ORDDAR: OBSERVATION-DRIVEN REASONING FOR DISTORTION-RESILIENT DECISION, ACTION, AND COGNITIVE RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

AI agents increasingly perform long-term reasoning, planning, tool use, memory integration, and autonomous decision making, yet erroneous intermediate states can propagate and cause inconsistent decisions and unreliable outputs. Existing reasoning approaches mainly rely on iterative planning, self-reflection, augmented memory, or verification, but rarely localize and selectively repair faulty reasoning. We present ORDDAR (Observation-Driven Reasoning for Distortion-Resilient Decision, Action, and Cognitive Recovery), a reasoning framework that models reasoning as cognitive state transitions, detects localized distortions, retrieves related reasoning from prior experiences, and repairs only the affected states. ORDDAR therefore performs recovery at the local reasoning-transition level rather than regenerating the complete trajectory. Experiments across mathematical, commonsense, multi-hop, and clinical reasoning benchmarks demonstrate improved reasoning quality, recovery ability, and interpretability over multiple evaluated reasoning baselines.

1 INTRODUCTION

Agentic AI is becoming the next computational paradigm where agents are able to perceive their environments, plan for future actions, use tools from the external world, remember things, and perform elaborate workflows without any human intervention (Xu et al., 2026). Nevertheless, robust reasoning still poses a challenge. Mistakes made at some stage of reasoning can cause errors like context drift, inconsistent memory usage, conflicting observations, and bad decisions. Recent benchmarking also indicates that even today's state-of-the-art agentic models fail to reason with complicated instructions and multiple constraints in realistic settings (Qi et al., 2026).

Several reasoning paradigms have been proposed to improve autonomous agent reasoning. ReAct integrates reasoning with environment interaction (Yao et al., 2022), Tree of Thoughts (ToT) explores multiple reasoning branches through deliberate search (Yao et al., 2023), and Reflexion enhances reasoning via self-reflection and episodic memory without model retraining (Shinn et al., 2023). Despite these advances, reasoning stability remains a major challenge, as errors in intermediate reasoning states can propagate through subsequent reasoning trajectories (Yeo et al., 2025; Gan et al., 2025). Existing approaches mainly rely on iterative self-correction by repeatedly refining or regenerating reasoning (Madaan et al., 2023), yet their effectiveness varies across models, tasks, and feedback settings (Kamoi et al., 2024). Furthermore, although modern AI agents demonstrate strong planning, tool-use, and memory capabilities, current architectures lack mechanisms for diagnosing localized reasoning failures and selectively repairing only corrupted reasoning states without regenerating the entire reasoning trajectory (Wang et al., 2024). Consequently, a principled framework for reasoning distortion detection, reasoning-state diagnosis, and selective cognitive recovery remains absent. To address this gap, we propose ORDDAR (Observation-Driven Reasoning for Distortion-Resilient Decision, Action, and Cognitive Recovery).

To address these limitations, we propose ORDDAR (Observation-Driven Reasoning for Distortion-Resilient Decision, Action, and Cognitive Recovery), a novel observation-driven reasoning system which considers reasoning as a series of cognitive state transitions (Fig. 1). The system identifies local reasoning distortions, analyzes the failure in the reasoning state with a Cognitive Mirror by

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.

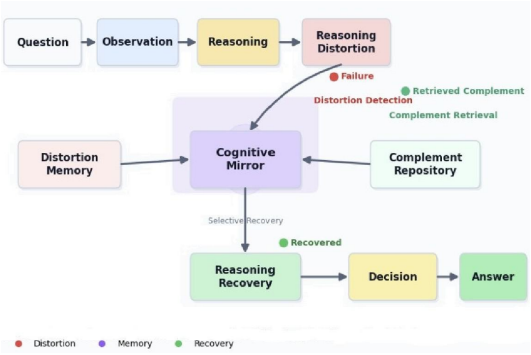


Figure 1: Overview of ORDDAR. The framework detects reasoning distortions, retrieves recovery complements through the Cognitive Mirror, selectively repairs corrupted reasoning states, and predicts the final decision.

combining distortion memory with complementary reasoning, and repairs only the faulty reasoning transitions without having to reconstruct the whole reasoning path. The key contributions of our study can be summarized as follows: First, we propose the ORDDAR model that treats reasoning as cognitive state transition and defines localized reasoning distortion. Second, we propose a Cognitive Mirror that integrates distortion memory and complementary reasoning retrieval to identify reasoning failures and recover only corrupted reasoning transition states. Third, we design an end-to-end reasoning pipeline that comprises of reasoning distortion detection, reasoning-state diagnosis, complementary reasoning retrieval, and selective cognitive recovery.

