

CoMATE: Consolidating Memory Across Trajectories for Long-Horizon Interactive Exploration on ARC-AGI-3

Anonymous authors
Paper under double-blind review

Abstract

As agents take on increasingly complex and demanding tasks, reliable long-horizon performance is becoming essential but remains difficult to achieve. This difficulty is evident in long-horizon interactive exploration, where ARC-AGI-3 remains very hard for frontier models in official direct-play evaluations. Parallel exploration at test time can improve performance by generating complementary experience, but shared notes alone do not establish which claims are supported, when they apply, or how conflicting accounts should be reconciled. We introduce CoMATE, an evidence-grounded memory consolidation framework. A fresh-context model examines agents' written memories and observer-view trajectories, producing collective entries with supporting evidence, epistemic status, and applicability conditions. Agents query this knowledge during subsequent exploration. Because results on ARC-AGI-3 depend heavily on the agent harness, we compare memory treatments under a shared parallel search protocol. Level-wise branching, checkpoint handoff, and retries are common to all conditions, so the no-consolidation baseline also retains a full parallel group at each new level. Under this protocol, CoMATE raises RHAE from 55.1 to 80.1 with six GPT-5.5 agents on the 25 public games, completing 22 games and 171 of 183 levels; nine agents reach 80.9 RHAE and 173 levels. With GPT-5.6 Sol, it raises RHAE from 87.6 to 97.0 and completes all 25 games and 183 levels. Action- and token-budget curves account for both gameplay and memory-related inference, while automated and human audits separately assess memory correctness and citation support. These results support cross-trajectory consolidation on ARC-AGI-3 and its potential for broader long-horizon interactive exploration.

1 Introduction

Agents are increasingly expected to carry out complex, demanding tasks rather than provide isolated responses. Progress in foundation models (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023; Touvron et al., 2023; Gemini Team, 2024; Grattafiori et al., 2024; Yang et al., 2024a; DeepSeek-AI, 2024) and their use in reasoning and tool-based interaction (Yao et al., 2023b; Schick et al., 2023; Wang et al., 2023a; Yang et al., 2024b) has expanded their applications. Reliable long-horizon performance requires more than handling each step well: agents must preserve useful experience, revise mistaken interpretations, and sustain progress across interdependent actions. Earlier observations and assumptions shape later decisions, complicating sustained progress.

The difficulty is especially apparent in long-horizon interactive exploration, where an agent must understand an unfamiliar environment while deciding how to act. Early observations may support competing explanations, allowing mistaken assumptions to persist through many actions. Extending an attempt can therefore prolong exploration under the same incomplete understanding. ARC-AGI-3 remains very hard for frontier models in official direct-play evaluations (ARC Prize Foundation, 2026; ARC Prize, 2026d). On the 25 public games used in this paper, ARC Prize's own runs with its standard single-agent harness score 0.73% RHAE for GPT-5.5 at high reasoning effort, completing 5 of 183 levels (ARC Prize, 2026b), and 13.33% for GPT-5.6 Sol at max effort, or about 5.2% at high

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.

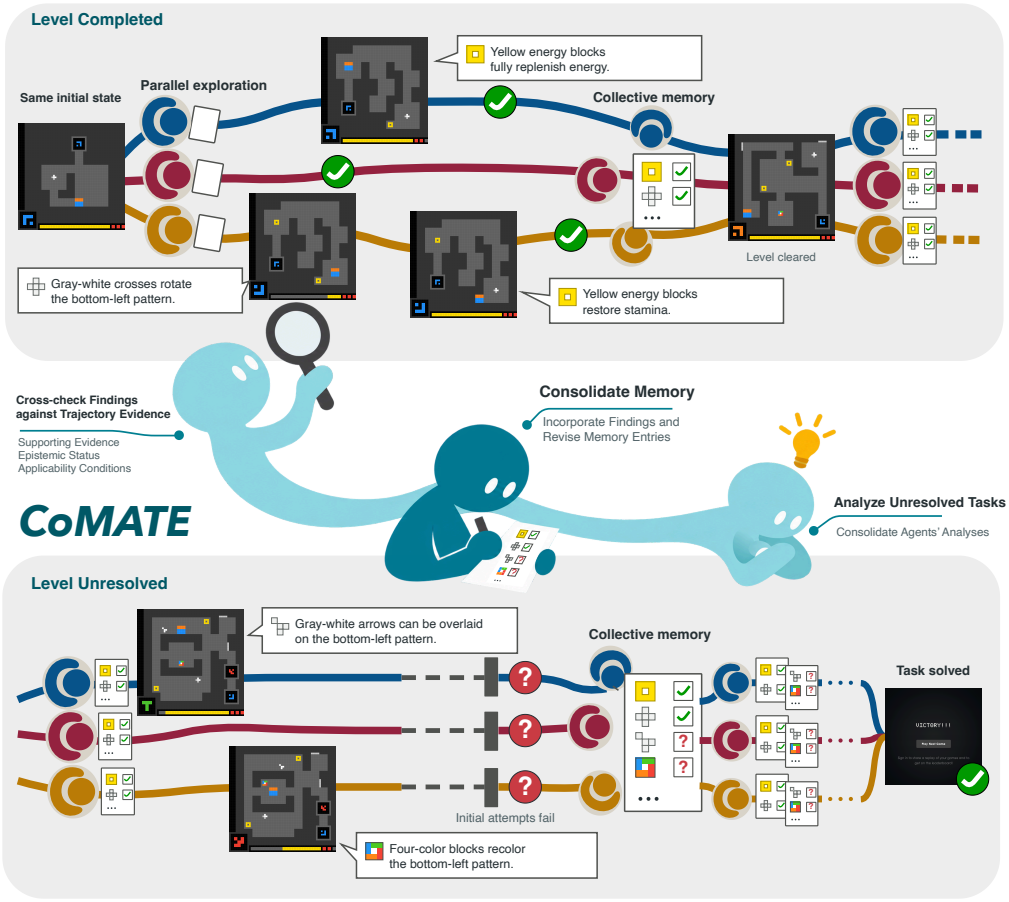


Figure 1: **Overview of CoMATE.** The shared level-wise search scaffold starts agents from the same checkpoint and hands a successful state forward to the next level. Agents develop their own trajectories and individual memories. CoMATE adds memory consolidation, which cross-checks their findings against recorded trajectory evidence and incorporates complementary knowledge into collective memory, with supporting evidence and epistemic status for each entry. Unresolved entries prioritize the best-performing agent's interpretation while preserving uncertainty. All agents can retrieve relevant entries. If all initial attempts end without solving the level, the same process consolidates their analyses into collective memory to guide every agent's further exploration.

effort (ARC Prize, 2026c) (Appendix A). Such results depend on the agent harness as well as the model. We therefore study the additional value of memory consolidation under an enhanced base harness and a parallel search protocol shared across the main comparisons, so that the evaluation centers on memory treatment rather than leaderboard standing.

