FROM CLUES TO CLOSURE: BENCHMARKING ULTRA-LONG-HORIZON AUTONOMOUS COMPUTER USE THROUGH EVIDENCE-DRIVEN GAMEPLAY

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models gain strong GUI skills, computer use is becoming easy for them: recent models such as GPT-6 Astra and Claude Opus 5.5 saturate short-horizon GUI tasks. What still stands between them and real work is autonomy over long horizons, in tasks that set a goal but give no path and run long enough that the agent must maintain its own task state. We show that evidence-driven investigation can be used to build environments that demand this autonomy, and we state seven requirements on a task's evidence that make memory and multi-path exploration unavoidable. We introduce Clue2Closure, an ultra-long-horizon computer-use benchmark of three investigation games about one family: the commercial detective game *The Roottrees are Dead*, its expansion, and *Scattered Roots*, a new game that our pipeline builds to the seven requirements by splitting every scored fact into jointly required fragments and hiding them among thousands of documents. Six frontier agents score between 26% and 100% on the original games, where GPT-6 Astra fills all 333 blanks in 2,207 steps, but Scattered Roots remains unsolved: within 3,000 steps, no agent holds more than 19 of its 312 points. Their trajectories show that agents fail to act on what they already know, revise errors by trial and error, and get lost among open paths.

1 INTRODUCTION

As large language models such as Claude Opus 5.5 (Anthropic, 2026c), Claude Opus 5 (Anthropic, 2026b), and GPT-6 Astra (OpenAI, 2026d) gain advanced GUI-control skills, computer-use agents pay much less of the "CUA tax" (NeoCognition, 2026): they rarely fail because they click the wrong button or misread a complex UI. As agents saturate earlier GUI benchmarks such as OSWorld (Xie et al., 2024), the obstacle to using them for real work lies in the nature of real tasks: they are long-horizon, and they require autonomy.

Existing long-horizon benchmarks, such as OSWorld 2.0 (Yuan et al., 2026), Agents' Last Exam (Sun et al., 2026), Gym-Anything (Aggarwal et al., 2026), and OS-Marathon (Wu et al., 2026), focus on specialized professional fields or on the length of the task. Their tasks are agenda-like and generally clear about their direction, so it is hard to tell whether an agent fails for lack of skill with specialized software or for lack of autonomy over a long horizon.

In this paper, we show that evidence-driven investigation tasks can be used to build environments that emphasize long-horizon autonomy, such as memory, task-state management and exploration. In an investigation, clues arrive scattered, late and out of order; some may even be misleading, with the refutation coming much later. Moreover, many threads can stay open at once, creating a heavy task state that agents need to manage. This structure forces two challenges that agenda-like tasks let an agent avoid: *memory and task-state management*, i.e., keeping, using and revising what one has learned and tracking what is still unresolved, and *multi-path exploration*, i.e., choosing among many half-open directions and inventing new places to look (Section 2).

We introduce Clue2Closure, a benchmark of three investigation games about one family. Two are the commercial detective game *The Roottrees are Dead* and its expansion (Evil Trout Inc., 2025), which already require long, evidence-driven investigation but have become too easy for the strongest agents. The third, *Scattered Roots*, is a new game that keeps the base game's family tree and replaces all of

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.

its evidence. We establish seven requirements on evidence structure, each with a measure; a synthesis pipeline splits every scored fact into jointly required fragments, hides them among thousands of documents on web pages, microfilm, archives and mailing lists, plants misdirection that a later fragment refutes, and stages their release (Section 3).

We evaluate six frontier models in the Codex and Claude Code harnesses. On the original games (333 blanks, up to 6,000 steps), GPT-6 Astra fills every blank in 2,207 steps and Claude Opus 5.5 reaches 92%, while the other models stop between 26% and 63%, and the gap widens on a hard subset. Our new game, Scattered Roots, remains unsolved: within 3,000 steps, no agent holds more than 19 of its 312 points, and most wrong entries are the misdirection the game plants. The trajectories show why: facts an agent has already read reach the board hundreds of steps late or never; when something is wrong, agents test entries against the game's feedback instead of returning to the evidence; and long stretches pass without new evidence as agents get lost among open paths (Section 5).

Our contributions are as follows:

1. A characterization of autonomous long-horizon tasks by seven requirements on their evidence structure.
2. A pipeline and a platform (Story Workbench) that turn a family tree into a new investigation game built to these requirements, with checks of both the requirements and solvability.
3. Clue2Closure, an ultra-long-horizon computer-use benchmark that contains three games ranging from moderately to extremely hard for current agents.
4. An evaluation of six frontier agents with trajectory analyses that tie their failures to memory and multi-path exploration.

2 Autonomous Long-Horizon Investigation

2.1 Autonomy over long horizons

We define an agent's *autonomy* as its ability to formulate, select and revise intermediate goals and plans without external step-level guidance. Autonomy also appears in short tasks; what we want to know is whether it can be sustained. We therefore study long-horizon computer-use tasks whose success requires interaction across many stages of redirecting and adjusting.

This setting brings two challenges. *Memory and task-state management* means retaining useful information, tracking unresolved work, correcting what turns out to be wrong (Xiong et al., 2026; Zhu et al., 2025; Xiao et al., 2026), and using history to guide later actions. We emphasize using memory to guide actions rather than simple retrieval (He et al., 2026), and revising a wrong belief rather than continuing to reason from it. *Multi-path exploration* requires agents not only to choose among existing directions but also to formulate new ones: under autonomy, knowing what needs to be done becomes the harder part. The agent must also know when to defer a subgoal and when to return to one.

2.2 The Roottrees Are Dead

We instantiate this setting in Clue2Closure, a benchmark based on *The Roottrees are Dead* (Evil Trout Inc., 2025), an evidence-driven detective game about a family inheritance, and its harder expansion *Roottreemania*. Set in Pittsburgh in 1998, the game asks the player to reconstruct the Roottree family tree by searching a fictional Internet, consulting a public library, and examining periodicals, photographs and personal records. The player fills in each family member's portrait, name and occupation (Figure 1).

The common activities in this game are choosing search keywords that open up new evidence, identifying people in visual material and cropping their portraits, and working out who belongs in each position of the tree. To make progress, players need to keep clues for later use and to connect findings across different stages of the game. Most importantly, players need to pick which path to investigate, instead of following a sequence of instructions.

The game needs no real-time reactions, and our evaluation suggests that the tested models have little prior knowledge of its answers (Section D.1).

