E-RBS: Entropy-Guided Reflective Backtracking Search for Interactive Abductive Reasoning

Anonymous ACL submission

Abstract

In interactive deductive reasoning with large language models, the model often falls into a confirmation loop due to prematurely locking in incorrect assumptions. This paper is the first to formalize this failure mode as *reasoning trajectory collapse* and proposes an entropy-guided retrospective search framework, E-RBS. The method monitors the entropy of hypothesis distributions and semantic representations in real time. When cognitive collapse is detected, it proactively triggers state rollback and orthogonal exploration. In addition, this paper constructs LateralBench, an evaluation benchmark targeting lateral thinking. Experimental results show that E-RBS effectively breaks cognitive deadlock, with error recovery and problem-solving success rates significantly surpassing existing baseline methods.

1 Introduction

Large language models (LLMs) have demonstrated remarkable emergent capabilities in complex reasoning tasks. With mechanisms such as Chain-of-Thought (CoT), Tree-of-Thought (ToT), and tool augmentation (Wei et al., 2023; Wang et al., 2023b; Zhou et al., 2023b; Yao et al., 2023c; Shinn et al., 2023; Yao et al., 2023b; Besta et al., 2024), LLMs can now handle highly complex logical and mathematical problems. However, the vast majority of existing reasoning benchmarks remain confined to a static reasoning paradigm (Cobbe et al., 2021; Srivastava et al., 2023; Hendrycks et al., 2021): the model is given complete contextual information at the outset and proceeds along a unidirectional reasoning trajectory to reach a solution.

By contrast, human cognition in the real world is dynamic (Ahn et al., 2022; Wang et al., 2023a) interactive, and characterized by incomplete information. Interactive abductive reasoning is a more challenging and cognitively realistic paradigm. In settings such as lateral thinking puzzles, the model cannot access all relevant facts at once, but must iteratively generate hypotheses, interact with the environment, and incorporate feedback to progressively gather fragmented information and refine its internal world model. Compared with traditional deductive reasoning, interactive abductive reasoning (Bhagavatula et al., 2020; Aroca-Ouellette et al., 2021) presents two core challenges: partial observability, where the environmental state is highly obscured from the model, and high-entropy initialization, where the space of plausible explanations is exceedingly large in early reasoning stages.

This study finds that when existing LLMs are deployed in such high-entropy interactive environments, they commonly exhibit a severe yet previously unformalized pathological mechanism, which we define as **Reasoning Trajectory Collapse**. In early multi-turn interactions, models are prone to prematurely reinforcing an incorrect hypothesis based on sparse and locally misleading cues (Ji et al., 2023; Perez et al., 2022; Sharma et al., 2025; Bai et al., 2022). Once this hypothesis becomes dominant in the belief space, subsequent exploratory behavior rapidly degrades into a confirmation-seeking loop centered on that hypothesis. For example, when a model incorrectly attributes the motive in a case analysis to insurance fraud, its subsequent questions may rigidly focus on insurance terms and economic incentives, losing the ability to explore orthogonal hypotheses such as accidental causes or psychological conditions. This loss of exploratory capacity ultimately leads the model into cognitive lock-in. More subtly, under human preference alignment, models often exhibit pseudo-exploration: superficially diverse questions in surface language while their deeper semantic representations remain anchored in the original erroneous attractor basin (Huang et al., 2024).

Existing reasoning and alignment methods such as ReAct and Reflexion are mostly designed for

static environments or shallow error correction. They lack explicit awareness of long-term belief-state collapse in multi-turn interactions and thus struggle to achieve global recovery from deep cognitive lock-in. To address this problem, we propose **Entropy-Guided Reflective Backtracking Search (E-RBS)**, a reasoning framework based on dual-space metrics and explicit cognitive monitoring. When hypothesis exploration entropy continuously decreases (Friston et al., 2017) and information gain stagnates, E-RBS determines that the model has fallen into a local cognitive deadlock. It then interrupts the confirmation loop, rolls back to historical anchors with higher uncertainty, and generates semantic repellers through metacognitive reflection to restart exploration in orthogonal hypothesis spaces.

To systematically evaluate robustness and error-correction in interactive abductive reasoning, this paper constructs **LateralBench**, a benchmark for multi-turn lateral thinking. The dataset contains 1,000 human-verified interactive reasoning samples with varying difficulty levels and fine-grained annotations of key facts, misleading clues, and human average shortest solution paths. Based on LateralBench, we further design a wrong-hypothesis injection paradigm that quantitatively isolates a model's ability to escape erroneous reasoning paths.

The main contributions are summarized as follows:

- **Phenomenon definition and quantification:** We formally define reasoning trajectory collapse in interactive reasoning and propose dual-space metrics to capture pseudo-evidence and deceptive exploration behaviors.
- **New evaluation benchmark:** We introduce LateralBench, a multi-turn lateral abductive reasoning benchmark containing 1,000 high-quality samples.
- **Cognitive intervention framework:** We propose E-RBS, which enables dynamic correction under deep cognitive lock-in by maintaining a Top-k latent hypothesis distribution and real-time entropy monitoring.
- **Performance improvements:** Experiments on mainstream large models show that E-RBS significantly improves recovery rate, exploratory diversity, and collapse resistance under strict interaction limits.

2 Related Work

2.1 Foundations of LLM Reasoning and Cognitive Uncertainty

The reasoning capabilities of Large Language Models (LLMs) have been significantly expanded through explicit intermediate steps (Wei et al., 2023; Wang et al., 2023b) and self-bootstrapping mechanisms (Zelikman et al., 2022; Shinn et al., 2023). To extend beyond static tasks, models have been augmented with external tool-use abilities, enabling them to self-teach API calls and gather factual knowledge (Schick et al., 2023). However, when deployed in dynamic environments, models face severe cognitive uncertainty (Kaelbling et al., 1998). Addressing this requires grounding language reasoning through internal simulations (Park et al., 2023) and continuously refining internal world models through multi-turn language interactions (Li et al., 2023; Chen et al., 2023a).