2 Related Work

With the fast progress in large language models, agents can now perform reasoning, planning, tool usage, integration of memory, and decision-making in changing environments (Wang et al., 2024; Xu et al., 2026). However, recent benchmarking efforts, such as AgentBench, TheAgentCompany, and AGENTIF, reveal that despite significant improvements, state-of-the-art agents still experience difficulties in following instructions, satisfying constraints, and performing consistent decision-making (Liu et al., 2024; Xu et al., 2026; Qi et al., 2026). Additionally, frameworks like OpenHands, AutoGen, and MetaGPT augment the capabilities of agents with techniques from software engineering, tool cooperation, and multi-agent coordination (Wang et al., 2025; Wu et al., 2024; Hong et al., 2024). Meanwhile, domain-specific benchmarks including MLAgentBench and ScienceAgentBench evaluate autonomous agents on machine learning and scientific discovery tasks, highlighting persistent challenges in reasoning robustness and consistency (Huang et al., 2023; Chen et al., 2025).

Iterative planning and self-correction have greatly contributed to the improvement of reasoning. ReAct combines reasoning with environment interactions, ToT explores various paths of reasoning, Reflexion and Self-Refine improve reasoning by reflecting and refining it (Yao et al., 2022; 2023; Shinn et al., 2023; Madaan et al., 2023). However, recently, researchers have found that longer reasoning paths do not always enhance reliability because intermediate reasoning errors tend to propagate and eventually contaminate subsequent reasoning steps (Yeo et al., 2025; Gan et al., 2025). In addition, current self-correction methods are extremely sensitive to the models, tasks, and feedback (Kamoi et al., 2024). As a result, most self-correction methods either generate or refine an entire reasoning path instead of fixing corrupted reasoning paths. Distortion of reasoning can be described as the presence of an unusual semantic shift between successive stages of cognition. Reasoning distortion does not preserve semantic integrity while introducing new information, but valid reasoning does so. Hence, semantic disparity is used as an indicator of localized reasoning distortion. The ORDDAR framework specifically identifies reasoning distortions and repairs corrupted reasoning paths instead of generating an entire reasoning path.

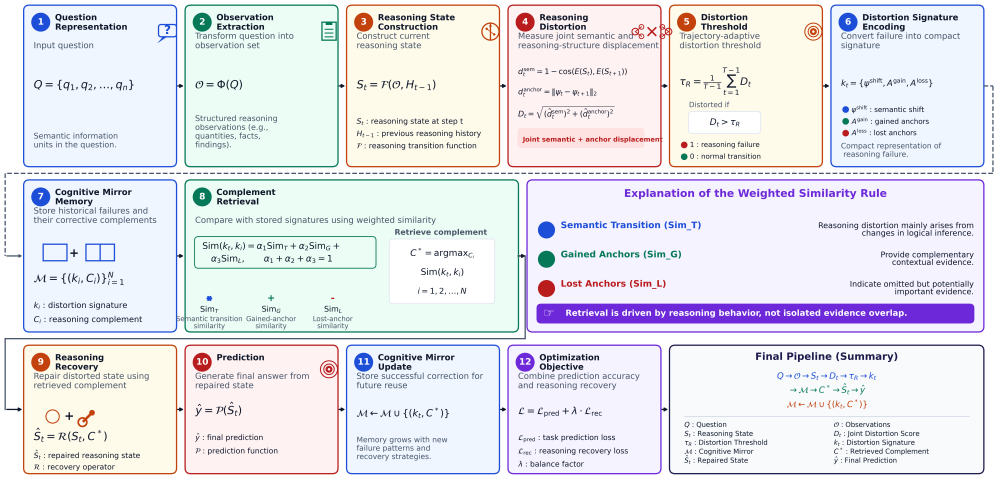


Figure 2: Overall mathematical workflow of ORDDAR. Starting from an input query, ORDDAR constructs observation-driven reasoning states, detects localized reasoning distortions, encodes compact distortion signatures, retrieves the most relevant recovery complement from the Cognitive Mirror using the proposed weighted similarity rule, selectively repairs distorted reasoning states, updates memory with successful recovery patterns, and performs the final prediction.

