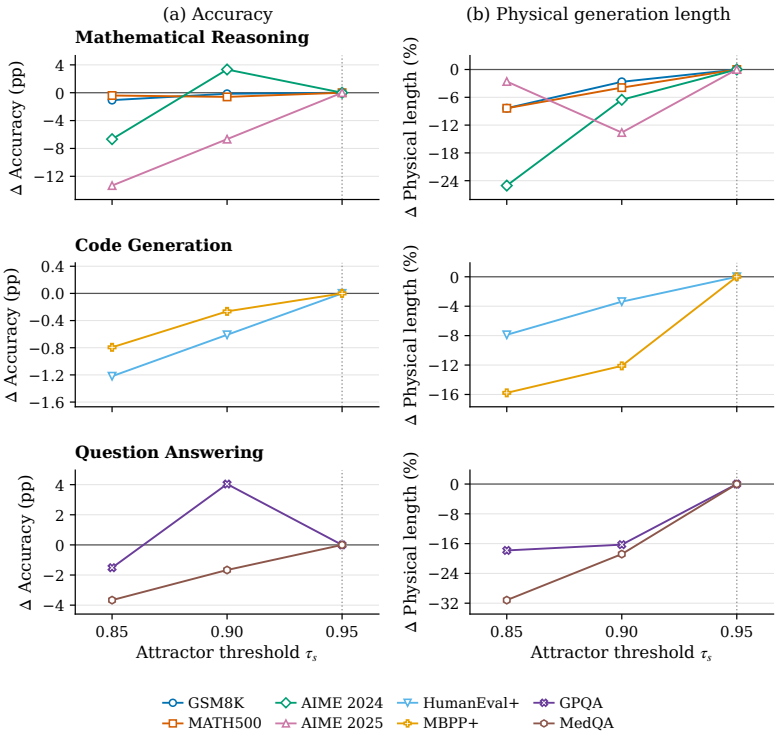


Figure 10: Attractor-threshold sensitivity of E-TRACE with Qwen3-8B. Rows group mathematical reasoning, code generation, and Q&A. The left and right columns show changes in accuracy and physical generation length, respectively, relative to $\tau_s = 0.95$; dotted vertical lines mark this fixed reference setting.

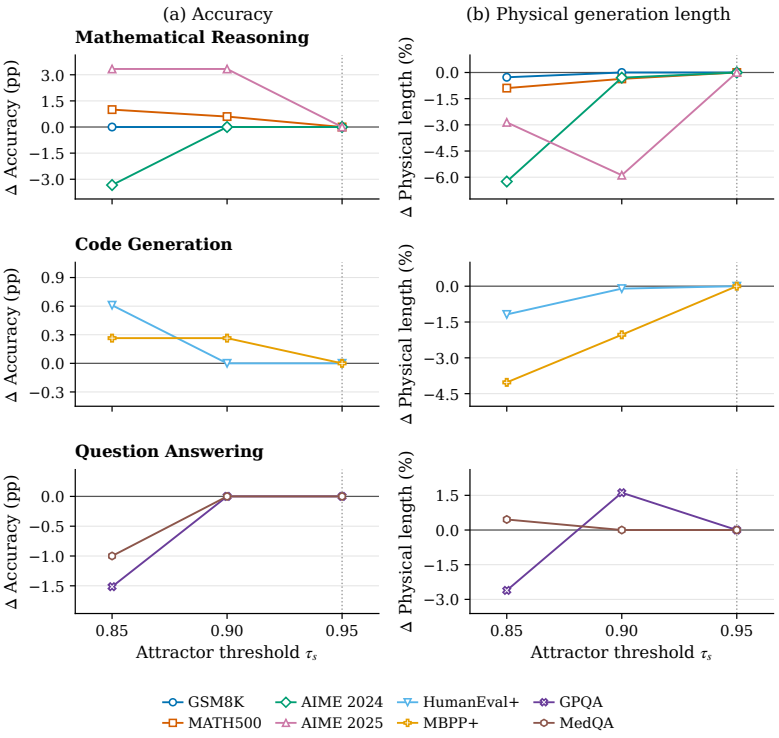


Figure 11: Attractor-threshold sensitivity of E-TRACE with Qwen3-32B; settings follow Fig. 10.