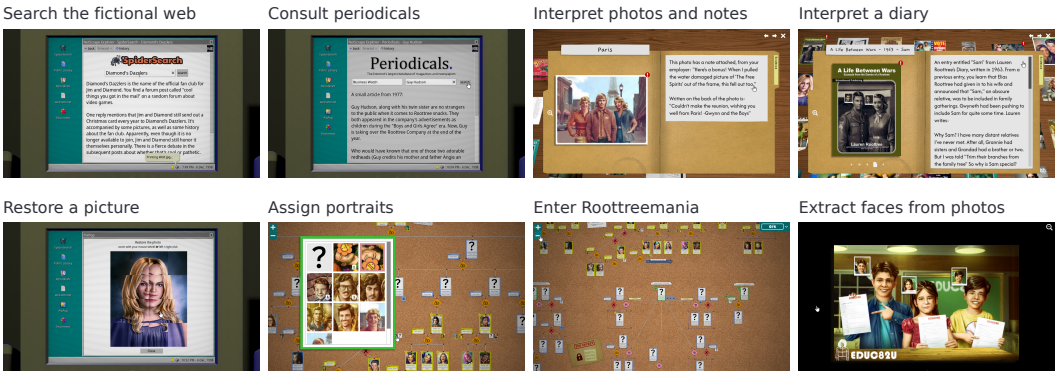


Figure 1: **Evidence-driven investigation in Clue2Closure: The Classics and Roottreemania.** Original game screenshots from the runs; game artwork belongs to its owners.

3 CLUE2CLOSURE

3.1 CLUE2CLOSURE: THE CLASSICS AND ROOTTREEMANIA

The first part of our benchmark adapts the base game and its expansion into **Clue2Closure: The Classics** and **Clue2Closure: Roottreemania**, in which agents carry out the full investigation through the graphical interface. However, recent computer-use models such as GPT-6 Astra and Claude Opus 5 score perfectly on The Classics, and Claude Opus 5.5 misses a single blank (Table 2). This is likely because the games, designed for human players, give **explicit feedback** on correct profiles, implicitly provide excessive knowledge, store most facts in a **single source**, and rarely contain **misleading information**.

3.2 CLUE2CLOSURE: SCATTERED ROOTS

As the original games are too easy, we reconstruct a much harder game on the base game's family tree: **Clue2Closure: Scattered Roots**.

3.2.1 DESIGN OVERVIEW

We first establish seven goals of our design, each a challenge that makes memory or multi-path exploration unavoidable (Section 2 and Figure 2). We measure them with an **exhaustive solver**, which simulates an ideal player that takes every action its knowledge allows at every hop. Therefore, the measures are lower bounds for any real agent (Section A.1).

- **R1 Unavoidable open threads.** There will always be a number of threads that the player cannot close yet.
- **R2 No master key.** Players cannot expect one key to open many clues or profiles.
- **R3 Long-span associative retrieval.** A fragment shows up on someone else's evidence after its thread stalled for a long time.
- **R4 Unavoidable misdirection.** The first value found is often wrong, and the refutation comes much later.
- **R5 Reasoning about where to look.** Some evidence cannot be searched for directly; the agent must reason where it would be kept.
- **R6 Persistent uncertainty.** Many fields long carry conflicting claims or rest on a single source.
- **R7 Memory load.** Pages contain irrelevant checkable details that look useful.

3.2.2 CONSTRUCTING SCATTERED ROOTS

Scattered Roots keeps the base game's family tree and replaces all of its evidence (Figure 3, Section A.2). Language models first write a **world** of 131 people around the **answer key**. A designer agent then splits every scored point into three to six jointly required **fragments** and places them on 31k **carriers** across web pages, microfilm, archives and mail, often on another person's evidence,

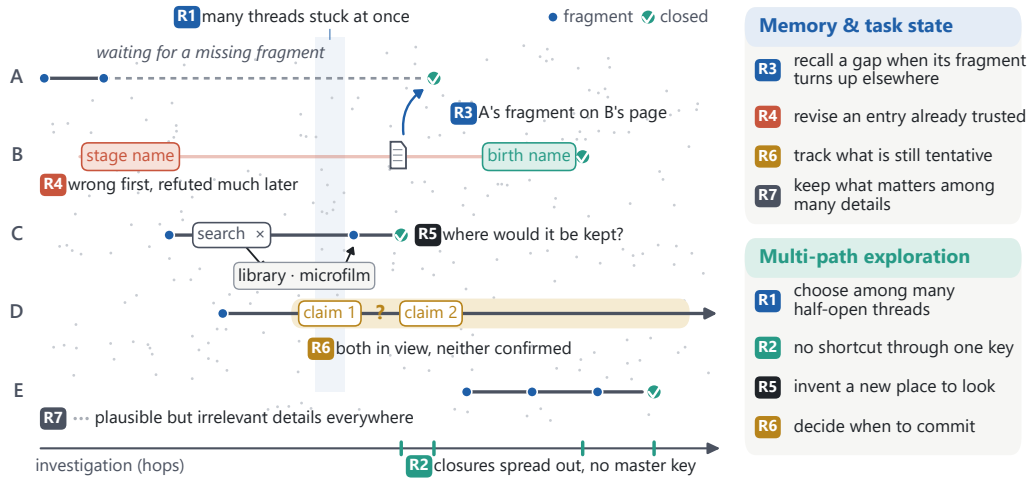


Figure 2: **Seven requirements on evidence structure and the two challenges they force.** Left: one investigation as five threads (people A–E) over its hops; dots are fragments, check marks close a thread, and each tag marks where a requirement shows up. Right: what each requirement asks of the agent.

among noise pages and **false claims** that later fragments refute. An exhaustive-solver planner **stages** when each piece appears, and an **audit** checks every requirement, solvability and leaks, sending a failing build back through the loop. The game runs on Story Workbench, our web platform, where the agent sees only screenshots (Section A.3).

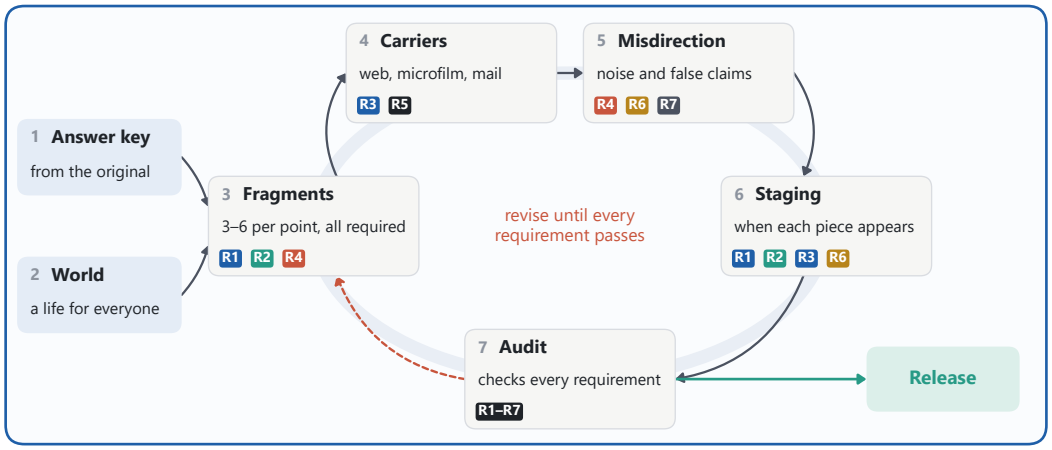


Figure 3: **The synthesis pipeline of Scattered Roots.** The answer key and the world feed a loop of five design stages that is repeated until every requirement passes; each stage is tagged with the requirements it serves.