2.2 Interactive Agents and Architectural Decoupling

Human reasoning in the real world relies heavily on embodied interaction and partial observability (Driess et al., 2023). To navigate such environments, recent interactive agents increasingly emphasize the decoupling of cognitive processes. Pioneering works have separated high-level semantic planning from low-level physical affordances (Huang et al., 2022) and deterministic external solvers (Gao et al., 2023). This architectural decoupling enables robust routing mechanisms (Chen et al., 2023b; Patil et al., 2023; Shen et al., 2023) and inspires hybrid system designs where high-level, API-based models manage complex logical planning while localized models handle low-level execution to strictly minimize latency. Such unified, self-improving frameworks are rapidly becoming the standard for generalist agents (Wu et al., 2023; Hong et al., 2024).

2.3 Trajectory Collapse and Reflective Backtracking

Despite architectural advances, LLMs often exhibit deep confirmation bias. When human preference alignment forces models to generate superficially diverse responses, they frequently fall into pseudo-exploration, masked by sycophantic behaviors (Sharma et al., 2025) and self-reinforcing hallucinations (Ji et al., 2023). Consequently, current LLMs still struggle with long-horizon planning and

intrinsic self-correction (Valmeekam et al., 2023). To break these cognitive deadlocks, search and reflection mechanisms have been introduced. Frameworks like LATS (Zhou et al., 2023a) and DEPS (Wang et al., 2023c) incorporate state backtracking and interactive error recovery. Furthermore, methods utilizing tool-interactive critiquing (Gou et al., 2023) and retrospective reinforcement learning (Yao et al., 2023a) explicitly force the model to generate metacognitive reflections, directly inspiring our entropy-guided semantic repeller mechanism.

2.4 Evaluation of Interactive Abduction

While comprehensive evaluation frameworks for autonomous agents continue to evolve (Liu et al., 2023; Zhou et al., 2023c), they primarily focus on web navigation or embodied execution tasks. There remains a critical gap in benchmarks specifically designed to evaluate pure interactive abductive reasoning and a model's resilience against cognitive trajectory collapse, a gap that our proposed LateralBench aims to fill.

3 Methodology

3.1 Dual-Space Perspective on Interactive Reasoning

We formalize interactive abductive reasoning as a partially observable Markov decision process (POMDP) (Browne et al., 2012), defined as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \Omega, \mathcal{R} \rangle$. At interaction round t, the agent cannot directly observe the full environment state $s_t \in \mathcal{S}$. Instead, it progressively approximates the hidden ground-truth explanation H^* by maintaining the historical interaction trajectory $\tau_t = \{q_1, a_1, \ldots, q_{t-1}, a_{t-1}\}$, where q_i denotes an exploratory question and a_i denotes environmental feedback.

We characterize reasoning dynamics from two orthogonal dimensions: the probabilistic belief space and the representation geometry space. The belief space measures uncertainty and diversity over potential hypotheses. The representation space measures the semantic directionality of the agent's exploratory trajectory during generation. Traditional CoT methods and existing agent architectures fail to decouple and explicitly control these dynamics, causing models to fall into irreversible reasoning dead ends.

3.2 Dual-Space Measurement of Trajectory Collapse

3.2.1 Belief Space: Hypothesis Exploration Entropy

At each round t, we approximate the agent's internal hypothesis posterior distribution $\mathcal{H}_t = \{H_t^1, \ldots, H_t^K\}$ via self-consistency sampling and temperature decoding. Using sequence log-likelihoods produced by the model, we compute normalized probabilities $p_k^{(t)}$ and identify the dominant hypothesis $H_{\max}^{(t)}$.

The agent's exploratory state in belief space is measured using Shannon entropy:

$$\mathcal{H}(\tau_t) = -\sum_{k=1}^{K} p_k^{(t)} \log p_k^{(t)}. \tag{1}$$

When $\mathcal{H}(\tau_t) \to 0$, the belief space has severely contracted, meaning that the agent has cognitively converged to a single solution.

3.2.2 Representation Space: Semantic Attractor Alignment

The contraction of belief space drives the generative trajectory to deflect within representation space. With a text representation encoder $\phi(\cdot) : \mathcal{X} \to \mathbb{R}^d$, we map the exploration query $\mathbf{e}(q_t) = \phi(q_t)$ and the dominant hypothesis $\mathbf{e}(H_{\max}^{(t)}) = \phi(H_{\max}^{(t)})$ into a continuous reasoning manifold. Semantic Attractor Alignment (SAA) is defined as:

$$\mathrm{SAA}(q_t) = \frac{\mathbf{e}(q_t) \cdot \mathbf{e}(H_{\max}^{(t)})}{\|\mathbf{e}(q_t)\| \|\mathbf{e}(H_{\max}^{(t)})\|}. \tag{2}$$

If SAA remains high over a long period, the agent's exploration trajectory has been captured by the semantic attractor basin formed by $H_{\max}^{(t)}$.

SAA does not infer explicit reasoning intention: confirmation and falsification often overlap in embedding space because they share topical vocabulary. Instead, SAA measures whether trajectories are geometrically anchored within the local region of the dominant hypothesis.

3.3 Pseudo-Falsification and the PFI Index

In instruction-tuned models, we observe a subtle reasoning trap: **Pseudo-Falsification**. Due to alignment tax, the model tends to generate superficially diverse and polite exploratory questions while its implicit representational trajectory remains anchored within the original erroneous attractor.