Parallel exploration offers one response at test time. By pursuing different plans and testing different hypotheses, agents can generate diverse candidate solutions (Wang et al., 2023b; Yao et al., 2023a; Snell et al., 2024). A level-wise search procedure can advance the whole group when one agent succeeds, preserving parallel exploration at later levels. Such a state handoff does not, however, reconcile what different agents learned. Their records can contain complementary observations and conflicting interpretations whose validity depends on different conditions. In LS20's second level, the best-performing agent reaches the exit without testing an energy block, leaving its replenishment mechanism undiscovered. This omission leads to repeated unsuccessful attempts in the third level before the agent rediscovers the mechanism. Other agents had already observed the effect, and CoMATE can consolidate their evidence to guide subsequent exploration. Passing a level validates progress, not the completeness of every agent's understanding.

Experience-derived knowledge and evolving memory already support agents, as illustrated by ExpeL and A-MEM (Zhao et al., 2024; Xu et al., 2025). ReasoningBank, ACE, and G-Memory accumulate memory across trajectories or agents, and the Symbolica harness we build on shares self-labeled memory among agents (Ouyang et al., 2026; Zhang et al., 2026b; 2025; Symbolica, 2026b). In these methods, whether a claim counts as established is decided by its writer, task outcomes, or usefulness signals rather than by checking it against recorded observations (Appendix B). The question here is not simply whether agents should share information, but how evidence-grounded consolidation can improve on making their notes jointly visible or compressing them into an unstructured summary. Agents may describe related observations differently, or reach apparently conflicting conclusions under different conditions. Their accounts need to be connected to the evidence that supports them before the resulting knowledge can guide further exploration. Memory must therefore distinguish observations from interpretations and make clear which findings remain unresolved and under what conditions they apply. We introduce CoMATE to meet this need through memory consolidation. It cross-checks individual memories against recorded interactions to fill gaps in agents' understanding. Collective entries retain evidence, epistemic status, and applicability conditions for later use by all agents. If initial exploration remains unresolved, the same process updates collective memory to guide further attempts (Figure 1).

Our contributions are threefold.

- We study cross-trajectory memory organization on ARC-AGI-3 under a shared level-wise search scaffold, distinguishing memory treatment from branching and state-handoff advantages.
- We introduce CoMATE, which consolidates individual memories and trajectory evidence into revisable collective memory with explicit epistemic status and applicability conditions.
- Under the shared search protocol on all 25 public ARC-AGI-3 games and 183 levels, CoMATE raises RHAE from 55.1 to 80.1 with six GPT-5.5 agents, completing 22 games and 171 levels. Nine agents reach 80.9 RHAE. With GPT-5.6 Sol, CoMATE raises RHAE from 87.6 to 97.0, completing 25 games and 183 levels. Resource curves include memory-related LM calls, and direct memory audits complement the task-level results. These ARC-AGI-3 findings may inform broader long-horizon interactive exploration.

2 CoMATE

2.1 Overview and Formulation

CoMATE augments a shared level-wise search scaffold with memory consolidation (Figure 1). At level l, N gameplay agents explore from a common entry checkpoint. A successful state is handed forward, and the next level again receives N branches. The no-consolidation baseline shares the branching, handoff, and retry trigger; memory processing adds no gameplay branches.

Agents record findings and questions in individual memory during exploration. Level completion triggers consolidation, anchored by environment feedback. For agent $i \in \{1, \ldots, N\}$, let $\tau_l^{(i)}$ be its recorded observation–action trajectory and $W_l^{(i)}$ its individual memory. Consolidation updates the existing collective memory K_l using these paired records:

$$\mathcal{B}_l = \{(\tau_l^{(i)}, W_l^{(i)})\}_{i=1}^{N}, \qquad K_l^{+} = \mathrm{Consolidate}(K_l, \mathcal{B}_l). \tag{1}$$

The superscript $+$ denotes a revised memory version available to all agents. The same operation also processes unresolved experience before further attempts.

2.2 Evidence-Grounded Cross-Trajectory Memory

A separate, fresh-context instance of the gameplay backbone executes consolidation. It receives the agents' notes and an observer-view trajectory containing visible observations, actions, and feedback, but excluding executed code and internal reasoning traces. Long records are processed in segments to extract observations and their evidence references, then aggregated and compared across agents. The consolidator examines existing evidence without conducting new environment tests. Related claims are checked against recorded actions, conditions, and outcomes, not agreement between generated accounts alone. Consistent evidence can support a finding; conflicting evidence can restrict its scope