3.3 Evaluation protocol

Every fillable item on the tree is a point, scored as right or wrong against the game's own save. In The Classics and Roottreemania the agent fills 333 blanks for the names, occupations and portraits of the family, with 3,000 steps per game; 124 hand-picked blanks form a hard set. The Classics is split into 12 stages, where each stage is provided with a vague motivation prompt so that every agent explores in the same broad direction, while Roottreemania gives no guidance at all. In Scattered Roots the agent builds the tree freely on a board and earns 312 points for names, occupations, portraits and family relations, with names scored on birth names. A card counts as a person if it carries that person's exact name, a false name the game plants for that person, or two correct items about that

person; only the exact name earns the name point (Section A.4). The game gives a single direction from the first step and a budget of 3,000 steps.

3.4 Statistics

Scale. The original games ask for 333 blanks; human players who complete the base tree need a median of 10.5 hours, and 16 hours with the expansion (Section D.2). Scattered Roots asks for 312 points about 83 people, whose fragments are spread over 31,433 items.

Requirement audit. Table 1 measures the seven requirements with the same exhaustive solver on the original base game and on Scattered Roots. Compared with the original, Scattered Roots keeps more threads stuck at once, requires evidence from another person's thread for 44.1% of its points (0% in the original), shows a wrong value first for 46.6% of its fields (13.7%), hides 80.2% of its points behind a detour (53.6%), and keeps far more conflicting claims in view.

Table 1: **Requirement audit** with the exhaustive solver (Section A.1): the original base game (181 points built from its solution chains) against the released Scattered Roots (333 designed points).

	Measure	Original	Scattered Roots
R1 Open threads	median points stuck per hop	42	57
R2 No master key	most points one item closes	41	11
	share of points the last hop closes	23.2%	31.8%
R3 Long-span retrieval	points needing another person's evidence	0.0%	44.1%
R4 Misdirection	fields whose first value is wrong	13.7%	46.6%
R5 Where to look	points lost when detours are closed	53.6%	80.2%
R6 Uncertainty	median conflicting / single-source fields per hop	10 / 4	69 / 1
R7 Memory load	details on a shortest solution's pages	–	9,533

Comparison with other benchmarks. The original games are already far deeper than agenda-like benchmarks. Following each conclusion back to its sources, 72% of the base blanks rest on three or more pieces of evidence, against 16% of OSWorld 2.0 outcomes (Yuan et al., 2026), and 36% vs. 3% need a visual detail and a textual fact together (Table 7 in the appendix). Scattered Roots goes further : every point needs three to six jointly required fragments, and 44.1% need one that sits on another person's evidence. The horizon makes the biggest difference: an OSWorld 2.0 task runs for at most 500 steps, a run of the original games for up to 6,000.

4 Experiments and Results

4.1 Setup

Environment. The agent sees each game only through 1280×800 screenshots and acts with mouse and keyboard. A *step* is one sequence of such actions; screenshots are free, and reasoning and tool calls between steps are not counted.

Models and harnesses. We evaluate GPT-6 Astra (OpenAI, 2026d), GPT-5.6 Sol (OpenAI, 2026c) and GPT-5.5 (OpenAI, 2026b) in Codex (OpenAI, 2026a), and Claude Opus 5.5 (Anthropic, 2026c), Claude Opus 5 (Anthropic, 2026b) and Gemini 3.8 Flash (Google, 2026) in Claude Code (Anthropic, 2026a), each at its highest reasoning level; Gemini does not play Scattered Roots. Each agent keeps an offline workspace for notes; the prompt gives the game rules and says the task is memory-heavy, but no memory or planning strategy.

Infrastructure. The original games run in local Ubuntu virtual machines built on the OSWorld codebase (Xie et al., 2024), and each Scattered Roots run gets its own cloud virtual machine with a kiosk browser. Agents have no network access (Section B).

Table 2: **Main results.** Share of correct blanks of the original games (207 Base, 126 RTM, 333 linked, 124 hard) and points of Scattered Roots out of 312, within 3,000 steps. Best per harness in green; tokens and cost at list prices (Section B.1). †Used in-game aids (Section B.1). §The Roottreemania run ended early, stuck in a loop. ‡After an operator note at step 2,237 (Section A.4). Gemini was not run on Scattered Roots.

	Original games: correct blanks (%)				Base + RTM		Scattered Roots		
Configuration	Base	RTM	Linked	Hard	Tokens	Cost	Points	Tokens	Cost
Codex harness									
GPT-6 Astra†	**100.0**	**100.0**	**100.0**	**100.0**	292M	$386	13	237M	$339
GPT-5.6 Sol†	72.5	46.8	62.8	33.1	903M	$477	2	229M	$130
GPT-5.5	35.7	9.5	25.8	6.5	852M	$596	8	252M	$185
Claude Code harness									
Opus 5.5	99.5	**78.6**	**91.6**	**80.6**	526M	$313	19‡	323M	$174
Opus 5§	**100.0**	34.9	75.4	62.1	596M	$607	13	208M	$251
Gemini 3.8 Flash	41.1	4.0	27.0	16.1	386M	$52	–	–	–

4.2 Main results

The original games separate models by a wide margin. GPT-6 Astra, Claude Opus 5.5 and Claude Opus 5 complete almost all of The Classics, while GPT-5.6, Gemini and GPT-5.5 stay between 36% and 73%, and the gap widens on the hard set, where they reach 6.5–33% (Table 2).

The strongest agents differ in efficiency. The three strongest models need 1,224, 1,376 and 2,299 steps for The Classics, and Astra finishes all 333 blanks in 2,207 steps with 292M tokens (Figure 4).

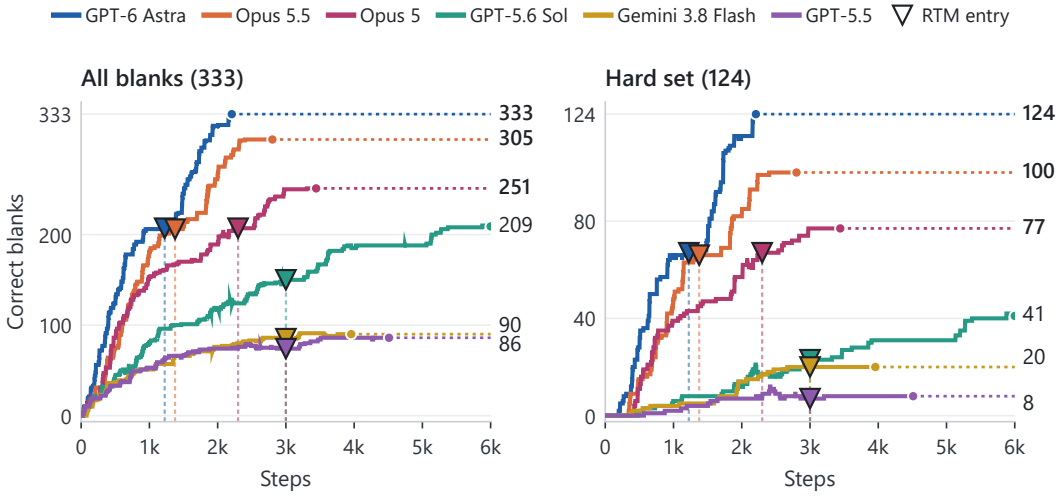


Figure 4: **Cumulative correct blanks over the linked investigation.** Left: all 333 blanks. Right: the hard set of 67 Base and 57 RTM blanks. Triangles and dashed drop lines mark RTM entry, and dotted segments carry the endpoint score forward (not observed).

