WORLDFORGE: HOW REASONING MODELS CHOOSE BETWEEN REMEMBERING, THINKING, AND LOOKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents must often act with incomplete world state, yet missing information can be handled in fundamentally different ways: an agent can spend more computation reasoning over what it already knows, or act to acquire new evidence. We introduce **WorldForge**, a simulated household platform for studying such decisions under controlled, replayable world conditions. WorldForge provides generated homes, certified physical interactions, matched task instances, and a shared interface for embodied agents, while supporting changes in information, task structure, and world dynamics. As one controlled study within this platform, we examine everyday object relocation and vary only whether target locations are given or must be recovered through interaction. Across two model–interface systems and multiple reasoning-effort settings, withholding locations reduces task completion. Agents respond by moving more, not by reasoning more per decision or opening many more containers. Increasing fixed reasoning does not reliably close the gap, and self-selected reasoning matches the best fixed-effort GPT completion at 78% lower cost without closing it either. Trajectory analysis shows that most failures occur before the target ever becomes observable, mostly for objects inside closed containers that the agent never reaches or never opens; an offline probe shows that models often rank the right furniture near the top, yet this knowledge does not by itself produce the action that exposes the target. These results expose a gap between allocating internal computation and allocating information-gathering actions, and illustrate how WorldForge can be used to isolate and study such failures in embodied agents.

1 INTRODUCTION

Ask a robot to bring the keys to the front door. They usually hang by the door, but today they are missing, and a resident has just dropped a bag on the kitchen counter. Should the robot recall where it last saw the keys, reason about where they are likely to be, or simply walk over and check?

A robot can know its home in three ways—by remembering what it saw, thinking about what it knows, or looking again—and each has a price. Memory is only as current as the last look; looking means walking to another room or opening a cabinet while other work waits; thinking takes time and cannot produce new evidence. Because reasoning models now make thinking adjustable, the robot must decide not only whether to think or to look, but also how much to think.

Recent work treats thinking as a costly resource that reasoning models over-spend (Cuadron et al., 2025; Wen et al., 2026) or learn to ration (Liu et al., 2026), and separately asks where an embodied agent should look and at what cost (Paudel et al., 2026; Zhou et al., 2025; Zhang et al., 2026a); each line ignores the other's cost. We ask whether frontier models, in tasks where looking is always an option, think more when information is missing, and whether that thinking substitutes for looking.

We study this question in WorldForge, a household platform with persona-generated homes, simulation-certified interactions, and exact replay on a shared clock. Here residents are held still and the clock pauses during model decisions, so thinking cannot change the world. Each task asks the robot to find and deliver two objects, sometimes alongside a timed appliance or behind a precondition, and is run from the same world state with and without the objects' initial locations (Known/Unknown). We evaluate GPT-6 Astra via Codex CLI and Claude Opus 5.5 via Claude Code at several native

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.