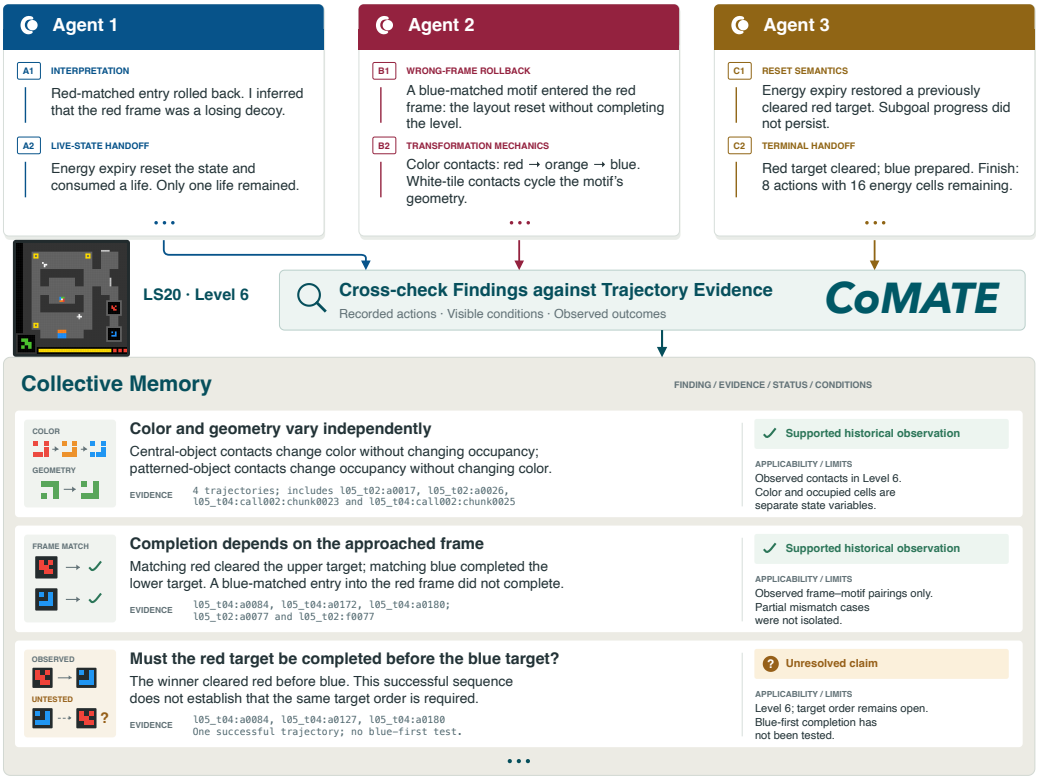


Figure 2: **Memory organization and use in CoMATE.** Selected individual memories and collective entries from LS20's sixth level connect complementary discoveries to trajectory evidence. Entries record findings, supporting evidence, epistemic status, and applicability conditions. A successful target order remains an observation, not an established requirement.

or leave it unresolved. Each memory entry is

$$e = (\phi, \mathcal{E}, s, \sigma), \tag{2}$$

where ϕ is the finding, $\mathcal{E}$ its supporting evidence, s its epistemic status, and σ its applicability conditions. These fields form one coupled representation rather than independent annotations: $\mathcal{E}$ comes from checking the claim against recorded trajectories, s is assessed against $\mathcal{E}$, and σ records the scope within which $\mathcal{E}$ supports ϕ. Entries convey mechanisms and their limits, not prescribed action sequences (Figure 2). For unresolved findings, the best-performing agent's interpretation receives provisional priority but cannot override contradictory trajectory evidence. Later observations can revise the entry as a whole while earlier versions and supporting records remain available.

2.3 Memory-Guided Exploration

Collective memory has no fixed entry-count limit and is queried rather than loaded in full. When a question remains after inspecting the current observation and individual notes, an agent supplies an action trigger, visible subject, and effect type. Structured and lexical filtering select bounded candidates; a stateless, read-only LLM retriever then collates relevant evidence, preserving conflicts and uncertainty. No match produces an explicit failure. Retrieved material describes historical findings rather than action plans, and the agent judges its applicability against the current observation. Several queries may precede an environment action. If all initial attempts end without solving the level, the same consolidation process extracts information from their unsuccessful experience. The level then receives one further attempt from its entry checkpoint, with updated collective memory and up to 1000 further actions per agent. This uses prior evidence to guide continued exploration while retaining the common branching and state-handoff mechanism.

Table 1: **Main results on ARC-AGI-3.** The six-agent no-consolidation baseline (B6) and CoMATE share the base harness and level-wise search protocol. Results are paired by backbone. All conditions use 500 initial actions per agent per level; a level all six agents fail receives one retry, with per-agent budgets shown. ∞^* marks a retry run until the agent concedes, not an unlimited inference budget.

Model	Method	Retry budget	Score	Levels	Clears
GPT-5.4	B6	∞^*	25.2/36.4	71/183 (51)	4/25 (2)
	CoMATE	1000	33.3/45.0	87/183 (69)	5/25 (4)
GPT-5.5	B6	∞^*	55.1/67.5	142/183 (92)	13/25 (6)
	CoMATE	1000	80.1/85.9	171/183 (132)	22/25 (12)
GPT-5.6 Sol	B6	∞^*	87.6/88.7	177/183 (132)	23/25 (14)
	CoMATE	1000	97.0/97.7	183/183 (164)	25/25 (23)

3 Experiments

3.1 Experimental Setup

We evaluate CoMATE on the 25 games and 183 levels of ARC-AGI-3's Public Demonstration Set (ARC Prize Foundation, 2026). The benchmark report presents this set as a community-facing entry point that does not cover every mechanic of the private set, which is designed to test generalization rigorously. We developed CoMATE blind to individual games: all 25 games use the same game-agnostic method, prompts, and protocol, with no per-game adjustment and no game-specific heuristics or mechanism rules. Development and evaluation used no game source code, recorded replays, or outside information: agents and harness code reach each game only through its environment interface, without web access or artifacts from previous runs, and every run starts from empty memory. Since B6 and CoMATE share each backbone, any pretraining exposure to these games affects both alike. Gameplay runs with GPT-5.4, GPT-5.5, and GPT-5.6 Sol all use `high` reasoning effort. The shared base harness extends Symbolica's framework (Symbolica, 2026a;b) with multimodal visual processing, basic memory access, and historical-state queries. It preserves visible frames for later inspection and controls action execution at level, terminal-state, and budget boundaries. These facilities and gameplay context handling are held fixed across memory conditions.

Main comparisons use six agents, and every condition gives each agent 500 initial actions per level. A level that all six agents fail receives one retry from its entry checkpoint, under the same concession rule in every condition. B6 retries afresh, without shared memory, and continues until each agent concedes, sometimes using thousands of actions (∞^*); CoMATE and the sharing controls retry with their shared memory and at most 1000 actions per agent beyond the initial 500. This cap is the only budget difference, so B6's retry is never more constrained than CoMATE's. These are environment-action allowances, not inference budgets; only the width study varies N.

3.2 Evaluation Metrics

We report Relative Human Action Efficiency (RHAE), the action-efficiency metric defined by ARC-AGI-3, alongside Linear Action Efficiency (LAE). Both score each completed level by the actions of the agent that completes it; for a level completed on retry, this count also includes that agent's unsuccessful initial attempt. LAE assigns a completed level $\min(100h/a, 100)$, where h is the reference action count and a this action count; unfinished levels receive zero. LAE averages levels equally within each game, then games equally. Score lists RHAE/LAE, with gray LAE. All scores are computed on the public games under our parallel protocol and serve as a controlled comparison of memory treatments, not as official benchmark performance. Official scoring counts the actions taken on first exposure to each game (ARC Prize Foundation, 2026), and competition mode allows only one `make` call per game and a single scorecard for the whole run (ARC Prize, 2026a). Our protocol instead runs parallel agents with checkpoint handoff and retries. Actions of the other agents do not enter RHAE or LAE, although their tokens are included in the resource accounting below.