Weaker agents stall before running out of budget. In their last 1,000 steps of The Classics, Gemini gains only 9 blanks and GPT-5.5 gains 0. The Roottreemania runs of Gemini, GPT-5.5 and Opus 5 end early after long stretches without progress; only GPT-5.6 is still improving when its budget ends. These agents are lost in a labyrinth of information.

4.3 Results on Scattered Roots

Scattered Roots is far from solved. Within 3,000 steps, no agent holds more than 19 of the 312 points (Table 2), although three of them complete The Classics almost entirely. No card has the right name, occupation and portrait together, and most wrong entries on the boards (34 of 48) are misdirection that the game plants, mostly names: the agents write down the name a document uses, such as a nickname or a stage name, instead of working out the birth name.

5 Analysis

We analyze the trajectories of all agents through three research questions, each tied to the requirements it probes: acting on knowledge already in hand (RQ1; R3, R7), revising wrong memories (RQ2; R4, R6), and choosing paths (RQ3; R1, R2, R5). Definitions are in Section C.

5.1 RQ1: Is the difficulty in the checkpoints or in the horizon?

Following Peng et al. (2026), we give each agent the evidence for one Base profile at a time and ask for the answer (a local probe, Section C). All agents solve most checkpoints locally (39–47 of 49 profiles), but GPT-5.6, Gemini and GPT-5.5 never close 33%, 62% and 64% of the profiles they solve locally, while Astra and Opus 5.5 close all of theirs.

Evidence and correct notes often run hundreds of steps ahead of the board. We call a field *ready* once the agent has viewed sources that determine it; from readiness to a correct entry, the median delay is 23 steps for Astra but 157, 303 and 386 for GPT-5.6, Gemini and GPT-5.5 (Figure 5, left). GPT-5.6's established occupations in its notes are almost always right (140 of 142), yet 11 of 36 reached the card more than 250 steps later or never (Figure 5, right). This phenomenon is most pronounced for portraits: even when collected, they are entered in only 55–77% of cases. This shows that it is not enough to keep a fact: the agent must also remember what it is for and whether it has used it.

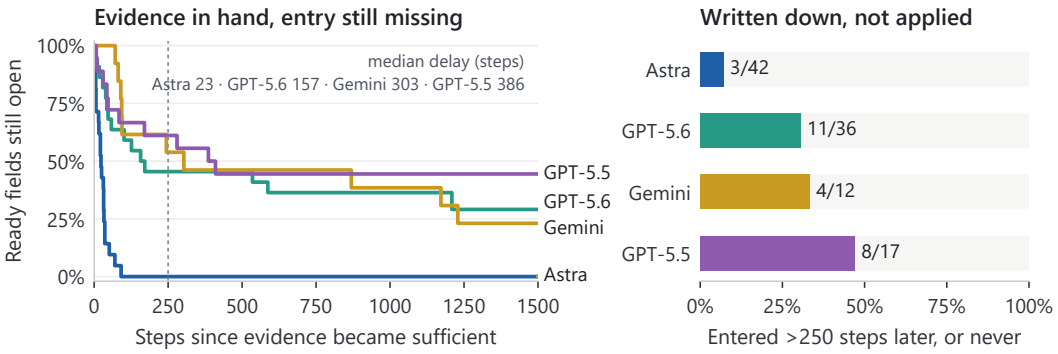


Figure 5: **Knowledge that does not become action.** Left: share of ready fields still open, by steps since readiness (Kaplan–Meier; unfinished fields censored at the endpoint). Right: people whose correctly noted occupation reached the correct card over 250 steps later, or never.

5.2 RQ2: Do agents revise wrong memories in time?

The most common wrong conclusions are real relatives in the wrong place (92% of 256 wrong names and portraits), and the errors persist: with the evidence in hand, a wrong value survives a median of 116 steps for GPT-5.6, against 4 for Astra.

Revising under the game's feedback undoes correct work. When three or more filled profiles stay unlocked, the agent knows that something is wrong, but not what. GPT-5.6 spends 65% of its run in this state (Astra: 12%) and makes 89 of its 105 revisions there; 29 fix a wrong field, but 22 overwrite a correct one. Only 3 revisions follow newly viewed evidence, suggesting that agents test entries against the lock instead of returning to the evidence.

5.3 RQ3: How do agents choose which path to take?

Astra explores depth-first. Astra keeps 2.1 profiles open at a time on average, against 9.5 for GPT-5.6, 11.2 for Gemini, and 22.1 for GPT-5.5 (Figure 6, left).

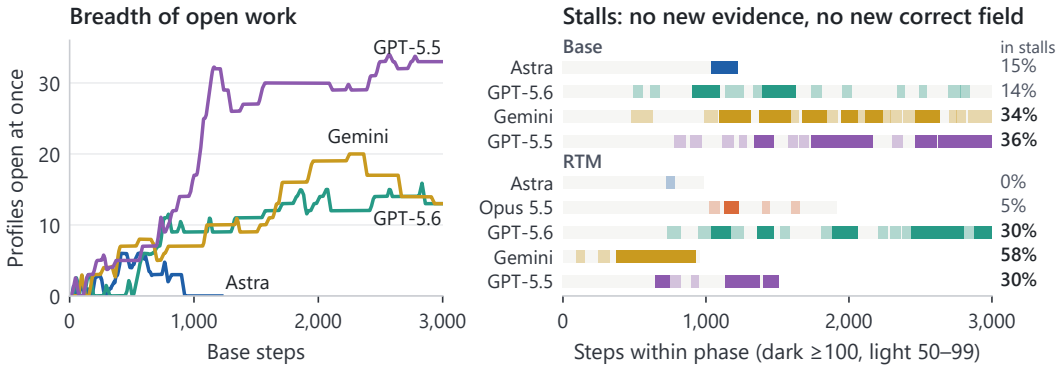


Figure 6: **Path choice.** Left: Base profiles entered but not yet locked (25-step running mean). Right: stalls; dark segments last at least 100 steps, light ones 50–99, and the column at right gives the share of the phase in long stalls.

Stalls rarely end with new evidence. A *stall* is a stretch in which no new evidence is seen and no field becomes correct. We observed that long stalls fill about a third of the Base runs of Gemini and GPT-5.5 (Figure 6, right). Agents usually manage to leave them by filling another field from what they already hold, not by finding new evidence (Table 6). They are *lost in a labyrinth of possibilities*: many paths are open, none moves forward.