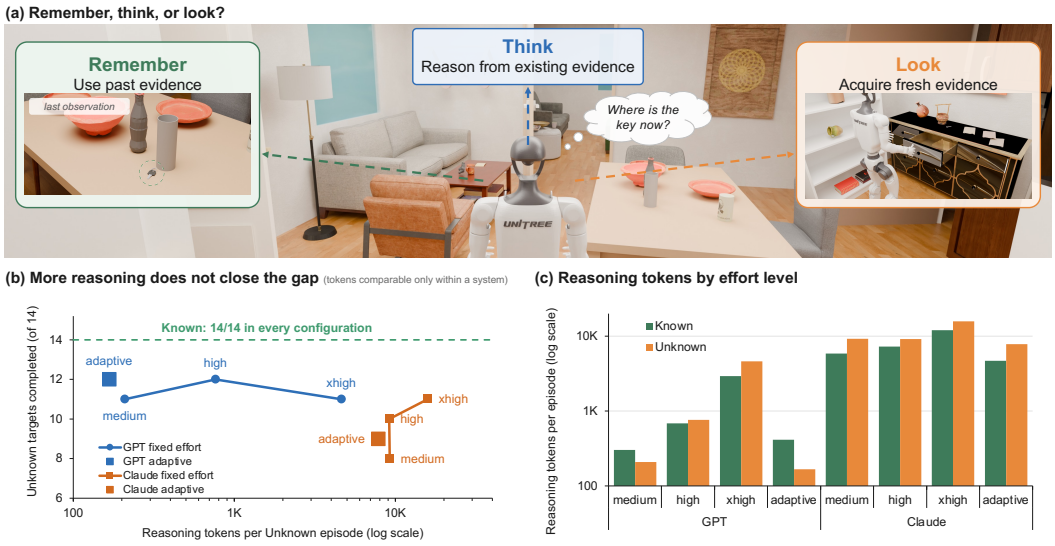

Figure 1: **Remember, think, or look?** (a) A robot missing its keys can use past evidence, reason from what it knows, or acquire fresh evidence; each has a cost. (b) With locations given (Known), both systems complete all 14 targets; with them withheld (Unknown), 22× more GPT reasoning leaves completion unchanged, and adaptive effort matches GPT's best at the lowest cost. (c) Unknown episodes use more tokens mainly because they take more decisions (Table 1). Tokens are comparable only within a system.

reasoning-effort settings, including an adaptive one in which the model picks its own effort at each step, and attribute each undelivered object to where it was lost: never seen, seen but not picked, or picked but not delivered.

Withholding locations consistently lowers success: on tasks both systems always solve when locations are given, GPT solves 4–5 of 7 without them and Claude 3–4. The agents do not give up—they take 20–51% more decisions—but the extra activity does not recover what is missing, and neither does extra thought: raising GPT's effort from medium to extra-high multiplies its reasoning tokens by 22× while success moves from 4 to 5 and back to 4. Given control of its own effort, GPT matches its best success with fewer tokens than any fixed setting. The failures lie elsewhere: most undelivered objects are never seen at all, and 12 of these 14 failures involve targets inside closed cabinets or drawers. Current models can think and can look; what they lack is the judgment to take the action that exposes the evidence.

Our contributions are:

- **A problem and a platform.** We frame embodied decision-making as a choice among remembering, thinking, and looking, and release WorldForge together with its homes, tasks, and full execution logs.
- **A protocol that isolates thinking.** Paired Known/Unknown tasks on identical world states across fixed and model-chosen reasoning effort, with per-step token accounting and stage-resolved failure attribution.
- **Findings on how models allocate thought.** Models do not think more when they lack information, gain little from more fixed reasoning, allocate cheaply when given the choice, and fail mainly by not taking the action that exposes the evidence.

2 Related Work

Reasoning as a budgeted action. Reasoning and acting have long been interleaved in language-model agents (Yao et al., 2023; Huang et al., 2023), but recent work increasingly treats reasoning itself

as a limited resource. Reasoning models may over-deliberate at the expense of interaction (Cuadron et al., 2025), struggle to allocate computation under shared budgets (Mia & Amini, 2026; Wang et al., 2026b), or act too late when the environment evolves during deliberation (Wen et al., 2026; Froger et al., 2026). Test-time scaling methods control or adapt reasoning effort (Muennighoff et al., 2025; Zhang et al., 2025), while embodied methods learn when to invoke deeper reasoning (Liu et al., 2026; Lin et al., 2026) or produce longer reasoning as tasks become harder (Zhang et al., 2026b). These works optimize or control how much an agent thinks. We instead use native reasoning effort as a diagnostic variable and ask whether frontier models allocate more thought when task-relevant information is actually missing.