Levels and Clears report completed levels and fully completed games as completed/total. Parentheses count full-score solutions (RHAE ≥ 100). Δ Levels and Δ Final Score include cleared blockers and

Table 2: **Detailed GPT-5.5 results on ARC-AGI-3.** B6 and CoMATE use six agents with the same level-wise branching, checkpoint handoff, and retry trigger. B6 omits cross-trajectory memory consolidation while retaining the same level-wise search structure. #L denotes the number of levels in each game. Green cells indicate completed levels with RHAE ≥ 100, and Yellow cells indicate completed levels below that threshold; both show action counts. F marks an unfinished level. Score lists RHAE/LAE. Within each game, the best Solved and score values are underlined.

Game	#L	B6												CoMATE											
		Solved	Level actions										Score	Solved	Level actions										Score
LP85	8	8/8	7	11	17	25	18	49	9	53			100.0/95.5	8/8	8	13	20	13	11	20	5	27			100.0/100.0
SB26	8	8/8	9	15	15	15	17	19	17	17			100.0/100.0	8/8	9	24	15	15	17	19	17	17			100.0/100.0
SU15	9	9/9	11	77	19	17	70	22	11	44	197		63.8/76.7	9/9	11	26	21	18	64	10	7	28	32		100.0/95.1
TU93	9	9/9	23	16	19	33	42	28	18	21	108		100.0/95.6	9/9	18	14	19	18	38	43	15	23	59		100.0/99.3
AR25	8	8/8	15	22	64	22	29	62	42	47			100.0/100.0	8/8	17	15	50	23	32	64	125	75			100.0/99.7
KA59	7	7/7	14	50	52	59	42	250	319				76.4/88.0	7/7	15	55	42	51	27	71	179				100.0/100.0
CN04	6	6/6	18	50	67	39	88	57					100.0/100.0	6/6	18	33	30	41	51	61					100.0/100.0
FT09	6	6/6	7	7	14	27	62	21					100.0/100.0	6/6	4	7	14	16	21	13					100.0/100.0
M0R0	6	5/6	25	43	87	32	74	F					71.4/80.2	6/6	25	33	95	15	51	51					100.0/100.0
SP80	6	4/6	12	41	27	47	F	F					47.6/65.4	6/6	20	35	20	35	89	80					100.0/100.0
TR87	6	6/6	35	80	28	93	317	166					54.7/71.9	6/6	29	33	32	21	30	24					100.0/100.0
CD82	6	6/6	15	28	64	26	31	37					48.7/68.3	6/6	15	15	32	14	13	16					100.0/92.2
RE86	8	6/8	25	36	54	78	139	161	F	F			58.3/73.3	8/8	28	36	49	58	63	67	197	396			96.9/94.2
VC33	7	4/7	10	13	28	340	F	F	F				22.7/41.1	7/7	7	12	28	29	58	70	49				94.9/92.7
SC25	6	6/6	23	5	20	132	312	109					51.4/75.8	6/6	19	5	29	116	162	39					94.0/93.3
TN36	7	7/7	9	13	27	113	20	65	61				85.8/88.0	7/7	14	15	10	28	37	36	97				87.7/92.1
DC22	6	4/6	32	47	48	155	F	F					40.5/60.5	6/6	38	45	59	67	634	698					80.6/89.0
WA30	9	7/9	44	154	80	340	523	381	46	F	F		37.4/54.9	9/9	40	370	88	66	482	112	60	505	1034		67.0/77.4
G50T	7	7/7	47	47	70	63	164	238	204				51.0/73.4	7/7	44	82	132	67	69	161	895				64.2/77.3
SK48	8	3/8	21	309	52	F	F	F	F	F			14.6/32.2	8/8	23	61	89	206	450	264	115	615			56.3/73.1
R11L	6	1/6	6	F	F	F	F	F					4.8/16.7	6/6	6	29	20	31	84	325					56.0/76.8
LF52	10	4/10	11	40	82	72	F	F	F	F	F	F	16.3/37.2	6/10	11	45	55	66	168	187	F	F	F	F	38.2/57.9
LS20	7	4/7	17	117	175	390	F	F	F				14.5/37.6	7/7	17	101	105	105	412	488	398				36.4/65.6
S5I5	8	2/8	18	100	F	F	F	F	F	F			7.6/23.6	3/8	18	44	82	F	F	F	F	F			16.7/37.5
BP35	9	5/9	22	48	77	228	304	F	F	F	F		9.0/31.1	6/9	37	61	92	207	37	537	F	F	F		14.5/34.1
All	183	**142/183**	**13/25 clears**										**55.1/67.5**	**171/183** (+29)	**22/25 clears** (+9)										**80.1** (+25.0)/**85.9** (+18.4)

subsequent progress through the end of continuation, with the latter using game-level aggregation. We report actions and total input-plus-output tokens, not monetary cost. Game-solving tokens include gameplay, failed attempts, consolidation, trajectory analysis, and retrieval.

3.3 Main Results

B6 is our six-agent no-consolidation baseline. It retains CoMATE's level-wise search scaffold and base memory facilities. Table 1 pairs backbones; Table 2 details the GPT-5.5 results. CoMATE improves coverage and solution quality. With GPT-5.5, it reaches 80.1 RHAE and 85.9 LAE, compared with 55.1 and 67.5 for the baseline. Completion rises from 142 to 171 levels and from 13 to 22 full game clears, while levels reaching RHAE ≥ 100 increase from 92 to 132. GPT-5.4 also improves from 25.2 to 33.3 RHAE. GPT-5.6 Sol's B6 score is 87.6 RHAE; CoMATE reaches 97.0 RHAE and 97.7 LAE, with 25 game clears and 183 completed levels. The GPT-5.5 comparison is stable across runs: B6 and CoMATE were each run five times, reporting the median-RHAE run,

Table 3: **Memory conditions under the same search protocol.** All rows use GPT-5.5, six agents, the same enhanced harness, and the budget and retry schedule of Section 3.1. Apart from B6's uncapped retry, the rows differ only in how agents' findings are passed between agents.