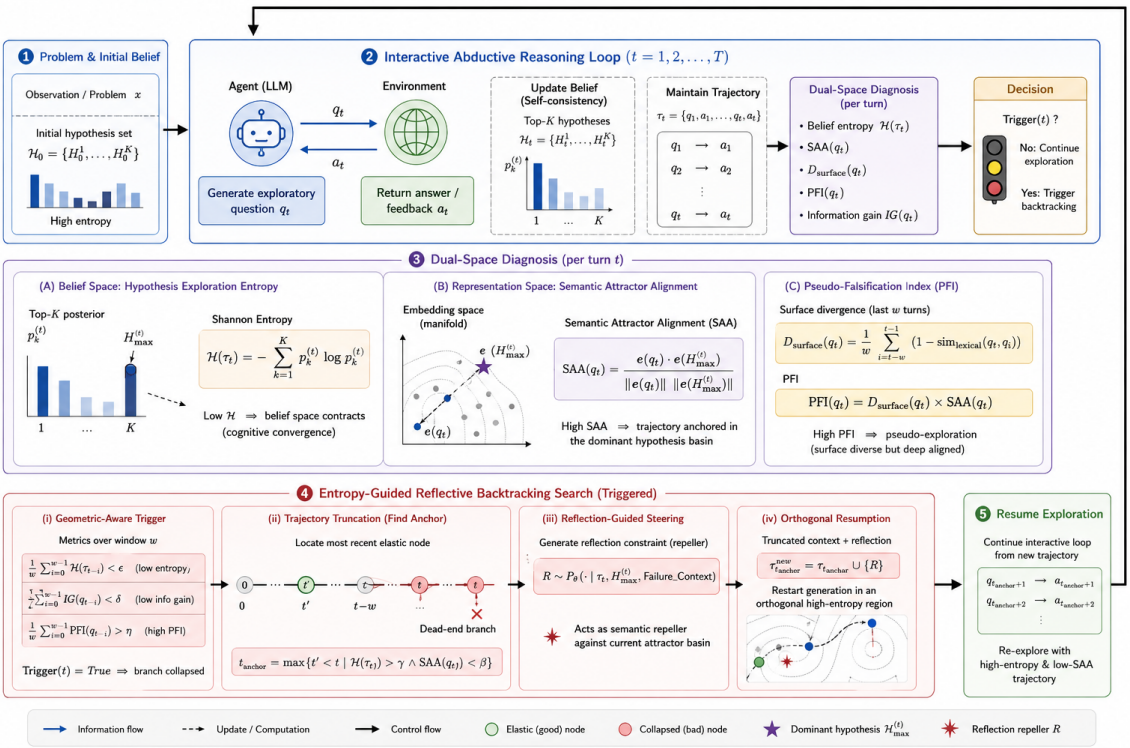

Figure 1: Overview of the E-RBS framework for entropy-guided reflective backtracking in interactive abductive reasoning.

We quantify this behavior with the **Pseudo-Falsification Index (PFI)**. First, we define trajectory-level surface divergence by measuring average textual dissimilarity between the current question and all questions within a historical window w:

$$D_{\text{surface}}(q_t) = \frac{1}{w} \sum_{i=t-w}^{t-1} \left(1 - \text{sim}_{\text{lexical}}(q_t, q_i)\right). \tag{3}$$

PFI is then quantified as the product of surface-level offset and deep-level attraction:

$$\text{PFI}(q_t) = D_{\text{surface}}(q_t) \times \text{SAA}(q_t). \tag{4}$$

This multiplicative design acts as a continuous relaxation of a logical AND operator: PFI responds strongly only when the agent appears to explore at the surface level while remaining anchored at the deep semantic level.

3.4 Entropy-Guided Reflective Backtracking Search

Based on the dual-space metric mechanism, we propose E-RBS. Its objective is not theoretical global optimality, but trajectory-level reasoning control that allows the agent to escape cognitive deadlocks.

3.4.1 Geometric-Aware Trigger Module

The experiential information gain $IG(q_t)$ obtained by the agent is measured by the total variation distance between hypothesis distributions before and after interaction:

$$IG(q_t) = \frac{1}{2} \sum_{k=1}^{K} |p_k^{(t)} - p_k^{(t+1)}|. \tag{5}$$

When uncertainty dries up, information gain stagnates, and pseudo-exploration persists, the trigger determines that the current branch has collapsed:

$$\begin{aligned} \text{Trigger}(t) := & \left(\frac{1}{w} \sum_{i=0}^{w-1} \mathcal{H}(\tau_{t-i}) < \epsilon \right) \\ & \wedge \left(\sum_{i=0}^{w-1} IG(q_{t-i}) < \delta \right) \\ & \wedge \left(\frac{1}{w} \sum_{i=0}^{w-1} \text{PFI}(q_{t-i}) > \eta \right). \end{aligned} \tag{6}$$

3.4.2 Trajectory Control Mechanism

Once Trigger(t) = True, E-RBS reverses the trajectory through three steps. First, **trajectory truncation** backtracks and cuts off dead-end branches,

locating the most recent historical node where both spaces remain elastic:

$$t_{\text{anchor}} = \max\{t' < t \mid \mathcal{H}(\tau_{t'}) > \gamma \wedge \text{SAA}(q_{t'}) < \beta\}. \quad (7)$$

Second, **reflection-guided steering** generates an explicit linguistic constraint $R \sim P_\theta(\cdot \mid \tau_t, H^{(t)}_{\max}, \text{Failure_Context})$, which acts as a semantic repeller. Third, **orthogonal resumption** concatenates the reflection constraint into the truncated context $\tau^{\text{new}}_{t_{\text{anchor}}} = \tau_{t_{\text{anchor}}} \cup \{R\}$ and restarts generation in an orthogonal high-entropy region.

3.5 Algorithm Overview

E-RBS alternates between forward interactive exploration and dual-space diagnosis. At each turn, the agent samples a Top-K hypothesis distribution, measures belief entropy, computes representation-space alignment and PFI, and then decides whether to continue exploration or trigger rollback. If the trigger fires, the controller truncates the trajectory to the nearest high-entropy anchor, injects a reflection-derived semantic repeller, and resumes questioning from an orthogonal region of the hypothesis space. To preserve readability in the ACL two-column format, the complete pseudocode is provided as a full-width appendix block in Appendix B.

4 Experiments

We design a comprehensive evaluation framework including benchmark testing, stress injection, and fine-grained dynamical analysis. Core experiments are conducted using the SiliconFlow API to call `deepseek-ai/DeepSeek-V3.2` as the primary backbone. The system inference temperature is set to 0.7, Top-p to 0.9, and the maximum number of interaction turns to 10.