Remembering and looking under partial observability. Embodied agents can compensate for missing information by using prior knowledge, memory, or active observation. Semantic priors guide search for unseen objects (Dorbala et al., 2024; Paudel et al., 2026; Yang et al., 2026), while memory-augmented agents reuse previously acquired spatial and task information (Sarch et al., 2023; Liu et al., 2025; Zhang et al., 2026c). Recent methods and benchmarks explicitly treat information acquisition as a decision, trading among exploration, memory retrieval, asking, or verification (Zhou et al., 2025; Wang et al., 2026c; Zhang et al., 2026a; Seo et al., 2026b); benchmarks with evolving task or user state further expose failures caused by outdated beliefs (Chao et al., 2026; Wang & Li, 2026). These works make looking or remembering adaptive, but generally hold the underlying reasoning regime fixed or train a dedicated policy. Our matched conditions independently vary supplied information and native reasoning effort, allowing us to ask whether additional thought substitutes for, or changes, embodied information gathering.

Generated and persistent household environments. Household simulators provide procedural scenes, interactive objects, physical tasks, and human activity at increasing scale (Deitke et al., 2022; Yang et al., 2024; Raistrick et al., 2024; Li et al., 2023; Puig et al., 2024; 2018). Dynamic environments additionally model asynchronous processes and human-driven change (Gonzalez-Pumariega et al., 2025; Wang et al., 2026a; Seo et al., 2026a). WorldForge instead targets controlled measurement of information allocation: certified interactions make looking a physical action with a cost, and checkpointed replay on a shared clock lets that cost and the cost of reasoning be measured within the same household execution.

3 WORLDFORGE

To study how embodied agents allocate reasoning under missing information, we need an environment in which information must sometimes be acquired through physical interaction, the same task can be replayed under controlled information conditions, and reasoning itself can be varied without changing the underlying world. We describe how WorldForge provides these properties through its generated homes and certified interactions (§3.1), its agent interface (§3.2), and its task and reasoning conditions (§3.3).

3.1 ENVIRONMENT AND CERTIFIED INTERACTIONS

WorldForge generates each home around its resident personas. A brief and a resident specification drive an LLM-based staged pipeline—floor plan, furniture, fixtures, interactive controls, manipulable objects—in which the residents' routines determine which rooms the home contains and which objects each room holds, so object locations are generated to follow household regularities. Articulated objects are drawn from a library of 1,706 ArtVIP and HSSD assets (Jin et al., 2026; Khanna et al., 2024), with joints and interaction specifications read from the assets rather than authored. Scene generation enforces collision, clearance, and robot-reachability constraints, so layouts are traversable and objects sit at robot-reachable positions (Figure 2a); scenes are exported to MuJoCo, from which the agent's egocentric RGB is rendered. Of the 50 released homes, 10 pass capability certification and form the evaluation set. Generation statistics and scene-construction details are given in Appendix **??**.

Crucially, interaction affordances are verified in simulation rather than assumed from asset labels. WorldForge tests five capabilities—movable, openable, container, surface, and switchable—using geometric checks and executable simulation probes (Figure 2b; Appendix **??**). Because containers are articulated bodies whose doors and drawers must be opened, obtaining hidden information may

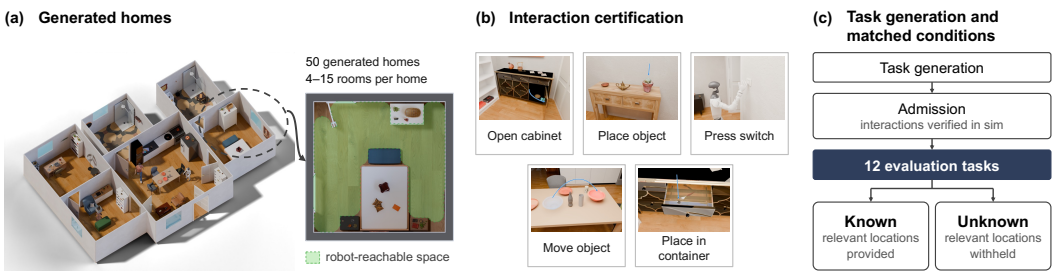

Figure 2: **WorldForge evaluation pipeline.** (a) Persona-generated homes with robot-reachable layouts. (b) Five interaction capabilities certified by simulation probes. (c) Tasks are admitted only if their interactions execute in simulation, then run under Known and Unknown conditions that differ only in whether target locations are stated.

itself require an embodied action: an object inside a closed cabinet is not rendered in any camera view until the agent reaches and opens that cabinet, so its location cannot be recovered by looking harder at the current image.