3 METHODOLOGY

A lot of advances have been made by autonomous AI agents in the areas of long-horizon planning, tool utilization, decision making, and lifelong learning, spanning a wide range of areas from embodied agents to web agents, software engineering to autonomous data science (Zhou et al., 2024; Zhang et al., 2025; Xie et al., 2024; Park et al., 2023; Luo et al., 2025; Wang et al., 2023). But their reasoning is prone to local errors that cause damage along their reasoning trajectories and compromise the ultimate decisions. In order to compensate for this weakness, ORDDAR proposes a new transition-based cognitive repair approach.

3.1 PROBLEM FORMULATION

Autonomous AI agents solve complex tasks through sequential reasoning, where observations are progressively transformed into cognitive states before decision making. Errors arising during these transitions propagate through subsequent reasoning and degrade final decisions. ORDDAR addresses this problem by recovering localized reasoning errors instead of regenerating the entire reasoning trajectory.

Given an input query $q \in \mathcal{Q}$, an autonomous AI agent generates an output $y \in \mathcal{Y}$ through sequential reasoning. ORDDAR models this process as a sequence of cognitive state transitions, where observations are progressively transformed into intermediate reasoning states. Rather than regenerating the entire reasoning trajectory after failure, ORDDAR identifies and selectively repairs localized reasoning distortions.

The input query is first transformed into a structured observation set

$$O = \phi(q) = \{o_1, o_2, \ldots, o_m\}, \tag{1}$$

where $\phi(\cdot)$ denotes the observation extraction function and o_i represents the i-th reasoning observation extracted from the input. Based on the observation sequence, the reasoning engine constructs a trajectory of cognitive states

$$R = \{s_1, s_2, \ldots, s_T\}, \tag{2}$$

where each cognitive state is recursively updated as

$$s_t = f(s_{t-1}, o_t), \tag{3}$$

with $f(\cdot)$ denoting the reasoning transition function. The objective of ORDDAR is to learn a recovery function

$$\mathcal{F} : R \to \hat{R}, \tag{4}$$

Where $\hat{R}$ represents the recovered reasoning path. Reasoning is defined as cognitive state changes driven by observations (Eqs. (2)–(3)). Intermediate reasoning can be captured and detected for distortion recovery by detecting reasoning distortions, acquiring complementary reasoning, and repairing cognitive states for prediction. For localized reasoning analysis, ORDDAR defines a cognitive window $W_t = \{s_t, s_{t+1}\}$ around the reasoning transition currently under inspection. The window moves sequentially along the trajectory, enabling distortion detection and recovery at the level of an individual transition while preserving the remaining reasoning states.

3.2 Reasoning Distortion Detection and Signature Encoding

ORDDAR detects localized reasoning failures by measuring the joint displacement of consecutive cognitive states in both semantic embedding space and cognitive-anchor space. Each cognitive state s_t is represented by a semantic embedding $E(s_t)$ and a normalized cognitive-anchor distribution ψ_t. For each transition $e_t = (s_t, s_{t+1})$, the semantic transition distance is defined as

$$d_t^{\text{sem}} = 1 - \cos\left(E(s_t), E(s_{t+1})\right), \tag{5}$$

where $E(\cdot)$ denotes the semantic embedding produced by a pretrained sentence encoder. We evaluate three off-the-shelf sentence encoders, namely MPNet, all-MiniLM-L6-v2, and BAAI/bge-base-en-v1.5, without task-specific fine-tuning. MPNet is adopted as the primary encoder based on its best preliminary retrieval performance, while MiniLM and BGE are evaluated to assess the robustness of the proposed distortion representation to the choice of semantic embedding space.