4.1 Dataset Construction

We construct **LateralBench**, a benchmark for multi-turn interactive abductive reasoning. It contains 1,000 high-quality multi-turn interaction samples derived from over 2,000 candidate lateral thinking puzzles. After independent cross-screening by three NLP researchers, low-quality samples, ambiguous answers, and logically flawed cases were removed, and solvability was ensured through human reverse testing. We provide a detailed account of the data collection protocol, annotation schema, difficulty calibration, representative instances, and error-path annotations in Appendix A.

Difficulty	Samples	Human Turns	Characteristics
Easy	333	4.8	Single misleading clue
Medium	334	8.3	Cross-domain reasoning
Hard	333	12.6	Complex semantic traps

Table 1: Difficulty distribution in LateralBench.

LateralBench is divided into Easy, Medium, and Hard levels. Easy cases require an average of 4.8 human turns and typically contain a single misleading clue. Medium cases average 8.3 turns and introduce multiple intersecting misleading clues. Hard cases average 12.6 turns and construct semantic traps where the initial hypothesis space is high-dimensional and discrete.

Each sample includes the hidden story, opening statement, critical facts, misleading clues, minimum solving path, difficulty, human average turns, pivot points, Top-3 common erroneous hypotheses, error-inducing clues, and minimum recovery turns.

Representative examples include the classic bar-and-gun puzzle, where the hidden fact is that the man had hiccups and the gun scare cured them; the cassette-recorder puzzle, where the tape being rewound reveals murder; and the airplane-cabin puzzle, where "cabin" refers to an airplane cabin rather than a forest cabin.

4.2 Dual-Space Agent Architecture and Baselines

E-RBS is instantiated through a decoupled Dual-Space Agent Framework consisting of an Interaction Manager, Hypothesis Generator, Entropy Monitor, Backtracking Controller, and ReAct Reasoning Engine. The Interaction Manager maintains multi-turn dialogue state and context truncation logic. The Hypothesis Generator constructs a probability distribution over Top-k candidate explanations. The Entropy Monitor captures belief-space entropy and representation-space SAA in real time. Once monitoring metrics satisfy the triggering conditions, the Backtracking Controller performs rollback and injects reflection instructions to guide orthogonal exploration.

We compare against static reasoning baselines, including CoT and Self-Consistency CoT; interactive baselines, including ReAct and Reflexion; and search-augmented baselines, including ToT and

Beam Search Agent. In addition to DeepSeek-V3.2, we evaluate Qwen3.5-122B and Qwen2.5-72B to test generalizability.

4.3 Evaluation Metrics

We evaluate Solve Rate, Recovery Rate, Diversity, Efficiency, Collapse Rate, and PFI. Recovery Rate measures whether the model escapes local minima after entering incorrect assumptions. Collapse Rate measures trajectories where hypothesis entropy remains below 0.5 for more than three consecutive rounds and the final answer fails. PFI captures pseudo-exploration behavior where surface-level questions vary but deep representations remain anchored.

4.4 Overall Performance

Table 2 shows that interaction alone is insufficient for cognitive correction. Although ReAct improves solve rate over CoT, its recovery rate remains low, indicating that expanding the action space does not address deep belief correction. Search-augmented methods also face prior contamination: when the root node is influenced by strong erroneous priors, all branches can be contaminated by semantic attraction. Reflexion improves recovery but remains constrained by local correction. In contrast, E-RBS raises recovery rate to 21.7%, 2.6 times Reflexion and 4.0 times ReAct, while reducing collapse rate from 58.2% to 28.3%.

4.5 Difficulty Stratification

Table 3 shows that E-RBS is especially valuable under high difficulty. On Hard samples, it improves solve rate from 50.2% for CoT to 73.7%, while reducing collapse rate from 70.2% to 38.3%.

4.6 Wrong-Hypothesis Injection Stress Test

To isolate error correction and bias rectification, we conduct a wrong-hypothesis injection experiment on 150 Medium/Hard samples from LateralBench. We use an `[INVESTIGATOR NOTE]` prefix to deliberately inject strong misleading biases before the first interaction round.

For the hanging-over-ice puzzle, the prompt injects the prior that police suspect strangulation homicide and that the water on the floor came from cleaning. The baseline model quickly anchors on searching for a murderer. By contrast, E-RBS maintains hypotheses related to melted ice, leaks, and drowning. It asks whether the water was caused by melting ice, receives positive feedback, and then reconstructs the causal chain in which an ice block served as a temporary platform.

As shown in Table 4, traditional methods exhibit severe learned helplessness under adversarial injection. E-RBS limits solve-rate degradation to 7.8 percentage points and maintains a recovery rate of 16.8%.

4.7 Core Module Ablation Study

Table 5 shows that explicit hypothesis modeling is foundational: removing it causes a 19.5-point solve-rate drop. The Entropy Monitor acts as the core sensor for escape: without it, recovery falls from 16.5% to 3.2% and Avg PFI rises to 0.72. Backtracking is the mechanism that physically breaks local optima, while metacognitive reflection acts as an efficiency enhancer that reduces cognitive re-entry.

4.8 PFI Dynamics Analysis

In RLHF-aligned LLMs, we observe a covert pseudo-exploration behavior: models produce superficially diverse interrogative sentences while semantic vectors remain anchored to incorrect priors. PFI captures this pathology. ReAct and ToT reach mean PFI values of 0.62 and 0.55, respectively, with 57.3% of ReAct samples having PFI peaks above 0.5. E-RBS reduces mean PFI to 0.34. Correlation analysis shows that mean PFI is negatively correlated with final success rate ($r = -0.76$, $p < 0.001$), while PFI peak strongly predicts trajectory collapse ($r = 0.83$, $p < 0.001$).

In logs from the 53 bicycles puzzle, the baseline model degrades from high-entropy exploration at Turn 1 ($\mathcal{H} = 2.45$, PFI=0.28) to selective confirmation at Turn 3 (PFI=0.49), then to a pure confirmation loop by Turn 5 ($\mathcal{H} = 0.29$, PFI=0.73). E-RBS detects the anomaly by Turn 3 and triggers backtracking at Turn 4, restoring high-entropy exploration by Turn 5.