WorldForge also maintains checkpointed world state and supports exact replay. Residents, appliances, robot actions, and model reasoning can share a common clock. In the present study, however, residents are held fixed and the clock is paused during model decisions, so changing the reasoning effort cannot alter the world trajectory. This deliberately removes exogenous world change: no failure to find or deliver an object can be attributed to the world moving, and what remains is the interaction between information availability, reasoning effort, and the agent's own actions.

3.2 Agent Interface

The embodied agent controls a kinematic Unitree G1 equipped with an egocentric RGB camera. At each decision it receives the task instruction, a static map of the home without movable-object locations, its observation history, the current RGB image, held-object state, and feedback from the previous action. No privileged simulator state or hidden object location is exposed.

The agent acts through a shared set of observation, navigation, manipulation, timing, and termination skills, including `observe`, `go_to`, `open`, `pick`, `place`, `press`, `wait`, and `finish`; the complete interface is listed in Appendix **??**. Each model decision produces one public action, which is validated and executed by the same skill executor for all systems.

The abstraction deliberately removes low-level control while preserving the information-gathering decisions of interest. In particular, the interface contains no `search`, `find`, or `verify` primitive. The executor may determine how to carry out a requested skill, but the model must decide *where* to go, *what* to inspect, whether a container should be opened, which object to manipulate, and when enough evidence has been gathered to continue the task.

3.3 Task Generation and Reasoning Conditions

Task generation. Rather than manually authoring individual evaluation episodes, we define task categories over the interaction capabilities supported by the simulator and instantiate tasks by sampling compatible homes, objects, receptacles, and interaction requirements. Candidate instances are then checked against the simulated scene and retained only when their required interactions are valid (Figure 2c; Appendix **??**).

This study evaluates 12 retained instances across five homes. Every task requires the agent to locate two target objects and deliver them to specified destinations: six Type-A tasks require only the two placements, two Type-B tasks add a timed household process, and four Type-C tasks add an interaction precondition such as opening or clearing the destination. Across the 12 instances, nine of the 24 target objects begin inside closed containers and 15 on exposed surfaces; the full task list is given in Appendix **??**.

Matched information conditions. Each instantiated task is evaluated under two information conditions that share the same underlying world checkpoint (Figure 2c). In **Known**, the instruction includes one sentence specifying the initial location of each target object. In **Unknown**, only these two location statements are removed. The task goal, physical world state, robot start, initial camera image, system prompt, and action interface are otherwise identical. Known and Unknown therefore differ in supplied information rather than in task difficulty or physical initialization.

Reasoning effort. Each system is evaluated at medium, high, and extra-high reasoning effort using its interface's native reasoning setting. Because these settings provide soft guidance rather than fixed token budgets, we record the reasoning tokens reported at every decision. Both systems are also run in an adaptive condition in which the model selects the reasoning effort for its next decision from low, medium, high, and extra-high while acting (Appendix **??**). The request is an explicit next-turn selector routed by the harness, not a provider-native automatic mode; it adds no call, and its instructions and output count toward measured tokens. We make no claim about undocumented provider-side reasoning caches.

4 EXPERIMENTS

We ask whether more reasoning helps embodied agents recover information withheld at the start of an episode and, when it does not, where they fail. We measure the Known–Unknown gap across reasoning efforts (§4.2), localize failures between reaching an information source and exposing the target (§4.3), test whether offline spatial priors become evidence-acquiring actions (§4.4), and let the model choose its own effort (§4.5).

4.1 SETUP

Systems and runs. We evaluate GPT-6 Astra through Codex CLI and Claude Opus 5.5 through Claude Code under the conditions of Section 3.3, one run per system–condition–effort configuration (execution details and budgets in Appendix **??**). Because the two interfaces serialize context differently, we treat each model–interface pair as a separate system: Known–Unknown and effort comparisons are made only within a system, cross-system contrasts describe where each system fails rather than attributing the difference to the model, and reported reasoning tokens are not normalized across systems.