$$d_t^{\text{anchor}} = \left\|\psi_t - \psi_{t+1}\right\|_2. \tag{6}$$

Because these two distances capture complementary aspects of a reasoning transition, they are normalized within the trajectory and combined using a root-sum-of-squares formulation:

$$D_t = \sqrt{\left(\tilde{d}_t^{\text{sem}}\right)^2 + \left(\tilde{d}_t^{\text{anchor}}\right)^2}, \tag{7}$$

where $\tilde{d}_t^{\text{sem}}$ and $\tilde{d}_t^{\text{anchor}}$ denote the normalized semantic and anchor-space distances, respectively. The resulting joint distortion score D_t captures changes in both the semantic content and the reasoning structure of consecutive cognitive states. For each reasoning trajectory, we define a trajectory-adaptive distortion threshold as the mean joint distortion score across its consecutive state transitions:

$$\tau_R = \frac{1}{T-1} \sum_{t=1}^{T-1} D_t. \tag{8}$$

A transition is considered distorted when

$$z_t = \begin{cases} 1, & D_t > \tau_R, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Here, τ_R denotes the trajectory-adaptive distortion threshold, and z_t indicates whether the transition is identified as distorted. Only transitions with $z_t = 1$ are forwarded for signature encoding. Each detected distortion is encoded as

$$k_t = \{\psi^{\text{shift}}, A^{\text{gain}}, A^{\text{loss}}\}, \tag{10}$$

where ψ^{shift} denotes the semantic transition shift, A^{gain} represents newly introduced reasoning anchors, and A^{loss} denotes discarded reasoning anchors. The Shift–Gain–Loss signature characterizes the nature of the detected distortion and supports recovery-complement retrieval. The resulting signature enables efficient retrieval and cognitive recovery (Fig. 2). An alternative design is to employ an additional drafter or verifier LLM to assess the correctness of intermediate reasoning states, similar in spirit to verifier-based speculative decoding. We considered this possibility but do not introduce an additional LLM verifier into ORDDAR. Unlike speculative decoding, where a drafter proposes candidate tokens for subsequent verification, ORDDAR operates at the reasoning-state transition level and aims to provide a lightweight, model-independent signal for localizing reasoning distortions. Introducing an additional verifier LLM would add another model-dependent correctness signal and could confound the contribution of the proposed distortion detector. Instead, ORDDAR separates distortion localization from correctness evaluation and evaluates the detector independently using ground-truth reasoning-transition annotations.

3.3 Cognitive Mirror Memory

To support efficient reasoning recovery, ORDDAR maintains a *Cognitive Mirror* that stores compact distortion signatures and their corresponding recovery complements, enabling efficient retrieval without preserving complete reasoning trajectories. The Cognitive Mirror is represented as

$$\mathcal{M} = \{(k_i, c_i)\}_{i=1}^{N}, \tag{11}$$

where k_i denotes the encoded distortion signature of the i-th reasoning failure and c_i denotes its associated reasoning complement. The Cognitive Mirror stores reusable distortion signatures and recovery complements for efficient indexing and cross-task reuse. Unlike conventional retrieval-augmented reasoning, retrieval in ORDDAR is conditioned on a localized distortion signature and is used specifically to repair the corresponding reasoning transition rather than to augment the entire input context. Each memory entry is created from a detected distorted transition. For a distorted transition e_t, ORDDAR encodes its distortion signature k_t and associates it with a recovery complement c_t. The signature k_t serves as the retrieval index, while c_t represents the reusable recovery pattern. Thus, the Cognitive Mirror stores compact distortion–recovery pairs rather than complete reasoning trajectories.