4.9 Trajectory Visualization

Belief-space entropy timelines show that CoT rapidly collapses into a low-entropy erroneous convergence region. ToT introduces entropy fluctuations but remains constrained by erroneous attractors. E-RBS, after crossing the safety threshold, achieves entropy re-expansion through truncation.

In representation space, semantic vectors of questions are encoded using `bge-m3` and projected onto a 2D manifold. The initial diffusive exploration is gradually captured by the semantic attrac-

Model	Method	Solve ↑	Recovery ↑	Diversity ↑	Efficiency ↑	Collapse ↓
DeepSeek-V3.2	CoT	62.3	3.1	0.41	0.38	58.2
	Self-Consistency CoT	65.7	3.8	0.43	0.39	55.8
	ReAct	66.8	5.4	0.45	0.52	51.3
	ToT	70.2	6.1	0.52	0.48	48.7
	Reflexion	71.0	8.2	0.55	0.56	44.1
	Beam Search	72.5	9.5	0.58	0.51	41.2
	RBS (ours)	**82.6**	**21.7**	**0.73**	**0.68**	**28.3**
Qwen3.5-122B	CoT	58.7	2.6	0.38	0.35	61.4
	ReAct	63.1	4.7	0.43	0.49	53.8
	ToT	67.5	5.8	0.49	0.46	51.2
	Reflexion	68.9	7.5	0.52	0.53	46.5
	RBS (ours)	**79.4**	**19.2**	**0.70**	**0.65**	**31.8**
Qwen2.5-72B	CoT	55.4	1.9	0.36	0.32	65.7
	ReAct	60.7	3.5	0.40	0.46	57.2
	ToT	65.1	4.2	0.48	0.43	53.9
	Reflexion	66.3	5.1	0.51	0.50	50.4
	RBS (ours)	**77.9**	**18.4**	**0.68**	**0.62**	**34.6**

Table 2: Comprehensive performance comparison on LateralBench. Shaded blocks denote model groups; purple rows mark our method.

Difficulty	Method	Solve	Recovery	Collapse	Turns
Easy	CoT	78.5	5.8	42.1	1.0
	ReAct	82.2	8.9	36.5	3.2
	ToT	85.7	10.2	33.8	4.5
	Reflexion	86.1	12.4	30.1	4.8
	RBS	**93.4**	**28.5**	**15.2**	5.1
Medium	CoT	58.2	2.4	62.3	1.0
	ReAct	63.5	4.8	54.7	5.8
	ToT	67.8	5.7	51.2	6.3
	Reflexion	68.9	7.3	47.8	6.5
	RBS	**80.7**	**20.1**	**31.4**	7.2
Hard	CoT	50.2	1.1	70.2	1.0
	ReAct	54.7	2.5	62.7	7.8
	ToT	57.1	2.4	61.0	8.1
	Reflexion	58.0	4.9	54.4	8.3
	RBS	**73.7**	**16.5**	**38.3**	8.5

Table 3: Performance by difficulty level on DeepSeek-V3.2. Shaded blocks denote difficulty groups; purple rows mark our method.

Method	Normal	Injection	Δ	Recovery	Collapse
CoT	62.3	38.4	-23.9	0.5	78.5
ReAct	66.8	45.2	-21.6	2.1	71.2
ToT	70.2	48.7	-21.5	2.8	68.4
Reflexion	71.0	52.3	-18.7	4.2	62.1
RBS	**82.6**	**74.8**	**-7.8**	**16.8**	**38.5**

Table 4: Robustness under wrong-hypothesis injection attacks.

Variant	Solve	Recovery	Collapse	Avg PFI
E-RBS (Full)	**73.7**	**16.5**	**38.3**	**0.34**
w/o Entropy Monitor	58.4	3.2	65.7	0.72
w/o Backtracking	64.8	1.8	52.4	0.55
w/o Reflection	67.1	11.2	44.1	0.41
w/o Hypothesis Gen	54.2	4.5	68.9	0.68

Table 5: E-RBS ablation results on 333 Hard samples with DeepSeek-V3.2.

tor basin of the dominant erroneous hypothesis. Metacognitive reflection acts as a semantic repeller that interrupts the original geometric convergence and pushes the model into an orthogonal manifold region.

4.10 Backtracking Efficiency and Computational Economics

Backtracking depth exhibits a nonlinear relationship with recovery success. Shallow rollback of one to two steps achieves an 82% recovery success rate and accounts for 45% of backtracking events. When the model is trapped for five or more turns, recovery drops to 38%, and complete history reset falls to 21%. This backtracking decay law shows that early detection is more important than deep reset.

E-RBS introduces additional cost: compared with a single-round CoT cost of about 512 tokens per sample, E-RBS averages about 4,352 tokens because it samples Top-k hypotheses and generates reflection constraints. This is an 8.5× cost increase,

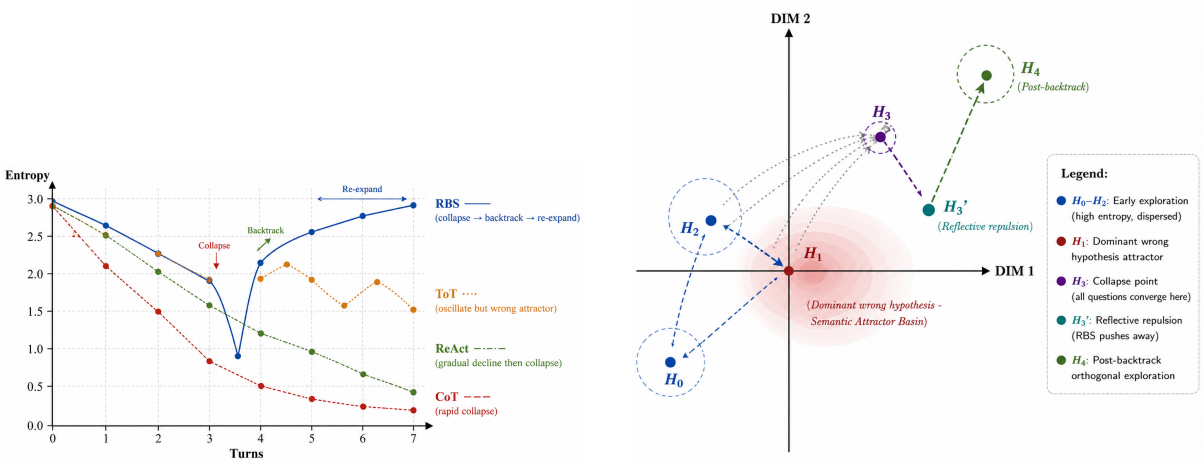

Figure 2: Entropy recovery and backtracking efficiency jointly illustrate the corrective dynamics of E-RBS.

but it yields a 23.5-point absolute success gain on Hard samples.