Cross-agent memory	Retry budget	Score	Levels	Clears
None (B6)	∞^*	55.1/67.5	142/183 (92)	13/25 (6)
Individual memory	1000	51.0/63.4	145/183 (77)	13/25 (6)
Plain summary	1000	64.6/71.4	157/183 (101)	17/25 (7)
Unstructured entries	1000	72.8/81.1	164/183 (112)	19/25 (9)
CoMATE	1000	80.1/85.9	171/183 (132)	22/25 (12)

Table 4: **System-level ablations of CoMATE.** All conditions use six agents under the same level-wise protocol. Bare6 plays directly without an agent framework, as in official evaluations; Symbolica6 uses Symbolica's original framework; B6 adds our harness extensions but omits consolidation. CoMATE w/o U omits unresolved-level consolidation and further exploration, so it has no retry (–). Retry budget is per agent; all conditions use 500 initial actions. Hard-12 comprises all twelve GPT-5.5 B6-uncleared games (90 levels).

Configuration	Retry budget	All games			Hard-12		
		Score	Levels	Clears	Score	Levels	Clears
Bare6	∞^*	1.3/5.8	13/183 (6)	0/25 (0)	1.1/5.2	6/90 (2)	0/12 (0)
Symbolica6	∞^*	46.5/60.4	131/183 (75)	11/25 (4)	21.5/39.0	42/90 (22)	0/12 (0)
B6	∞^*	55.1/67.5	142/183 (92)	13/25 (6)	28.7/46.2	49/90 (29)	0/12 (0)
CoMATE w/o U	–	78.1/84.2	164/183 (132)	18/25 (12)	58.9/71.5	72/90 (49)	6/12 (2)
CoMATE	1000	80.1/85.9	171/183 (132)	22/25 (12)	63.1/74.9	78/90 (49)	9/12 (2)

and their ranges (maximum minus minimum) are at most 5 levels and under 3 RHAE for CoMATE, whose clears never vary, and 6 levels, under 5 RHAE, and 2 clears for B6. Every CoMATE run therefore exceeds every B6 run in RHAE, levels, and clears.

Coverage gains include R11L, from one of six levels to all six, and SK48, from three of eight to all eight. Both methods clear CD82 and solve its first level in 15 actions, before consolidation has produced any collective memory; CoMATE then uses fewer actions at every later level. Consolidation supports progress and efficient solutions under the same search topology.

3.4 From Memory Sharing to Consolidation

Table 3 compares CoMATE with four simpler ways of passing on what agents learn. From top to bottom, each row adds one part of CoMATE, with the backbone, agents, harness, budgets (Section 3.1), and memory update points fixed. In **None (B6)**, each agent keeps its own memory, as in the base framework, and cannot see the others'. **Individual memory** shows each agent's native memory, including the CONFIRMED and HYPOTHESIS labels its writer assigned, unchanged to the other five, loaded in full into their context; it extends the base harness's self-labeled sharing across agents. **Plain summary** gives a fresh-context instance of the backbone the same inputs as CoMATE's consolidator, the agents' notes and observer-view trajectories, and passes its single paragraph in full to every agent. **Unstructured entries** matches CoMATE in inputs, entry-level storage, and retriever, but its consolidator merges findings into short free-form entries without checking them against the trajectories or recording evidence, epistemic status, and applicability conditions.

Sharing self-labeled memory brings no gain: individual memory reaches 51.0 RHAE, against 55.1 for B6, possibly because unverified claims mislead other agents or because loading every raw entry crowds their context. Each further part of CoMATE then adds to performance. Rewriting the memories into one summary raises RHAE to 64.6. Storing separate entries that agents retrieve on demand adds 8.2 points (72.8), keeping detail that a single compressed paragraph loses. Checking claims against the trajectories and recording evidence, epistemic status, and applicability conditions

Table 5: **Immediate and downstream benefits of further exploration.** The cohort contains all seven games left blocked by the initial GPT-5.5 CoMATE pass, with six agents per case. Continuation extends the existing attempts and Direct retry starts fresh, both running until agents concede; CoMATE uses collective memory from the preceding experience and at most 1000 further actions per agent. The gains Δ Levels and Δ Final Score are measured through the end of continuation, including cleared blockers and the later levels they unlock.

Method	Blocked levels cleared	Successful agents	Δ Levels	Δ Final Score
Continuation	0/7	0/42	+0	+0.0/+0.0
Direct retry	1/7	3/42	+2	+0.7/+0.6
CoMATE	5/7	18/42	+7	+2.0/+1.7

Table 6: **Effect of parallel exploration width in CoMATE.** All use GPT-5.5 and 500 initial actions per agent per level. Mean actions averages the scored action counts over completed levels.

N	Score	Levels	Clears	Mean actions
3	65.6/72.1	140/183 (113)	15/25 (10)	116.2
6	80.1/85.9	171/183 (132)	22/25 (12)	92.3
9	80.9/86.6	173/183 (136)	22/25 (12)	92.6

adds another 7.3, reaching 80.1. Over five runs of unstructured entries (median-RHAE run reported), RHAE varies by 1.4 and completed levels by at most one from 164, so every CoMATE run exceeds every unstructured-entry run in both RHAE and levels. This last step distinguishes CoMATE from existing memory systems, which already store and retrieve separate entries; its gain over a variant already far stronger than the plain summary shows that making each claim's support and scope explicit helps beyond entry-level memory. Trajectory checking and the three fields enter together because they form one dependency chain (Section 2.2): checking yields the evidence, the status is assessed against it, and the conditions record where it holds. Removing any link leaves the others undefined (a status with no evidence to assess, for instance), so per-field deletions would not yield a coherent method; the unstructured-entries control instead removes the whole chain with inputs, storage, and retrieval fixed, testing the evidence-grounded design as a unit.

3.5 System-Level Ablations

Table 4 separates the harness from memory consolidation. Under the same six-agent protocol, direct play without an agent framework (Bare6) reaches 1.3 RHAE and Symbolica's original framework (Symbolica6) 46.5; our harness extensions raise this to 55.1 (B6), the enhanced baseline against which CoMATE is compared. Points above this baseline are harder to earn: beyond every level B6 completes, CoMATE completes 29 more that lie late in their games (79% of the way through on average, versus 50%), have 2.4 times longer human baselines (179 versus 73 actions), and, as RHAE weights later levels more, carry 22.5 of the 100 attainable points. Consolidating during progress alone gives CoMATE w/o U 78.1 RHAE. Handling unresolved levels raises this to 80.1, adding seven levels and four game clears. The same consolidation process thus supports both accumulated knowledge reuse and recovery from blocked search. Hard-12 contains all twelve GPT-5.5 B6-uncleared games (90 levels). Here CoMATE reaches 63.1 RHAE versus 28.7 for B6, completing 78 versus 49 levels and clearing nine games versus none. The benefit therefore reaches the complete set of games left unfinished by the baseline.