3.4 Complement Retrieval

At runtime, the current cognitive window is encoded using the same distortion-signature representation. When a distortion is detected, its signature is used to retrieve the most compatible recovery complement from the Cognitive Mirror, which is applied only to the affected reasoning state. For a distortion signature k_t, ORDDAR retrieves the most relevant recovery complement from the Cognitive Mirror by computing similarity with stored signatures:

$$S(k_t, k_i) = \alpha_1 S_{\text{shift}} + \alpha_2 S_{\text{gain}} + \alpha_3 S_{\text{loss}}, \qquad \alpha_1 + \alpha_2 + \alpha_3 = 1, \quad \alpha_j \geq 0. \tag{12}$$

The weights are constrained to be non-negative and sum to one, where α_1, α_2, and α_3 control the relative contributions of transition shift, gain, and loss similarities respectively. Sensitivity analysis identifies $(\alpha_1, \alpha_2, \alpha_3) = (0.6, 0.2, 0.2)$ as the best-performing configuration, achieving the highest average signature matching score. We therefore adopt the 60–20–20 configuration for all subsequent experiments.

$$c^* = \arg \max_{(k_i, c_i) \in \mathcal{M}} S(k_t, k_i), \tag{13}$$

where c^* is the retrieved recovery complement.

3.5 Localized Reasoning Recovery

ORDDAR repairs the distorted state using the retrieved complement while preserving the rest of the reasoning trajectory, enabling localized recovery instead of complete regeneration. The restored reasoning state can be obtained as

$$\hat{s}_t = \mathcal{R}(s_t, c^*), \tag{14}$$

where s_t denotes the distorted reasoning state, c^* is the retrieved reasoning complement obtained from the Cognitive Mirror, and $\mathcal{R}(\cdot)$ denotes the complement-guided recovery operator.

The recovered reasoning trajectory is

$$\hat{R} = \{s_1, \ldots, \hat{s}_t, \ldots, s_T\}, \tag{15}$$

where only the distorted reasoning state is replaced while the remaining reasoning trajectory is preserved. This localized recovery avoids unnecessary regeneration of valid reasoning states.

3.6 Prediction and Cognitive Mirror Update

The recovered reasoning trajectory is used for final prediction:

$$\hat{y} = P(\hat{R}), \tag{16}$$

where $\hat{R}$ denotes the recovered reasoning trajectory and $P(\cdot)$ is the prediction function. Following localized recovery, the recovered trajectory is validated before the Cognitive Mirror is updated:

$$\mathcal{M} \leftarrow \mathcal{M} \cup \{(k_t, c^*)\}, \tag{17}$$

where k_t is the current distortion signature and c^* is the retrieved recovery complement. Only validated recovery patterns are eligible for online memory updates. During evaluation, the current test instance is queried against the existing Cognitive Mirror and is not inserted into memory before prediction, preventing evaluation-time contamination. To prevent data contamination, evaluation trajectories are not inserted into the Cognitive Mirror before prediction; only validated recovery patterns are eligible for subsequent memory updates. The complete Cognitive Mirror workflow, including memory construction and indexing, local cognitive-window formation, distortion-signature retrieval, localized reasoning recovery, and validated online memory update, is summarized in Algorithm 1.

The workflow is task-agnostic: although representations and validation may vary across tasks, cognitive-window construction, distortion-trace encoding, complement retrieval, and recovery remain unchanged. The Cognitive Mirror is initialized from validated trajectories; recovery occurs only when a compatible validated complement exists, otherwise the trajectory is preserved. Successful recovery may update memory for future queries, while evaluation instances are excluded before prediction.

4 Experimental Setup

Experiments are conducted on four reasoning benchmarks: GSM8K, HotpotQA, StrategyQA, and CKDQA, covering mathematical, multi-hop, commonsense, and clinical reasoning, respectively. All datasets are publicly available except CKDQA, which was constructed from hospital cases following ethical clearance. For fair comparison, all methods use Qwen2.5-7B-Instruct with its standard tokenizer and chat template, deterministic greedy decoding (`do_sample=False`), and a maximum generation length of 512 tokens. Each evaluation uses 100 test instances sampled with random seed 42. For semantic retrieval, ORDDAR evaluates MPNet, all-MiniLM-L6-v2, and BAAI/bge-base-en-v1.5, with MPNet selected based on preliminary retrieval performance; independent ablations evaluate the contributions of the individual ORDDAR components.

5 Results

5.1 Overall Performance

Comparison of ORDDAR with CoT, Self-Refine, Reflexion, Tree of Thoughts (ToT), and ProCo is provided in Table 1. ORDDAR achieves the highest reported performance among the compared methods. The difference between the baselines suggests that the reasoning refinement process is highly dependent on task-specific features and intermediate errors. Whereas CoT and Self-Refine depend on a consistent reasoning flow, Reflexion, ToT, and ProCo may still be influenced by intermediate errors or ineffective corrections. ORDDAR identifies reasoning distortions and retrieves complementary reasoning patterns to repair the affected reasoning state.