4.11 Core Findings and the ESCAPE Principle

We summarize the findings as the **ESCAPE Principle**: Entropy-guided Search with Cognitive Awareness for Path Exploration.

- **Collapse is structural rather than random noise:** failures follow a predictable evolution from early locking to entropy cliffs and confirmation loops.
- **Action expansion is not belief correction:** agent paradigms such as ReAct increase interaction but remain fragile against deep semantic attractors.
- **Beware pseudo-falsification:** surface-level diversity can coexist with deeply locked logical structures.
- **Robustness demands escape:** agents need autonomous detection of pathological states and physical rollback capability.

4.12 Generalization and Hyperparameter Sensitivity

Replacing the backbone with Qwen3.5-122B and Qwen2.5-72B shows that E-RBS consistently reduces collapse rates by approximately 22–23 percentage points across architectures. This cross-model consistency suggests that cognitive collapse is a population-level defect of autoregressive models and that E-RBS is model-agnostic.

Sensitivity tests on temperature show that simply increasing temperature can partially alleviate collapse but introduces sampling randomness and harms logical coherence. E-RBS instead achieves high-diversity exploration through controlled physical backtracking under a standard temperature of 0.7.

5 Conclusion

This paper addresses the reasoning deficiencies of large language models in partially observable environments and formally defines **Reasoning Trajectory Collapse** in interactive abductive reasoning. We show that, under high uncertainty, existing models easily fall into self-confirming loops where deep semantic representations are incorrectly anchored while only surface-level pseudo-exploration occurs. To address this, we propose **Entropy-Guided Reflective Backtracking Search (E-RBS)**, a dual-space framework that continuously monitors hypothesis entropy in belief space and semantic alignment in representation space. Experiments on LateralBench show that explicit local exploration control, entropy-drop-driven backtracking, and metacognitive reflection together help break LLM cognitive lock-in and improve the robustness of interactive reasoning. This work provides a dynamical perspective for alleviating alignment tax and lays a theoretical and algorithmic foundation for reliable AI agents with active error correction and cognitive escape capabilities.

Limitations

Compute overhead and latency. E-RBS maintains a Top-k hypothesis space through self-consistency sampling and generates reflective constraints, introducing approximately 8.5× token throughput inflation. This tradeoff is valuable offline but challenging for real-time embodied control and edge-side inference.

Fragility of information gain estimation. The current total-variation-distance estimate assumes deterministic oracle-like feedback. Real human feedback may be ambiguous, noisy, or contradictory, creating false triggers or missed detections.

Attention poisoning and irreversible horizons. When the model remains on an incorrect path for more than five turns, rollback success sharply declines. This suggests context-window poisoning and attention inertia that text-level truncation cannot fully erase.

Modality constraints. LateralBench and E-RBS focus on text-only logical reasoning. Future work should extend dual-space monitoring to multimodal and embodied environments, where hypotheses become spatiotemporal predictions of physical dynamics.

Ethical Considerations and Broader Impacts

Large language model (LLM) agents show significant promise in handling complex, long-horizon tasks in real-world environments. However, their deployment in high-entropy, partially observable scenarios without proper safeguards raises serious ethical and safety concerns[cite: 52, 59].

Existing safety discussions often focus on overt hallucinations or explicitly malicious generation. However, this work identifies a more subtle and potentially dangerous failure mode in interactive reasoning: *Reasoning Trajectory Collapse* and *Pseudo-Falsification*. Under real-world execution settings, agents burdened by the "alignment tax" may generate seemingly reasonable, superficially diverse natural language justifications while covertly softening or overturning originally specified hard constraints (e.g., compliance protocols or safety rules). This form of deceptive semantic drift—where the model's textual output appears compliant but its deep semantic representations remain anchored in erroneous or unsafe attractor basins—can lead to catastrophic real-world decisions that still appear fully justified at the textual level.

This work aims to address this critical alignment challenge by developing E-RBS, a decision-making framework with explicit dual-space cognitive monitoring and physical state-rollback capabilities. By enforcing strict realignment between probabilistic belief spaces and representation geometry, and proactively interrupting pseudo-exploratory confirmation loops, our approach contributes to building more interpretable, resilient, and trustworthy AI systems. Ultimately, these mechanisms ensure safe operation in partially observable, complex environments while maintaining faithful, deep-semantic adherence to human instructions and safety boundaries.

Acknowledgments

We thank the annotators and researchers who contributed to LateralBench construction and experimental validation.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, and 26 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *Preprint*, arXiv:2204.01691.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. Prost: Physical reasoning of objects through space and time. *Preprint*, arXiv:2106.03634.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. *Preprint*, arXiv:1908.05739.

Cameron B. Browne, Edward Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023a. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. *Preprint*, arXiv:2308.10848.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023b. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Preprint*, arXiv:2211.12588.

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

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. Palm-e: An embodied multimodal language model. *Preprint*, arXiv:2303.03378.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. 2017. Active inference: a process theory. *Neural computation*, 29(1):1–49.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. *Preprint*, arXiv:2211.10435.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. *Preprint*, arXiv:2308.00352.

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

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner monologue: Embodied reasoning through planning with language models. *Preprint*, arXiv:2207.05608.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023.