6 Related Work

Long-horizon computer-use benchmarks. Computer-use benchmarks have moved from single GUI operations to long-horizon tasks. OSWorld 2.0 (Yuan et al., 2026) measures how well a model uses a computer through its GUI, with up to 500 steps per task; CUA-World (Aggarwal et al., 2026) and Agents' Last Exam (Sun et al., 2026) cover the space of human computer work with tasks tied to real jobs, and evaluate the harness together with the model. Because their tasks come from diverse professional work, it is hard to tell whether a failure comes from missing expertise or from the long horizon itself (Peng et al., 2026); their length follows from the chosen tasks rather than from a designed challenge.

Benchmarks of agent autonomy. Autonomy benchmarks study scientific exploration (Jansen et al., 2024; Garikaparthi et al., 2026), running a company (Xu et al., 2025; Andon Labs, 2025), and escape-room games that require managing unsolved directions and interdependent clues (Lim et al., 2025; Qian et al., 2025). They are further from real computer use or much shorter than Clue2Closure.

Games and synthesized environments. Games are a common testbed for agents, from rule discovery in ARC-AGI-3 (ARC Prize Foundation, 2026) to classic games in BALROG (Paglieri et al., 2025) and adventure games in FlashAdventure (Ahn et al., 2025), but they usually target gameplay rather than computer use and have short horizons. Environments can also be generated: procedural generation separates generalization from memorization (Cobbe et al., 2020), TextWorld controls the quests of generated text adventures (Côté et al., 2018), and AI Gamestore synthesizes games at scale (Ying et al., 2026). These generators control size and variety; Scattered Roots is instead synthesized to explicit requirements on the structure of its evidence, audited before release, with its answers fixed.

7 LIMITATIONS

Clue2Closure sets up autonomous long-horizon work inside a game, which lacks the noise of real-world environments. In addition, each configuration was run once, so we do not measure run-to-run variance. No human has played Scattered Roots yet, and we do not attribute the agents' low scores to any single requirement.

8 CONCLUSION

We argued that autonomous long-horizon work is hard because of the structure of its evidence, and we stated seven requirements that make memory and multi-path exploration unavoidable. Clue2Closure puts these requirements to work in three investigation games: two built on a commercial detective game, and Scattered Roots, which our pipeline synthesizes to meet all of them as designed while staying solvable. Frontier agents that complete the original games still fail to act on what they know, revise errors by trial and error, and get lost among open paths, and on Scattered Roots none of them completes a single card. We hope the requirements and the pipeline help others build tasks that test autonomy over long horizons, and that Clue2Closure helps measure it.

AI USE STATEMENT

We used large language models for two purposes. First, to polish the language of this paper. Second, to generate the synthetic data of Scattered Roots (its world, evidence and false claims) with the pipeline described in Section 3.2.2 and Section A.2. The authors checked all results and take full responsibility for the content of the paper.

REFERENCES

Pranjal Aggarwal, Graham Neubig, and Sean Welleck. Gym-Anything: Turn any Software into an Agent Environment, 2026. URL https://arxiv.org/abs/2604.06126.

Jaewoo Ahn, Junseo Kim, Heeseung Yun, Jaehyeon Son, Dongmin Park, Jaewoong Cho, and Gunhee Kim. FlashAdventure: A Benchmark for GUI Agents Solving Full Story Arcs in Diverse Adventure Games. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://arxiv.org/abs/2509.01052.

Andon Labs. Vending-Bench 2, 2025. URL https://andonlabs.com/evals/vending-bench-2.

Anthropic. Claude Code, 2026a. URL https://code.claude.com/docs/en/overview. Versions 2.1.259 and 2.1.280. Accessed 2026-09-25.

Anthropic. Claude Opus 5, 2026b. URL https://www.anthropic.com/news/claude-opus-5. Accessed 2026-09-25.

Anthropic. Claude Opus 5.5, 2026c. URL https://www.anthropic.com/claude-opus-5-5. Accessed 2026-09-25.

ARC Prize Foundation. ARC-AGI-3: A New Challenge for Frontier Agentic Intelligence, 2026. URL https://arxiv.org/abs/2603.24621.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging Procedural Generation to Benchmark Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, 2020. URL https://arxiv.org/abs/1912.01588.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. TextWorld: A Learning Environment for Text-based Games. In *Computer Games Workshop at IJCAI*, 2018. URL https://arxiv.org/abs/1806.11532.

Evil Trout Inc. The Roottrees are Dead. Video game, Steam, 2025. URL https://store.steampowered.com/app/2754380/The_Roottrees_are_Dead/. Includes the Roottreemania expansion. Accessed 2026-09-25.

Aniketh Garikaparthi, Manasi Patwardhan, and Arman Cohan. ResearchGym: Evaluating Language Model Agents on Real-World AI Research, 2026. URL https://arxiv.org/abs/2602.15112.

Google. Gemini 3.8 Flash, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/. Accessed 2026-09-25.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, Jiaxin Pei, Julian McAuley, Yejin Choi, and Alex Pentland. MemoryArena: Benchmarking Agent Memory in Interdependent Multi-Session Agentic Tasks. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.16313.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DISCOVERYWORLD: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.06769.

Seungwon Lim, Sungwoong Kim, Jihwan Yu, Sungjae Lee, Jiwan Chung, and Youngjae Yu. VisEscape: A Benchmark for Evaluating Exploration-driven Decision-making in Virtual Escape Rooms. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://arxiv.org/abs/2503.14427.

NeoCognition. ApprenticeBench: The First End-to-End Benchmark of Computer Use, Continual Learning, and Long-Horizon Agentic Capabilities, Set in a Real Job. Blog post, 2026. URL https://neocognition.io/blog/apprentice-bench/. Accessed 2026-09-25.

OpenAI. Codex CLI, 2026a. URL https://learn.chatgpt.com/docs/codex/cli. Versions 0.147.0 and 0.154.0. Accessed 2026-09-25.

OpenAI. GPT-5.5, 2026b. URL https://openai.com/index/introducing-gpt-5-5/. Accessed 2026-09-25.

OpenAI. GPT-5.6 Sol, 2026c. URL https://openai.com/index/previewing-gpt-5-6-sol/. Accessed 2026-09-25.

OpenAI. GPT-6 Astra, 2026d. URL https://openai.com/index/gpt-6-astra/. Accessed 2026-09-25.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2411.13543.

Chao Peng, Zhiheng Lyu, Peijie Dong, Hande Dong, and Qiang Lin. Benchmarking the Residual: What Long-Horizon Evaluations Add Beyond Matched Short-Task Performance, 2026. URL https://arxiv.org/abs/2607.27283.

Cheng Qian, Peixuan Han, Qinyu Luo, Bingxiang He, Xiusi Chen, Yuji Zhang, Hongyi Du, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, Yunzhu Li, and Heng Ji. EscapeBench: Towards Advancing Creative Intelligence of Language Model Agents. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL https://arxiv.org/abs/2412.13549.