Information-analysis subset. Results over all 12 tasks are in Appendix **??**. For mechanism analysis we keep the seven Type-A and Type-C tasks (three homes; five Type-A, two Type-C) whose Known condition succeeds in every completed configuration of both systems, so that their Unknown failures primarily reflect information acquisition and use rather than task difficulty. This gives 14 targets per configuration, or 13 for analyses that depend on a target's staged location, which exclude the IF20_03 controller because an approved equivalent satisfies its goal. Both Type-B process tasks are excluded, so we make no claim about process tasks. Unless noted otherwise, §4.3 pools the three fixed efforts (42 target–configuration runs per system under Unknown); the adaptive condition is analyzed in §4.5.

Measures. A task is complete when both target-placement predicates hold in the final world state (accepting the approved equivalent controllers in IF20_03), which is distinct from success under all task constraints. For each unfinished target, evaluator-side records never exposed to the agent give the furthest stage reached: *acquisition* if the target never appears in a delivered image (whether or not the model recognizes it), *use* if it is seen but never picked, and *execution* if it is picked but not delivered; for targets inside closed containers we further record whether the agent *reached* the source furniture and *opened* the compartment holding the target.[1] We also record decisions, reasoning tokens, navigation distance (public `go_to` path lengths, including partial movement in failed calls), and executed skills; costs are means over the seven subset episodes of each configuration, including failures, and repeated evaluations of the same target across configurations are target–configuration runs, not independent targets.

[1] Operationally, a target is seen when a delivered observation carries a positive visibility flag for it, a source is reached when a `go_to` call returns a successful arrival at that furniture, and a compartment is opened when an effective `open` acts on a joint that controls it (Appendix **??**).

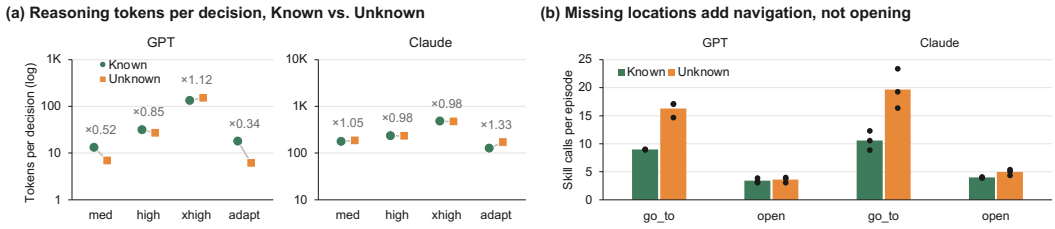

Figure 3: **Missing locations add movement more than deeper reasoning or opening.** (a) Reasoning tokens per decision under Known and Unknown (log scale); labels give the Unknown/Known ratio. (b) Mean `go_to` and `open` calls per episode; bars average the three fixed efforts, dots mark individual efforts. Tokens are comparable only within a system.

4.2 Does More Reasoning Recover Missing Information?

Table 1: Completion and cost on the seven-task subset. All configurations complete 7/7 tasks and 14/14 targets under Known, so completion is shown for Unknown. Costs are Known/Unknown means over the seven episodes, including failures; reasoning tokens are interface-reported and comparable only within a system. Bold: each system's best Unknown completion and, among GPT settings attaining it, the lowest reasoning cost.

		Unknown completed		Decisions		Nav. distance (m)		Reasoning tokens	
System	Effort	Tasks (/7)	Targets (/14)	K	U	K	U	K	U
GPT	medium	4	11	22.7	30.4	31.1	58.2	302	209
	high	**5**	12	21.7	28.3	30.5	51.4	687	764
	extra-high	4	11	21.7	30.6	30.1	54.8	2,917	4,615
	adaptive	**5**	12	22.9	27.4	31.2	49.1	413	**167**
Claude	medium	3	8	32.9	49.6	32.5	67.8	5,843	9,222
	high	3	10	30.7	39.3	37.3	60.6	7,256	9,149
	extra-high	**4**	11	24.9	33.6	28.9	54.6	12,005	15,842
	adaptive	3	9	36.9	46.1	36.0	53.1	4,695	7,830

By construction, both systems complete all seven Known tasks in every configuration (Table 1). Removing only the target locations opens a persistent gap: GPT completes 4–5 of 7 Unknown tasks and Claude 3–4.