3.6 Further Exploration on Blocked Levels

The initial GPT-5.5 CoMATE pass clears 18/25 games. We include the first level left unresolved by all six agents in each of the remaining seven games: all eligible cases, not a subset chosen by outcome or logs. Table 5 compares continuing existing attempts, fresh retries without new consolidation, and fresh attempts using updated collective memory. Continuation and Direct retry run until each agent concedes, whereas CoMATE caps further attempts at 1000 actions per agent; gains are measured

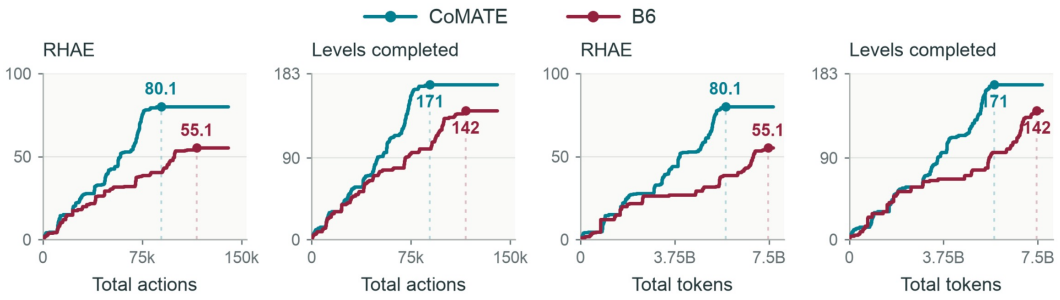


Figure 3: **Performance versus environment-action and token budgets.** GPT-5.5 results with six agents on all 25 games. The left two panels plot RHAE and completed levels against environment actions, and the right two against input-plus-output tokens. Tokens include all gameplay and memory-related calls, including failures. Markers denote first attainment of final scores; plateaus retain achieved performance.

through the end of continuation. Continuation resolves no blocker, Direct retry one, and CoMATE five, with 18 successful agents out of 42. Those five blockers unlock two later levels, yielding seven new completions, 2.0/1.7 RHAE/LAE points, and a final 22/25 game clears. The value of consolidation thus extends beyond the immediate obstacle by restoring access to the rest of the game.

3.7 Parallel Exploration Width

Table 6 varies $N \in \{3, 6, 9\}$ with GPT-5.5 and 500 initial actions per agent per level. Six agents capture most of the gain: nine add only 0.8 RHAE and two levels, with no further game clears. The six-agent configuration also shortens completed solutions on average; the slight increase at nine reflects longer solutions for its additional completions.

3.8 Interaction and Token Efficiency

At nearly every action or token budget, with tokens including memory-related model calls, CoMATE's curves lie above B6's in both RHAE and completed levels (Figure 3), a comparison independent of where each condition's retry allowance ends its run. The benefit therefore persists with consolidation and retrieval counted, not only under an environment-action budget.

3.9 Memory Reliability

An automated audit of all 5,523 entries by GPT-5.6 Terra, a different model from the GPT-5.5 backbone that wrote them, and human evaluation of 500 sampled entries judge factual support and citation entailment (Table 7, Appendix C). Yes rates are 89.9% and 81.9% for the audit and 95.6% and 90.2% for humans, with No and Uncertain judgments kept in the denominator. These assessments complement task outcomes by examining whether stored claims are supported and adequately cited, rather than assuming that successful gameplay implies reliable memory.

4 Conclusion

CoMATE organizes parallel agents' experience into evidence-grounded collective memory with explicit epistemic status and applicability conditions. The method operates on the same level-wise branching, state-handoff, and retry scaffold as the no-consolidation baseline. Under this shared protocol on the public ARC-AGI-3 games, six GPT-5.5 agents reach 80.1 RHAE, complete 171 levels, and clear 22 games; nine agents reach 80.9 RHAE and 173 levels. GPT-5.6 Sol reaches 97.0 RHAE with 25 game clears and 183 completed levels. The standard six-agent configuration adds 29 levels and nine game clears over the shared-scaffold baseline. Action and total-token comparisons account for memory-related model calls, while direct memory audits distinguish factual support from the sufficiency of citations. We demonstrate these gains on ARC-AGI-3 and believe that the consolidation mechanism may inform broader long-horizon interactive exploration.

AI USE STATEMENT

We used GitHub Copilot and Claude Code as writing assistance tools to improve the clarity, conciseness, and grammatical correctness of the text, including proofreading for typographical errors, rephrasing complex sentences for readability, and ensuring consistent terminology throughout the paper. Claude Code was also used to search for related work; the authors verified the resulting text.

ETHICS STATEMENT

The human memory audit (Appendix C) involved 16 volunteers recruited through the authors' personal networks, who acted as annotators of model-generated memory entries about ARC-AGI-3 game mechanics. Participation was voluntary and unpaid, and participants were informed of the task and of how their judgments would be used. The evaluated material contains no personal or sensitive content, and no personally identifiable information about the participants was collected.

REPRODUCIBILITY STATEMENT

Section 2 describes the consolidation executor, observer-view trajectories, segmented processing, and retrieval. Section 3 specifies the common search scaffold, budgets, comparison conditions, and complete difficult-game subsets. The appendix gives the sources of the official reference scores, reviews related work, and details memory evaluation. Interaction traces, memory artifacts, frozen prompts, key configurations, source code, and scoring scripts will be released upon acceptance to support reproduction.

REFERENCES

ARC Prize. Competition mode. ARC-AGI-3 toolkit documentation, 2026a. URL https://docs.arcprize.org/toolkit/competition_mode.

ARC Prize. ARC-AGI-3 model scorecard: openai-gpt-5-5-2026-04-23-high, 2026b. URL https://arcprize.org/scorecards/model/openai-gpt-5-5-2026-04-23-high. Published 2026-05-01.

ARC Prize. GPT-5.6 Sol – ARC-AGI results, 2026c. URL https://arcprize.org/results/openai-gpt-5-6-sol.

ARC Prize. ARC Prize Leaderboard, 2026d. URL https://arcprize.org/leaderboard.

ARC Prize Foundation. ARC-AGI-3: A new challenge for frontier agentic intelligence. Technical report, arXiv:2603.24621, 2026. URL https://arxiv.org/abs/2603.24621.

Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.14165.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. URL https://arxiv.org/abs/1911.01547.