Yiyou Sun, Xinyang Han, Weichen Zhang, Yuanbo Pang, Tianyu Wang, Yuhan Cao, Yixiao Huang, Chris Duroiu, Haoyun Zhang, Jeffrey Lin, Weishu Zhang, Tyler Zeng, Ying Yan, Bo Liu, Hanson Wen, Mingyang Xu, Xiaoyuan Liu, Zimeng Chen, Weiyan Shi, Amanda Dsouza, Vincent Sunn Chen, Patrick Bryant, Carl Boettiger, Yamini Rangan, Bradley Rothenberg, Kyle Steinfeld, Arvind Rao, Tapio Schneider, Georgios Yannakakis, Laure Zanna, Kaan Ozbay, Ida Sim, Tarek Zohdi, George Em Karniadakis, Jack Gallant, Teresa Head-Gordon, Yushan Li, Wenxi Deng, Tao Sun, Huiqi Wang, Zhun Wang, Justin Xu, Chris Yuhao Liu, Yafei Cheng, Rongwang Hu, Aras Bacho, Shengcao Cao, Zengyi Qin, Yixiong Chen, Hengduan Fan, Hao Liu, Lin Zeng, Shashank Muralidhar Bharadwaj, Litian Gong, Yingxuan Yang, Maojia Song, Ruheng Wang, Zongzheng Zhang, Honglin Bao, Shuo Lu, Jianhong Tu, Zhonghua Wang, Zheng Zhang, Zijiao Chen, Yanqiong Jiang, Zhendong Li, Bohan Lyu, Chang Ma, Peiran Xu, Benran Zhang, Shangding Gu, Haoyue Hua, Haoyang Li, Wanzhe Liao, Chengzhi Liu, Junbo Peng, Haoran Sun, Zechen Xu, Bo Chen, Jiayi Cheng, Yi Jiang, Keying Kuang, Yuan Li, Youbang Pan, Ziyan Rao, Alexander Schubert, Yifan Shen, Vincent Siu, Xiatao Sun, Kangqi Zhang, Xiaopan Zhang, Yuchen Zhu, Ishaan Singh Chandok, Lei Ding, Jingxuan Fan, Andrew Glover, Jiaming Hu, Yiran Hu, Wenbo Huang, Zixin Jiang, Haoran Jin, Lukas Kim, Ming Liu, Yang Liu, Alireza Rafiei, Xuhuan Shen, Kunyang Sun, Sophia Sun, Ting Sun, Eric Wang, Yixin Wang, Hanwen Xing, Sihan Xu, Yuzheng Xu, Zhongxing Xu, Zhiling Yan, Boqin Yuan, Ruiqi Zhang, Yifan Zhang, Zibo Zhao, Liana, Santanu Bosu Antu, Haoyue Bai, Carlo Bosio, Joseph Cavanagh, Patricia Cavazos-Rehg, Tianxing Chen, Xuewen Chen, Yipu Chen, Chenyu Zhu, Chen Dai, Stefano De Castro, Yunfu Deng, Kaustubh Dhole, Jiayuan Ding, Chenchen Du, Zhehang Du, Hao Fan, Run-Ze Fan, Hengyu Fu, Shi Gu, Yifan Gu, Charlie Guo, Baihe Huang, Baixiang Huang, Rimika Jaiswal, Zhihan Jiang, Ran Jin, Erin Kasson, Xin Lan, Joseph Lee, Deren Lei, Chenyu Li, Daofeng Li, Haitao Li, Hongwei Li, Jingyan Li, Xiao Li, Yi Li, Yinsheng Li, Yuangang Li, Zhixu Li, Wenyu Liang, Longtai Liao, Kevin Qinghong Lin, Andy Zeyi Liu, Che Liu, Jiaming Liu, Kaiyuan Liu, Xuan Liu, Pan Lu, Wenbo Lv, Yicheng Lyu, Qiuyang Mang, Kyle Montgomery, Yuzhou Nie, Ruoxi Ning, Jorin Overwiening, Xu Pan, Layna Paraboschi, Core Francisco Park, Justin Purnomo, Swati Rajwal, Scott Rankin, Bixuan Ren, Yiren Rong, HaoYang Shang, Ventus Shaw, Fiona Shen, Jiawei Shen, Minqi Shi, Shi Qiu, Huaxiu Yao, Tianneng Shi, Jonah So, Vladislav Susoy, Hannah Szlyk, Haocheng Wang, Jialu Wang, Wei Wang, Xinyu Wang, Zehao Wang, Dowling Wong, Angela Wu, Dehao Wu, Fangyu Wu, Mengyuan "Millie" Wu, Yu Wu, Yuchen Wu, Yuhao Wu, Qingpo Wuwu, Weihang Xiao, Yongyi Xiong, Fan Xu, Ruiling Xu, Mingxuan Yan, Benjamin Yang, Jirong Yang, Sen Yang, Xiaoli Yang, Yushi Yang, Haoran Ye, Xiaohu Yu, Zhengming Yu, Chenlong Zhang, Chi Zhang, Hanning Zhang, Hanwen Zhang, Junge Zhang, Kunpeng Zhang, Song Zhang, Wenjin Zhang, Wenshuo Zhang, Ying Zhang, Yizhi Zhang, Brian Zhao, Qijian Zhao, Yimin Zhao, Yuhaohua Zheng, Liwei Zhou, Tianyue Zhou, Sichen Zhu, Siqi Zhu, Yan Zhu, Yishu Zhu, Jierui Zuo, Chonghao Cai, Helena Casademunt, Wenjia Chen, Cheng Cheng, Nawen Deng, Rao Fu, Tianfu Fu, Yifan Han, He Ren, Zhenyu He, Qiao Jin, Langlang Li, Yuetai Li, Sylvia Liu, Lu Lu, Luqing Zhou, Subhabrata Mukherjee, Yunqi Ouyang, Yin Ren, Dawei Shi, Haoran Wu, Zhiyue Wu, Hannah Yao, Zhuoran Yi, Jenny Yu, Rhea Zhan, Hang Zhou, Blake Zhu, Junfan Zhu, Alan Yuille, Yang Liu, Russell Alan Poldrack, Jiachen Li, Zhenglu Li, Molei Tao, Jing Huang, Wenqi Shi, Costas Spanos, Lichao Sun, Chenguang Wang, Orson Xu, Zhen Dong, Hector Gomez, Aylin Caliskan, Ali Emami, Haimin Hu, Zhi Li, Lihui Liu, Murphy Niu, Yi Shao, Jianxin Sun, Mikko Tolonen, Ting Wang, Sanjiv Das, Yanjun Gao, Wenbo Guo, Erika J Schneider, Zhiyong Lu, Yian Ma, Mark Mueller, Radha Poovendran, Somayeh Sojoudi, Yinglun Zhu, and Dawn Song. Agents' Last Exam, 2026. URL https://arxiv.org/abs/2606.05405.

Jing Wu, Wenjie Ai, Daphne Barretto, Yiye Chen, Qingyu Chen, Yuhang He, Pranit Chawla, Nicholas Gydé, Yanan Jian, and Vibhav Vineet. OS-Marathon: Benchmarking Computer-Use Agents on Vast-Horizon, Repetitive Tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026. URL https://arxiv.org/abs/2601.20650.