The higher performance on CKDQA suggests that localized errors in clinical findings or staging criteria can have a significant influence on the final decision. In contrast, HotpotQA features multi-hop

Algorithm 1 Cognitive Mirror Construction, Trace-Guided Retrieval, and Recovery

Require: Validated trajectories $\mathcal{D}$, query q, threshold $\tau_{\mathcal{R}}$, weights $\alpha_1, \alpha_2, \alpha_3$
Ensure: Prediction $\hat{y}$
Initialize Cognitive Mirror $\mathcal{M} \leftarrow \emptyset$
Memory Construction
for each validated trajectory $\mathcal{R}_j = \{s_1, \ldots, s_T\} \in \mathcal{D}$ **do**
  for $t = 1$ to $T - 1$ **do**
    $W_t \leftarrow \{s_t, s_{t+1}\}$; compute joint distortion D_t
    if $D_t > \tau_{\mathcal{R}_j}$ **then**
      $k_t \leftarrow \{\psi_{\text{shift}}, A_{\text{gain}}, A_{\text{loss}}\}$
      Store validated pair (k_t, c_t) in $\mathcal{M}$
    end if
  end for
end for
Runtime Detection and Recovery
Generate $\mathcal{R} = \{s_1, \ldots, s_T\}$ from q
for $t = 1$ to $T - 1$ **do**
  $W_t \leftarrow \{s_t, s_{t+1}\}$; compute joint distortion D_t
  if $D_t > \tau_{\mathcal{R}}$ **then**
    Encode distortion trace $k_t \leftarrow \{\psi_{\text{shift}}, A_{\text{gain}}, A_{\text{loss}}\}$
    Search $\mathcal{M}$ for a compatible distortion trace and retrieve its complement c^* using $\alpha_1 S_{\text{shift}} + \alpha_2 S_{\text{gain}} + \alpha_3 S_{\text{loss}}$
    if a compatible trace–complement pair exists **then**
      $\hat{s}_t \leftarrow R(s_t, c^*)$
      Replace only the affected state
    else
      Preserve the current reasoning trajectory
    end if
  end if
end for
$\hat{\mathcal{R}} \leftarrow$ recovered reasoning trajectory
$\hat{y} \leftarrow P(\hat{\mathcal{R}})$
Validated Memory Update for Future Queries
if recovery is validated and memory updating is permitted **then**
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{(k_t, c^*)\}$
end if
return $\hat{y}$

Table 1: Performance comparison of CoT, Self-Refine, Reflexion, ToT, ProCo, and ORDDAR. Values denote accuracy (%) unless otherwise indicated. For HotpotQA, both EM and F1 are reported.

Dataset	CoT	Self-Refine	Reflexion	ToT	ProCo	ORDDAR
GSM8K	81	87	80	–	50	92
HotpotQA (EM)	63	66	24	–	–	67
HotpotQA (F1)	75	78	30	–	–	79.22
StrategyQA	81	80	60	50	70	86
CKDQA	67	60	53.33	30	25	94.38

evidence dependencies, resulting in more distributed reasoning errors. This indicates that refinement effectiveness depends on both the error type and the nature of the reasoning task. ORDDAR addresses these inconsistencies through localized state detection and recovery.

5.2 Ablation Study

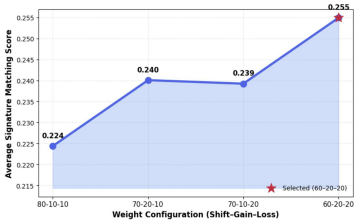
The contributions of the Complement Memory (CM), Distortion Memory (DM), and Recovery Memory (RM) are evaluated through independent ablations (Table 3). Removing any component consistently degrades performance, confirming their complementary roles. On GSM8K, removing DM causes the largest drop (92%→85%), highlighting the importance of accurate distortion localization. On StrategyQA, removing CM or DM reduces accuracy from 86% to 81%, while removing

Table 2: Comparison of embedding models used for Cognitive Mirror retrieval. Values denote accuracy (%) unless otherwise indicated. For HotpotQA, F1 is reported.