DeepSeek-AI. DeepSeek-V3 technical report. arXiv:2412.19437, 2024. URL https://arxiv.org/abs/2412.19437.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. arXiv:2305.14325, 2023. URL https://arxiv.org/abs/2305.14325.

Alexis Fox, Junlin Wang, Paul Rosu, and Bhuwan Dhingra. PRO-LONG: Programmatic memory enables long-horizon reasoning. arXiv:2607.20064v2, 2026. URL https://arxiv.org/abs/2607.20064v2.

Pengyu Gao et al. An efficient context-dependent memory framework for LLM-centric agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Industry Track*, pp. 1055–1069, 2025. URL https://aclanthology.org/2025.naacl-industry.80/.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530, 2024. URL https://arxiv.org/abs/2403.05530.

Aaron Grattafiori et al. The Llama 3 herd of models. arXiv:2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06770.

Greg Kamradt. Analyzing GPT-5.5 & Opus 4.7 with ARC-AGI-3. ARC Prize blog, 2026. URL https://arcprize.org/blog/arc-agi-3-gpt-5-5-opus-4-7-analysis.

Seth Karten, Alex L. Zhang, Kevin Thomas, Sebastian Müller, and Prime Intellect Team. Prime agent: A self-improving RLM agent. *Prime Intellect Blog*, August 2026a. URL https://www.primeintellect.ai/blog/prime-agent.

Seth Karten, Joel Zhang, Tersoo Upaa, Jr., Ruirong Feng, Wenzhe Li, Chengshuai Shi, Chi Jin, and Kiran Vodrahalli. Continual harness: Online adaptation for self-improving foundation agents. arXiv:2605.09998, 2026b. URL https://arxiv.org/abs/2605.09998.

Xiao Liu et al. Agentbench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2308.03688.

Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The ConceptARC benchmark: Evaluating understanding and generalization in the ARC domain. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=8ykyGbtt2q.

OpenAI. GPT-4 technical report. arXiv:2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

Long Ouyang et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.02155.

Siru Ouyang et al. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.25140.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. arXiv:2310.08560, 2023. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. 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*, 2023. URL https://arxiv.org/abs/2304.03442.

Alireza Rezazadeh et al. Collaborative memory: Multi-user memory sharing in LLM agents with dynamic access control. arXiv:2505.18279, 2025. URL https://arxiv.org/abs/2505.18279.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2302.04761.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 8634–8652, 2023. URL https://arxiv.org/abs/2303.11366.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. arXiv:2408.03314, 2024. URL https://arxiv.org/abs/2408.03314.

Symbolica. From 0% to 36% on day 1 of ARC-AGI-3. *Symbolica Blog*, March 2026a. URL https://www.symbolica.ai/blog/arc-agi-3.

Symbolica. ARC-AGI-3-Agents, arcgentica branch. GitHub repository, 2026b. URL https://github.com/symbolica-ai/ARC-AGI-3-Agents/tree/symbolica/arcgentica.

Xiangru Tang et al. Agent KB: Leveraging cross-domain experience for agentic problem solving. arXiv:2507.06229, 2025. URL https://arxiv.org/abs/2507.06229.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288, 2023. URL https://arxiv.org/abs/2307.09288.

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. arXiv:2305.16291, 2023a. URL https://arxiv.org/abs/2305.16291.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/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. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv:2308.08155, 2023. URL https://arxiv.org/abs/2308.08155.

Tianbao Xie et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. arXiv:2404.07972, 2024. URL https://arxiv.org/abs/2404.07972.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2502.12110.

An Yang et al. Qwen2 technical report. arXiv:2407.10671, 2024a. URL https://arxiv.org/abs/2407.10671.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent–computer interfaces enable automated software engineering. arXiv:2405.15793, 2024b. URL https://arxiv.org/abs/2405.15793.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023a. URL https://arxiv.org/abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=WE_vluYUL-X.

Alex L. Zhang, Tim Kraska, and Omar Khattab. Recursive language models. arXiv:2512.24601v3, 2026a. URL https://arxiv.org/abs/2512.24601v3.

Guibin Zhang et al. G-Memory: Tracing hierarchical memory for multi-agent systems. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.07398.

Qizheng Zhang, Changran Hu, et al. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2510.04618.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. URL https://arxiv.org/abs/2308.10144.

Shuyan Zhou et al. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.13854.

A Official Reference Scores

The reference scores in Section 1 are ARC Prize's own evaluations on the same 25 public games with its standard single-agent harness. They count every action of first-contact play, so they indicate how difficult the games remain under direct play rather than serving as a baseline under our protocol.

For GPT-5.5 at high reasoning effort, the ARC Prize scorecard released on 2026-05-01 reports 0.73% RHAE, completing 5 of 183 levels and no game with 4,409 actions in total (ARC Prize, 2026b). ARC Prize's analysis post links this scorecard as its public-demo replays; the 0.43% quoted in the post is measured on the semi-private set (Kamradt, 2026). For GPT-5.6 Sol, the ARC Prize results page reports 13.33% on the public set at max reasoning effort and 7.78% on the semi-private set (ARC Prize, 2026c). Because our runs use high effort, we also compute the high-effort public score from the per-game results on that page, which gives about 5.2%. We quote public-set scores because our experiments use the public games; semi-private scores are measured on different games.

B Related Work

Interactive-agent research examines observation, reasoning, and tool use over extended action sequences. AgentBench, WebArena, SWE-bench, and OSWorld cover general agent capabilities, web interaction, software engineering, and desktop use (Liu et al., 2024; Zhou et al., 2024; Jimenez et al., 2024; Xie et al., 2024). ARC and ConceptARC examine abstraction and generalization (Chollet, 2019; Moskvichev et al., 2023), while ARC-AGI-3 studies unfamiliar tasks through interactive exploration (ARC Prize Foundation, 2026). Recent agent frameworks also examine this setting (Symbolica, 2026a). These benchmarks extend the study of reasoning from isolated responses to sustained behavior in environments where earlier observations influence later decisions.