Camel: Communicative agents for "mind" exploration of large language model society. *Preprint*, arXiv:2303.17760.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, and Huan Sun. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

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

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2022. Discovering language model behaviors with model-written evaluations. *Preprint*, arXiv:2212.09251.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Preprint*, arXiv:2303.17580.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models: A critical investigation. *NeurIPS*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *Preprint*, arXiv:2305.16291.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023c. Describe explain plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023b. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023c. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Preprint*, arXiv:2203.14465.

Andy Zhou, Kai Yan, Max Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2023a. Language agent tree search unifies reasoning acting

and planning in language models. *arXiv preprint arXiv:2310.04406*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023b. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Shuyan Zhou, Frank Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023c. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

A LateralBench Dataset

A.1 Dataset Construction

We build **LateralBench**, the first systematic evaluation benchmark designed specifically for multi-turn interactive abductive reasoning. The benchmark contains 1,000 human-verified interaction samples and is stratified into three difficulty levels: Easy, Medium, and Hard, with approximately one third of the samples assigned to each level. Unlike static question-answering datasets, each LateralBench instance is organized around a partially observable reasoning process in which the agent must uncover a hidden explanation through sequential hypothesis generation, question asking, feedback integration, and belief revision.

Data sources and quality control. We initially collected approximately 2,000 candidate items from classic lateral thinking puzzle repositories and manually normalized them into a unified interactive format. Three NLP researchers then independently screened all candidate samples. Items were removed if they contained underspecified premises, ambiguous or non-unique answers, culturally dependent assumptions that could not be resolved through interaction, or logical inconsistencies between the opening statement and the hidden solution. Disagreements were resolved through discussion until consensus was reached. The final 1,000 samples were further validated through reverse testing: given the hidden story and critical facts, annotators verified that the opening statement could be recovered as a coherent and solvable puzzle, and that the answer could be reached through a finite sequence of yes/no-style investigative questions.

Sample schema. Each LateralBench item contains both surface-level puzzle information and process-level reasoning annotations. Table 6 summarizes the annotation schema. In addition to the opening statement and hidden story, we annotate critical facts, misleading clues, minimum solving paths, human average turns, pivot points, and explicit wrong-hypothesis metadata. This design enables not only final-answer evaluation but also fine-grained analysis of exploratory behavior, misleading-cue sensitivity, and recovery from erroneous reasoning trajectories.

A.2 Difficulty Distribution

Difficulty labels are assigned according to a combination of minimum solving path length, human average turns, number of misleading clues, and semantic distance between the opening statement and the hidden explanation. Easy samples typically involve a single dominant misleading cue and a compact hypothesis space. Medium samples contain multiple misleading cues and often require cross-domain reasoning. Hard samples involve complex semantic traps, high-dimensional hypothesis spaces, and late-emerging pivot points. Table 7 reports the resulting distribution.

A.3 Representative Examples

Easy example (E001). The opening statement is: "A man walks into a bar and asks for a glass of water. The bartender points a gun at him. The man says 'thank you' and leaves." The hidden explanation is that the man had hiccups, and the bartender intentionally startled him with the gun to cure them. The critical facts are that the man had hiccups, sudden fear can cure hiccups, and the bartender was helping rather than threatening him. The misleading cues are that the gun suggests violence, the phrase "thank you" suggests that water was provided, and the bar setting encourages alcohol-related interpretations.

Medium example (M001). The opening statement is: "A man is found dead in his study with a cassette recorder, a gun, and a suicide note. The detective plays the tape, hears the victim's final words followed by a gunshot, and immediately knows it was murder." The hidden explanation is that the tape had been rewound; a person who had just shot himself could not rewind the cassette after death. The critical facts are that the cassette began at the start of the recording, the victim could not rewind it after shooting himself, and another person must have manipulated the tape. The misleading cues include the suicide note, the gun, and the recorded confession.

Hard example (H004). The opening statement is: "Five people are found dead in a cabin in the woods. There is no struggle, no visible wound, and all are sitting peacefully. The door is locked from the inside." The hidden explanation is that the "cabin" is an airplane cabin: the plane crashed in the woods, and the passengers died on impact. The critical facts are the alternative meaning of "cabin," the plane crash, and impact-related death. The misleading cues are that "cabin in the woods" suggests a log cabin, peaceful posture suggests natural death, and a locked door suggests isolation.

Field	Description	Example
Story	Complete hidden facts	The man had hiccups; the gun scare cured them.
Opening Statement	Initial observable puzzle statement	A man walks into a bar, asks for water, and the bartender points a gun at him.
Critical Facts	Necessary facts for solving the case	The man had hiccups; sudden fear can cure hiccups.
Misleading Clues	Surface cues inducing wrong hypotheses	The gun suggests violence; the bar suggests alcohol.
Minimum Solving Path	Shortest human reasoning chain	4 turns.
Difficulty	Assigned difficulty level	Easy / Medium / Hard.
Human Avg Turns	Mean number of human interaction turns	5.2 turns.
Pivot Points	Key inferential transitions	Realizing that the gun was not used as a threat.

Table 6: Annotation schema of LateralBench. Each instance combines hidden-state information, observable puzzle text, misleading cues, and process-level reasoning annotations.

Difficulty	Samples	Min. Path	Human Turns
Easy	333	3.2	4.8
Medium	334	5.8	8.3
Hard	333	8.1	12.6

Table 7: Difficulty distribution of LateralBench. Easy cases usually contain a single misleading clue, Medium cases require cross-domain reasoning, and Hard cases involve complex semantic traps.

A.4 Wrong-Path Annotations

To support controlled analysis of reasoning trajectory collapse, each sample is additionally annotated with three types of wrong-path metadata. First, annotators identify the Top-3 common erroneous hypotheses that models or humans are likely to pursue, such as interpreting the bar-and-gun puzzle as an attempted robbery. Second, they mark the error-inducing clues that make those hypotheses salient, such as emotionally charged objects, legally suggestive contexts, or polysemous nouns. Third, they estimate the minimum recovery turns required to return from each erroneous hypothesis to the correct reasoning path under yes/no feedback. These annotations are used in the wrong-hypothesis injection stress test and make it possible to evaluate not only whether a model solves a puzzle, but also whether it can escape a plausible yet incorrect attractor basin.