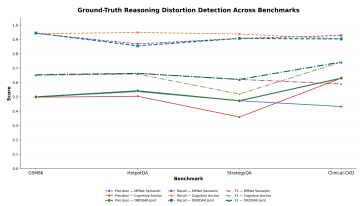
Dataset	MPNet	MiniLM	BGE
GSM8K	92	91	92
HotpotQA (F1)	79.22	79	79.1
StrategyQA	86	85	85
CKDQA	94.38	87	87

Table 3: Ablation study of the Complement (CM), Distortion (DM), and Recovery (RM) mechanisms. Values denote accuracy (%) unless otherwise indicated. For HotpotQA, both EM and F1 are reported.

Dataset	ORDDAR	w/o CM	w/o DM	w/o RM
GSM8K	92	90	85	88
HotpotQA (EM)	67	65	64	62.5
HotpotQA (F1)	79.22	79.0	71.0	74.03
StrategyQA	86	81	81	83.3
CKDQA	94.38	81	78.05	79.79



(a) Validation of the 60–20–20 weighting.



(b) Ground-truth distortion detection.

Figure 3: Validation of the proposed distortion representation and its ground-truth detection performance.

RM yields 83.3%. Similarly, on HotpotQA, removing CM reduces EM from 67% to 65%, DM reduces F1 from 79.22% to 79.0%, and RM produces the lowest EM (62.5%). The largest degradation occurs on CKDQA, where removing CM, DM, and RM reduces performance from 94.38% to 81%, 78.05%, and 79.79%, respectively. These results show that complement retrieval, distortion localization, and localized recovery each contribute to ORDDAR's effectiveness.

5.3 Ground-Truth Validation of Similarity-Based Distortion Detection

To validate the proposed distortion representation, we examine the relative contribution of transition similarity, logical gain, and logical loss. Sensitivity analysis identifies the 60–20–20 weighting as the best-performing configuration, assigning 60% to transition similarity and 20% each to gain and loss, as shown in Fig. 3(a). We evaluate embedding-based distortion detection on an independently labeled GT set across GSM8K, HotpotQA, StrategyQA, and Clinical-CKD. Table 4 reports precision, recall, and F1 for MPNet Semantic, Cognitive Anchor, and ORDDAR Joint. The cross-benchmark results in Fig. 3(b) show that similarity provides a meaningful distortion signal, with anchor-based representations providing complementary evidence.

5.4 Qualitative Case Study and Retrieval Analysis

A representative Clinical-CKD recovery case is shown in Fig. 4(a). For a patient with creatinine 3.6 mg/dL and eGFR 16.6 mL/min/1.73m^2, the initial reasoning incorrectly assigns G3, whereas 15–29 corresponds to G4 under KDIGO (Stevens et al., 2024). ORDDAR localizes the distortion

Table 4: Ground-truth validation of reasoning distortion detection. P, R, and F1 denote precision, recall, and F1-score, respectively.

Benchmark	MPNet Semantic			Cognitive Anchor			ORDDAR Joint		
	P	R	F1	P	R	F1	P	R	F1
GSM8K	.499	.942	.652	.498	.940	.651	.500	.945	.654
HotpotQA	.537	.869	.663	.504	.949	.659	.542	.855	.664
StrategyQA	.473	.907	.622	.359	.938	.520	.473	.907	.622
Clinical-CKD	.433	.928	.590	.630	.904	.743	.630	.904	.743



(a) Clinical-CKD reasoning recovery case.

(b) Cognitive Mirror retrieval space.

Figure 4: Qualitative analysis of localized reasoning recovery and Cognitive Mirror retrieval.

(0.874), retrieves a Cognitive Mirror complement (0.940), and repairs only the erroneous transition, consistent with prior findings on intrinsic self-correction limitations (Tyen et al., 2024; Kamoi et al., 2024). The corresponding UMAP visualization in Fig. 4(b) shows that retrieved complements remain semantically close to detected distortions, suggesting that distortion signatures preserve useful information for recovery-pattern retrieval.