Inference-time reasoning and memory provide complementary approaches to organizing extended computation. Reasoning traces, candidate sampling, and search have been studied as ways to improve problem solving (Wei et al., 2022; Wang et al., 2023b; Yao et al., 2023a; Snell et al., 2024), including through multi-agent interaction (Wu et al., 2023; Du et al., 2023). Work on agent memory and reflection investigates how experiences, retrieved information, and lessons from feedback can support subsequent decisions (Park et al., 2023; Packer et al., 2023; Shinn et al., 2023). ExpeL extracts reusable natural-language knowledge from experience (Zhao et al., 2024), while A-MEM studies dynamically organized and evolving agent memory (Xu et al., 2025). Recent research further explores recursive context processing, continual agent adaptation, and programmatic memory (Zhang et al., 2026a; Karten et al., 2026b;a; Fox et al., 2026). Other systems draw on several trajectories or agents: ReasoningBank contrasts multiple trajectories for the same query, ACE curates itemized context entries from execution feedback, and G-Memory organizes multi-agent collaboration experience (Ouyang et al., 2026; Zhang et al., 2026b; 2025). The harness we extend already shares one memory store among its agents and asks each writer to separate confirmed findings from hypotheses (Symbolica, 2026b), a form of sharing that Section 3.4 evaluates directly. Across these methods, whether a stored claim counts as established is settled by the model that writes it, by task outcomes, or by usefulness signals. CoMATE instead assigns this decision to a separate consolidator that checks each claim against the recorded observations. The rest of this appendix details this comparison; descriptions follow the cited papers, and the description of the Symbolica harness follows its released code.

Memory from an agent's own experience. Reflexion stores verbal self-reflections across repeated trials of the same task and keeps only the most recent ones; an evaluator scores the task outcome, but the reflections themselves are not checked (Shinn et al., 2023). ExpeL extracts general insights by comparing successful and failed trials, and a separate extraction model up- or down-votes existing insights, removing those whose count reaches zero (Zhao et al., 2024). A-MEM stores each interaction as a note and lets an LLM link and rewrite notes over time (Xu et al., 2025), and CDMem summarizes past trials into context-indexed guidance (Gao et al., 2025). In these methods, an insight is kept according to outcome signals, votes, or the memory model itself, and derived insights do not cite the specific steps that support them.

Memory across trajectories, agents, and tasks. ReasoningBank distills free-text memory items from trajectories that an LLM judges successful or failed, and its parallel test-time scaling contrasts

Table 7: **Evidence-based audit of GPT-5.5 CoMATE memory.** Automated judgments cover 5,523 entries; 16 volunteers evaluated 500 sampled memories. Factual and Evidence are evaluator scores that report Yes percentages, including No and Uncertain in the denominator. Revision is a corpus-level ground-truth statistic computed from memory history, shared by the two evaluator rows; it is not a rater score or an error rate. The gray column marks this different quantity.

Subset	Evaluator	Factual score (%)	Evidence score (%)	Revision (GT corpus %)
All-25	GPT-5.6 Terra	89.9	81.9	20.4
	Human	95.6	90.2	
Hard-12	GPT-5.6 Terra	85.8	76.5	18.1
	Human	92.4	84.8	

several sampled trajectories for the same query (Ouyang et al., 2026). ACE maintains itemized context entries with helpful and harmful counters, written by reflector and curator roles from execution feedback and de-duplicated by embedding similarity (Zhang et al., 2026b). G-Memory organizes multi-agent collaboration trajectories across tasks and links each insight to the queries it came from (Zhang et al., 2025); Agent KB distills experience from different agent frameworks into a shared knowledge base curated by an LLM ranker and a utility score (Tang et al., 2025); and Collaborative Memory tags fragments shared among users and agents with provenance used for access control (Rezazadeh et al., 2025). None of these checks a stored claim against the observations behind it, and conflicting entries are merged, replaced, or left unaddressed rather than kept unresolved.

The base harness. The Symbolica framework that our harness extends gives its orchestrator and subagents one shared memory store, asks writers to separate confirmed findings from hypotheses in free text, and answers queries with an LLM instructed to return only what the memories support (Symbolica, 2026b). No step checks an entry against the game trajectory, so the writer decides what counts as confirmed. The individual memory condition in Table 3 shares these native memories, with their writer-assigned labels, across the six agents, and therefore tests self-labeled sharing directly: it reaches 51.0 RHAE, below the 55.1 of B6.

What CoMATE adds. CoMATE assigns this decision to a separate consolidator. It reads the agents' notes together with observer-view trajectories, checks each claim against the recorded actions and outcomes, attaches the supporting records as evidence, keeps conflicting findings unresolved or restricts their scope, and records the conditions under which a finding held (Section 2.2). Outcome-based filtering cannot make this decision during exploration. A level can be passed on a partly wrong or incomplete interpretation, while an agent that fails may still have observed a correct mechanism, as in the LS20 example of Section 1. Whether a claim is established therefore depends on the observations behind it rather than on the outcome of the attempt that produced it. Because each claim carries its evidence, later observations can revise it, and its support can be audited directly (Section 3.9).

C Memory Evaluation

We assess the GPT-5.5 CoMATE memory corpus using two questions with different evidence packets. The automated audit covers all 5,523 memory entries, using a fresh GPT-5.6 Terra context and a single question for each judgment; the evaluator differs from the GPT-5.5 backbone that wrote the entries. The human audit covers 500 sampled memory entries evaluated by 16 volunteers from diverse occupational backgrounds recruited through the authors' personal networks.

Factual correctness. The rater receives the complete finding, its citations, visible before/after frames, and additional recorded interactions from the same relevant exploration phase. Additional material is selected by a fixed rule, with at most 300 action-index records and four extra transitions. The question is whether the complete claim holds within its stated scope and conditions, not whether a plausible fragment can be supported. A visible contradiction receives No; conditions not covered by the material or uninterpretable evidence receive Uncertain.

Evidence entailment. The rater receives the finding and the interaction records referenced by its own citations, without the extra material used for the factual question. Yes requires support for the subject, trigger, conditions, effect, and stated limitations. Partial support, mismatched references, or overgeneralization receive No; missing or unreadable cited material receives Uncertain. A finding can be correct while its citations are insufficient, so these judgments are kept separate.

Reporting. For both questions, the Yes percentage retains No and Uncertain judgments in the denominator. The human sample comprises 500 memory entries, not merely 500 rating events. A non-Yes judgment is therefore not automatically a factual error. Table 7 reports All-25 and the complete Hard-12 subset separately; the human Hard-12 subset comprises 250 of the 500 sampled entries. The sampled human assessment complements the full-corpus automated audit; it is not presented as a paired measure of rater agreement.

Revision history. The reported revision percentages are 20.4% overall and 18.1% on Hard-12. These describe memory maintenance, not a rater judgment or factual-error rate: revisions can refine evidence, scope, or epistemic status without reversing a finding. The audit thus distinguishes the validity of claims, the sufficiency of their citations, and changes to the stored record.