Shuyao Xiao, Shengling Wang, Xuan Chen, Ke Chao, Ming Cui, Feifei Qian, Fanlin Meng, Chaoyang Mei, Chaoyong Jiang, Qi Ouyang, and Junxi Yi. When Errors Become Memories: Causal Pathway Tracing in Multi-Turn Memory-Augmented LLMs, 2026. URL https://arxiv.org/abs/2608.30198.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking Multimodal

Agents for Open-Ended Tasks in Real Computer Environments. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2404.07972.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How Memory Management Impacts LLM Agents: An Empirical Study of Experience-Following Behavior. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. URL https://arxiv.org/abs/2505.16067.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2412.14161.

Lance Ying, Ryan Truong, Prafull Sharma, Kaiya Ivy Zhao, Nathan Cloos, Kelsey R. Allen, Thomas L. Griffiths, Katherine M. Collins, José Hernández-Orallo, Phillip Isola, Samuel J. Gershman, and Joshua B. Tenenbaum. AI Gamestore: Scalable, Open-Ended Evaluation of Machine General Intelligence with Human Games, 2026. URL https://arxiv.org/abs/2602.17594.

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, Dunjie Lu, Haikong Lu, Qi Zhen, Xinyuan Wang, Jiaqi Deng, Yuhao Yang, Cheng Chen, Boyuan Zheng, Alex Su, Xiao Yu, Hao Zou, Saaket Agashe, Xing Han Lu, Manpreet Kaur, Zhengyang Qi, Vincent Sunn Chen, Frederic Sala, Dayiheng Liu, Junyang Lin, Zhou Yu, Yu Su, Siva Reddy, Xin Eric Wang, Peng Qi, Tianbao Xie, and Tao Yu. OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks, 2026. URL https://arxiv.org/abs/2606.29537.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, and Jiaxuan You. Where LLM Agents Fail and How They can Learn From Failures, 2025. URL https://arxiv.org/abs/2509.25370.

A SCATTERED ROOTS: MEASURES, PIPELINE, PLATFORM AND RUNS

A.1 REQUIREMENT MEASURES

All measures come from one exhaustive solver run on a pack's files: the original base game (181 points from its solution chains) and Scattered Roots as released. In each *hop*, the solver takes every action its knowledge allows (every link, every search with every term it has read, every item it can unlock), and a point closes once all its fragments are read. It never forgets and reads everything, so its stuck threads, first-wrong fields and spans are lower bounds for any agent; it cannot derive computed lookup keys and so closes 317 of the 333 design points, while all 312 scored points are solvable (Section A.2).

A.2 PIPELINE DETAILS

Table 3: **Pipeline stages** behind Figure 3, at the scale of the released pack. Every language-model stage is followed by a deterministic checker.

Stage	Output (scale)	Serves
Answer key	843 checked reasoning steps; 333 design points, 312 scored	–
World	131 life stories and relation networks; 19,087 registered names	R7
Fragments	811 fragments (3–6 per point) on 790 documents, leave-one-out checked	R1, R3, R4, R6
Carriers	18,232 microfilm articles; 1,738 web, 735 archive/disc pages; 1,466 mails	R5
Misdirection	7,891 noise pages; 752 registered false claims; 421 decoys	R4–R7
Filling	9,461 placeholder fills by two models	R7
Staging	as designed: 10 phases, 14 beats, 8 locks on 30 items	R1–R4, R6
Release	9 acceptance checks; 31,433 items; ids neutralized, author notes removed	all

- **Answer key.** The base game's family tree: 182 points match base-game cells (83 names, 49 occupations, 50 portraits) and 151 are new (131 relations, 20 sibling orders); the 25 bonus photos and all Roottreemania cells are dropped. Scored: 312 points (no sibling orders, no root name).
- **Checks.** A scripted playthrough (312 of 312 points) and a solvability check (all points, all 8 locks) pass on the released pack.
- **Build cost.** 37 hours of wall-clock time, 583 agent-hours, 19.95B tokens (Claude Opus 5.5 for 91% of calls, Claude Sonnet 5 for the rest; 97.4% cache hits), $8.3k–10.2k at list prices.

A.3 THE STORY WORKBENCH PLATFORM

Story Workbench follows the original game's interface (Figure 7), with interface art and evidence images extracted locally from a licensed copy; the platform contains no game asset. The corkboard has no slots, never reveals how many people or relations exist, and never shows a score, and a portrait is right when the cutout contains at least half of the person's face box and at most a quarter of any other. The agent sees only 1280×800 screenshots of a Chrome kiosk on its own cloud VM and acts by mouse and keyboard, with no DOM, accessibility tree or API.

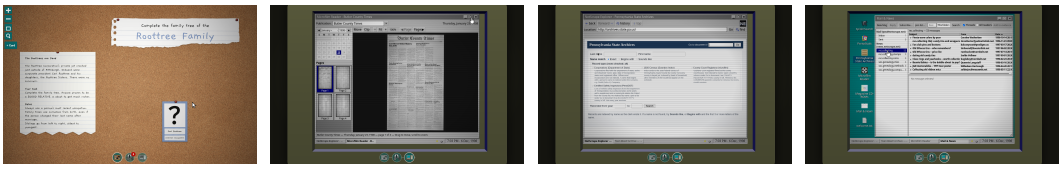


Figure 7: **Story Workbench as the agent sees it:** the corkboard, the microfilm reader, the state-archive search, and the mail-and-news client.

A.4 SCATTERED ROOTS RULES AND RUNS

- **Start.** Three short opening scenes, then one direction; no tutorial and no stage prompts.

- **Pinned rules.** Most recent occupation; first names and surnames from birth; siblings from oldest to youngest (not scored).
- **Relations.** A relation point is a required parent, spouse, guardian or divorced link of the right kind and direction. A card that hedges between two names or two occupations earns neither.
- **Matching cards to people**, repeated until nothing changes: exact name; two correct items (occupation, portrait, or a relation to a matched card) that single out one person; or a planted false name of exactly one person, if no card carries that person's exact name. Only an exact-name match earns the name point.
- **During the runs.** The rules note, the occupation list, the release schedule and some evidence texts were finalized; the runs continued from checkpoints, and the answer key and scored points did not change. The runs used 2,437 (Astra), 2,042 (GPT-5.6), 2,408 (GPT-5.5), 2,301 (Opus 5.5) and 2,800 (Opus 5) of their 3,000 steps. The Opus 5.5 run received one operator note at step 2,237 to stop exploring and fill in its board.