6 Conclusion

ORDDAR was introduced as an observation-based framework for detecting localized reasoning distortions and performing cognitive recovery. By representing reasoning as a sequence of cognitive state transitions, ORDDAR detects distorted transitions, retrieves their complements from the Cognitive Mirror, and recovers only the affected states while preserving correct reasoning. Experiments across mathematical, commonsense, multi-hop, and clinical reasoning consistently demonstrate the superiority of ORDDAR over the evaluated reasoning baselines, particularly on CKDQA. Ablation results confirm that distortion detection, complement retrieval, and recovery each contribute to the final performance. Retrieval analysis further demonstrates that distortion signatures can identify relevant recovery patterns. Overall, the results show that targeted cognitive recovery can provide an effective alternative to complete reasoning regeneration, while handling higher-order multi-hop dependencies remains an important direction for future research. In the current clinical setting, recovery complements incorporate locally provided KDIGO knowledge. Future work will extend this design to a hybrid recovery scheme combining Cognitive Mirror retrieval with dynamically retrieved clinical knowledge for broader and more adaptive domain-specific recovery.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and limited code assistance during the preparation of this manuscript. These tools were not used to generate experimental results, figures, perform the reported evaluations, or make final scientific decisions. All AI-assisted outputs were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all claims, analyses, results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The clinical data used in this study were collected under the appropriate institutional and hospital ethical approvals. The study procedures followed the ethical requirements and guidelines of the participating institution and hospital. Appropriate measures were taken to protect patient privacy and confidentiality, and no personally identifiable information is reported in the manuscript or supplementary materials. The CKDQA dataset is private and is therefore not publicly released.

REPRODUCIBILITY STATEMENT

We provide the implementation and configuration details required to reproduce the reported experiments. The source code covers the ORDDAR framework, reasoning distortion detection, Cognitive Mirror retrieval, localized recovery, and the evaluation procedures across GSM8K, HotpotQA, StrategyQA, and CKDQA. The accompanying `README.txt` and `requirements.txt` document the required environment, dependencies, and execution procedure. Code for all benchmark experiments, including CKDQA, can be provided as anonymous supplementary material.

AUTHOR CONTRIBUTIONS

The authors contributed to the conceptualization, methodology, experiments, analysis, and preparation of the manuscript.

ACKNOWLEDGMENTS

We sincerely thank the clinicians and doctors who supported this work by providing the clinical data and valuable domain expertise required for the CKDQA experiments. We also gratefully acknowledge the AI research community for valuable knowledge, insights, and discussions on AI agents that contributed to the development of the ideas explored in this work. We further thank everyone who provided helpful discussions and feedback during the development of this research.

REFERENCES

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pp. 96934–96990, 2025.

Zeyu Gan, Yun Liao, and Yong Liu. Rethinking external slow-thinking: From snowball errors to probability of correct reasoning. *arXiv preprint arXiv:2501.15602*, 2025.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2023.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K Qiu, and Yuqing Yang. Agent lightning: Train any ai agents with reinforcement learning. *arXiv preprint arXiv:2508.03680*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, Lei Hou, and Juanzi Li. Agentif: Benchmarking large language models instruction following ability in agentic scenarios. *Advances in Neural Information Processing Systems*, 38, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Paul E Stevens, Sofia B Ahmed, Juan Jesus Carrero, Bethany Foster, Anna Francis, Rasheeda K Hall, Will G Herrington, Guy Hill, Lesley A Inker, Rümeyza Kazancıoğlu, et al. Kdigo 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney international*, 105(4):S117–S314, 2024.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Yuanzhu Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908, 2024.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023. *URL https://arxiv. org/abs/2305.16291*, 2(11), 2023.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations*, volume 2025, pp. 65882–65919, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Frank Fangzheng Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. *Advances in Neural Information Processing Systems*, 38, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

Shaolei Zhang, Ju Fan, Meihao Fan, Guoliang Li, and Xiaoyong Du. Deepanalyze: Agentic large language models for autonomous data science. *arXiv preprint arXiv:2510.16872*, 2025.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, volume 2024, pp. 15585–15606, 2024.