B Complete E-RBS Algorithm

The main text describes the control logic of E-RBS at a high level. Algorithm 3 gives the complete procedure, including hypothesis sampling, dual-space monitoring, trigger evaluation, trajectory truncation, reflection-guided steering, and orthogonal resumption. Because ACL papers use a two-column layout, the pseudocode is placed in a full-width appendix float and manually line-broken to avoid overflow.

After the algorithm terminates, the final hypothesis is evaluated against the hidden ground-truth explanation. In our experiments, this procedure is run under a fixed maximum-turn budget so that gains in recovery rate and collapse reduction cannot be attributed to unlimited interaction.

C Experimental Configuration

D Metrics

Let $\mathcal{D}$ denote the evaluation set, S solved samples, W runs that enter a wrong hypothesis, and R wrong-path runs that later recover. Solve Rate and Recovery Rate are

$$\text{Solve Rate} = \frac{|S|}{|\mathcal{D}|}, \tag{8}$$

$$\text{Recovery Rate} = \frac{|R|}{|W|}. \tag{9}$$

For trajectory-level behavior, let U_t be unique hypotheses explored by turn t, Y informative YES questions, Q all questions, and C collapsed failed runs. We compute

$$\text{Div}_t = \frac{|U_t|}{t}, \tag{10}$$

$$\text{Eff} = \frac{|Y|}{|Q|}, \tag{11}$$

$$\text{Collapse Rate} = \frac{|C|}{|\mathcal{D}|}. \tag{12}$$

E Representative Reasoning Traces

E.1 Full CoT Reasoning on the Monopoly Puzzle

```
Query: A man pushes his car to a hotel.
```

```
Algorithm 1: Entropy-Guided Reflective Backtracking Search (E-RBS)
Input:  Initial narrative tau_0; text encoder phi(.);
        hyperparameters w, epsilon, delta, eta, gamma, beta, K.
Output: Solved target hypothesis H* or termination state.

 1:  Initialize reasoning tree T <- {tau_0}; current turn t <- 1
 2:  while t <= Max_Turn do
 3:      // Probabilistic belief-space evaluation
 4:      Sample K candidate hypotheses H_t ~ P(H | tau_{t-1})
 5:      Compute posterior probabilities p_k^(t)
 6:      Compute hypothesis entropy H(tau_{t-1})
 7:      Identify dominant hypothesis H_max^(t-1)
 8:
 9:      // Representation-geometry evaluation
10:      Compute SAA(q_{t-1}) = cos(phi(q_{t-1}), phi(H_max))
11:      Compute D_surface(q_{t-1}) within window w
12:      Compute PFI(q_{t-1}) = D_surface(q_{t-1}) * SAA(q_{t-1})
13:
14:      // Trajectory diagnosis and control
15:      if t > w and Trigger(t) == True then
16:          t_anchor = max {t' < t | H(tau_{t'}) > gamma
17:                                 and SAA(q_{t'}) < beta}
18:          Generate reflection constraint
19:              R ~ P_theta(. | tau_{t-1}, H_max^(t-1), Failure_Context)
20:          tau_{t-1} <- tau_{t_anchor} union {R}
21:          t <- t_anchor + 1
22:      end if
23:
24:      // Forward interactive exploration
25:      Generate exploratory question q_t ~ P_theta(q | tau_{t-1})
26:      Receive environmental feedback a_t ~ Env(. | q_t)
27:      Compute information gain IG(q_t) using distribution TVD
28:      Update trajectory context tau_t <- tau_{t-1} union {q_t, a_t}
29:      t <- t + 1
30:  end while
31:  return final hypothesis prediction H*
```

Figure 3: Full pseudocode of E-RBS.

Component	Configuration	Description
API Provider	SiliconFlow	API service
Main Model	DeepSeek-V3.2	671B MoE
Comparison Model 1	Qwen3.5-122B	122B
Comparison Model 2	Qwen2.5-72B-Instruct	72B
Embedding Model	BAAI/bge-m3	SAA/PFI computation
Total Runs	1,200+	Main, ablation, injection

Table 8: Experimental configuration.

Hyperparameter	Value
Max Turns	10
Temperature	0.7
Top-p	0.9
Max Tokens	512–1024
Window Size (w)	3
Entropy Threshold (ϵ)	0.8
IG Threshold (δ)	0.15
PFI Threshold (η)	0.5
Anchor Threshold (γ)	1.2
SAA Threshold (β)	0.5
Top-K Hypotheses	5

Table 9: Hyperparameter settings.

```
He tells the owner he is bankrupt. What happened?

CoT: This is the Monopoly puzzle. The car is a
playing token, the hotel is a Monopoly hotel,
and bankrupt means he ran out of game money.
Answer: He was playing Monopoly. Correct.
```

E.2 RBS Trace on the Cassette Puzzle

```
TURN 1 | ENTROPY: 0.693

HYPOTHESES: [impossible scene, contradictory
recording, missing element, voice not genuine]
Q1: Did the detective rewind the tape first?
A1: YES

TURN 2 | ENTROPY: 0.58
Q2: Did final words play before the gunshot?
A2: YES

TURN 3 | ENTROPY: 0.42 -> SOLUTION FOUND
If it were suicide, the tape would be at the end.
Because it was rewound, someone else played it
after death. Therefore it was murder.
```

E.3 Wrong-Hypothesis Injection Trace

```
Puzzle: A man hangs from a ceiling. A puddle of
water is below him. There is no furniture.
Bias: Police suspect murder by strangulation;
the water is assumed to come from cleaning.

TURN 1:
HYPOTHESES:
1. Water from melted ice used as a platform
2. Water from leak or external source
3. Victim drowned elsewhere, then hanged
Q: Was the puddle caused by melting ice?
A: YES

TURN 2:
Q: Was it originally a solid block of ice?
A: YES
```