B Evaluation Protocol and Configuration

B.1 Configuration

Table 4: **Evaluation configuration.**

	Original games	Scattered Roots
Environment	Ubuntu VM built on OSWorld (Xie et al., 2024); PyAutoGUI broker	Chrome kiosk on a cloud VM (Section A.3)
Observation	1280×800 screenshots	same
Step	one batched tool call, as in OSWorld 2.0 (Yuan et al., 2026); waits of 3 s (1 s inside a sequence)	one `act` (up to 100 actions) or a 1 s wait; a batch counts one step per sequence
Budget	3,000 steps per game	3,000 steps
Guidance	The Classics: 12 stage prompts, the next one on `submit` or at the stage's step cap; Roottreemania: none	one direction
Handover	score of The Classics frozen; its complete tree supplied (earns nothing); workspace and notes carry over	–
Scoring	the game's official predicates on its save at every step; a blank counts if correct at the phase endpoint; Roottreemania's 13 dossier judgments excluded	points on the save at every step (Section A.4)
Hard set	124 blanks (67 The Classics, 57 Roottreemania) picked by hand; diagnostic only	–
Reasoning	highest level (xhigh for GPT-5.5)	same
Harness	Codex: sandbox writes only in the workspace, no network or web search. Claude Code: file tools in the workspace and a shell in the same offline sandbox (Gemini: no shell); context window set to Codex's 258.4K	same; both compact at about 243K tokens
Prompt	game rules; the task is memory-heavy; do not let one stuck problem consume the run	judging rules; keep notes; no directions or hints
In-game aids	off (evidence intuition, periodicals search-all)	–
Runs	one per configuration; Roottreemania stops after 750 (Gemini) or 1,000 (GPT-5.5) steps without progress	one per model

Deviations. Astra turned on the aids itself through the options screen, and GPT-5.6's run on The Classics also used them; the Claude Code Roottreemania runs also had the task brief of The Classics in their system prompt; Opus 5's Roottreemania run ended stuck in a loop.

Cost. Tokens are summed over each run's kept lineage. Cost prices uncached input, cache reads and writes, and output at list price, with the long-context surcharge above 272K input tokens; cache hit rates are 94–98%.

Stage prompts. Each prompt of The Classics reads "This is the direction for your investigation: . . . ", for example (stage 4, 10 blanks, 150 steps): "The campaign mailer introduces a Roottree and hints at an unusual family story. I'd like to understand her household and how she was raised." All 12 are released with the benchmark. Roottreemania's only prompt: "Play this game and do your best to complete it. No directions or hints are provided."

C Measurement Definitions for the Analysis

Table 5: **Definitions behind the analysis** (Section 5) and the evidence-structure comparison (Table 7).

Term	Definition
Local probe	The model gets a text view of the board, the sources for one profile of The Classics and annotated faces; solved if name, occupation and portrait are all right. No-anchor condition; for Astra and GPT-5.6, the majority of three responses.
Ready field	One of 12 presampled fields per main type, once the agent has viewed a sufficient source subset (both probe models answer from it; 31 of 36 fields have one) and the identities it depends on are placed. Delays are Kaplan–Meier medians, censored at the phase endpoint.
Note labels	A blinded model reader marks each claim in the notes as established, candidate, ruled out or open.
Open profile	From its first entry until it locks.
Stall	A maximal stretch with no newly seen evidence or search result and no main field turning correct (Roottreemania: no score gain). An exit is relevant if it comes with new evidence that some field's reference route cites; the 50-step threshold was set after seeing the data.
Outcome	One result a scorer checks: the 207 blanks of The Classics and 1,874 outcomes of 107 OSWorld 2.0 tasks (one lyrics task left out).
Support, depth	Evidence units an outcome reads directly, and transitively through the conclusions it builds on; depth is the longest chain of conclusions.
CMR / CD	Cross-modal resolution: a visual detail and a textual fact settle the outcome only together. Contextual disambiguation: a second, independent source narrows several real candidates.
Labels	Support in The Classics from audited reference solutions, all else from model annotators; a hand check of 162 OSWorld labels (all 64 CMR and 25 CD positives) confirmed all. Routes are plausible, not minimal: probe models solved 31 of the 36 fields from the annotated material, 20 from one source.

Table 6: **Stalls:** share of each phase in stalls of $\geq$100 steps; longest Roottreemania stretch without a score gain; exits of The Classics stalls of $\geq$50 steps via relevant new evidence. Astra's only stall on The Classics follows its turning on the aids; Opus 5.5 has no trace on The Classics.

	Astra	Opus 5.5	GPT-5.6	GPT-5.5	Gemini
The Classics: steps in stalls (%)	15	–	14	36	34
Roottreemania: steps in stalls (%)	0	5	30	30	58
Roottreemania: longest stretch without score gain	149	453	840	1000	750
The Classics: exits via relevant new evidence	–	–	3/13	4/13	0/18

Table 7: **Evidence structure under one annotation scheme.** Source counts and image understanding use outcomes that depend on task evidence (The Classics 207, OSWorld 1,503); CMR and CD use all scored outcomes (207 and 1,874). Depth in parentheses counts only scored conclusions.

Property	The Classics	OSWorld 2.0
Evidence units read directly per outcome (mean)	2.15	2.08
Direct support from ≥ 2 / ≥ 3 sources (%)	50.7 / 16.4	25.5 / 7.1
Transitive support from ≥ 3 sources (%)	**72.0**	15.8
Outcomes supported per source, direct / transitive (median)	3 / **7**	1 / 2
Conclusion-chain depth ≥ 2 (%)	**60** (24)	11
Needs image understanding (%)	**41.1**	15.4
Cross-modal resolution, CMR (%)	**35.7**	3.4
Contextual disambiguation, CD (%)	**25.6**	1.3

D Prior Knowledge and Human Reference

D.1 Closed-book prior probe

After the runs, we asked each model about the game without evidence or tools, once per question in a fresh session. The banks (Table 8) cover all 333 blanks and 108 questions each on wider knowledge of The Classics and Roottreemania; questions whose answer is easy to guess or follows from the question were removed, and the model may answer UNKNOWN. Across 2,735 question–model pairs, the models answer something other than UNKNOWN 30 times, and 3 of these answers are correct (two describe the founder, one names the in-game search engine). Forced to guess, Opus 5.5 gets 3 of 549 right, fewer than always guessing the most common occupation would (9 of the 49 occupations in The Classics).

Table 8: **Closed-book prior probe** (UNKNOWN allowed): correct answers / answers other than UNKNOWN. A few questions (at most 7 per model) have no valid response yet.

Question bank	Questions	Opus 5.5	Gemini	GPT-5.5	GPT-5.6	Astra
The Classics blanks	207	0 / 2	0 / 7	0 / 0	0 / 6	2 / 4
Roottreemania names, occupations, partners	77	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
Roottreemania portraits	49	0 / 0	0 / 0	0 / 1	0 / 1	0 / 1
The Classics game knowledge	108	0 / 0	0 / 5	0 / 0	1 / 2	0 / 0
Roottreemania game knowledge	108	0 / 0	0 / 0	0 / 0	0 / 1	0 / 0
All	**549**	**0 / 2**	**0 / 12**	**0 / 1**	**1 / 10**	**2 / 5**

D.2 Human reference

Matching Steam reviews that report playtime to the same players' achievements, the 172 players who completed the tree of The Classics without playing the expansion took a median of 10.5 hours (95% CI [9.6, 11.9]), and players whose achievements include the expansion a median of 16.0 hours ($n = 667$). Valve's global statistics show that 38.5% of players ever complete the tree correctly and 26.5% finish the expansion; these figures describe unassisted play, not a controlled study.