Raising fixed reasoning effort does not reliably close it (Figure 1b): GPT's reasoning use rises 22×, from 209 to 4,615 tokens per Unknown episode between medium and extra-high, while task completion goes $4/7 \rightarrow 5/7 \rightarrow 4/7$, and Claude completes 3, 3, and 4 tasks (8, 10, and 11 of 14 targets). More computation sometimes helps individual targets but does not consistently remove the gap, and differences between adjacent settings are within run-to-run variation.[2]

Nor do agents stop acting: Unknown episodes take 20–51% more decisions and 57–109% more navigation distance in every configuration. Reasoning per decision does not rise with them: at fixed effort, Unknown decisions use 0.52–1.12× the reasoning tokens of Known decisions (Figure 3a), so the extra tokens in Unknown episodes (Figure 1c) come from more decisions rather than deeper ones. The extra interaction is mostly movement rather than opening (Figure 3b). Under the three fixed efforts, mean `go_to` calls per episode rise from 8.9–9.0 to 14.7–17.1 for GPT and from 8.9–12.3 to 16.4–23.4 for Claude, whereas `open` calls rise only from 3.4 to 3.6 and from 4.0 to 5.0 (means over 21 episodes per condition), and neither system ever calls `observe`; §4.3 shows that on closed-container tasks even this additional opening does not reach the compartments that hold the targets.

Table 2: Furthest stage reached by Unknown targets on the admitted tasks, pooled over the three fixed efforts (42 target–configuration runs per system). Per-configuration counts are in Appendix **??**.

Stage	GPT	Claude	Total
Completed	34	29	63
Acquisition failure	4	10	14
Use failure	3	2	5
Execution failure	1	1	2
Acquisition failures on closed-container targets	4 of 4	8 of 10	12 of 14

Table 3: Target-level funnel under Unknown on four of the five closed-container targets (IF20_06, IF20_07, IF20_15, IF20_20). Each row starts from 4 targets; the last column counts sources reached without the correct compartment being opened.

System	Effort	Reached	Opened	Visible	Completed	Reached, not opened
GPT	medium	3	3	3	2	0
	high	3	3	3	3	0
	extra-high	3	2	2	2	1
Claude	medium	3	1	1	1	2
	high	4	1	1	1	3
	extra-high	4	2	2	2	2

4.3 Where Do Failures Occur?

Most failures occur before any evidence is obtained (Table 2): 14 of the 21 failed target runs are acquisition failures, in which the target never appears in a delivered image. They do not diminish with effort (1–2 per configuration for GPT, 3–4 for Claude at every level), and 12 of the 14 fall on the five admitted targets that start inside a closed container, so the extra opening that Unknown induces (§4.2) does not reach the compartments that matter.

Table 3 follows each closed-container target through reaching the source, opening the compartment, exposing the target, and completing placement (definitions in §4.1); the two systems break at different stages. GPT fails to reach: it reaches 3 of 4 sources at every effort and, once there, opens the correct compartment in all but one case. Claude fails to open: it reaches 3–4 sources per effort but opens the correct compartment for only 1–2, passing a reached source without opening it seven times, five without any opening attempt and two by opening a different drawer of the same furniture. GPT's three use failures are of a third kind: on IF20_06 it picks a visually similar paperback instead of the target at every effort.

Targets that are found are not found much later: over target–effort pairs seen under both conditions (35 for GPT, 29 for Claude), withholding locations delays the first sighting by 4.7 decisions on average for GPT (median 2) and 2.2 for Claude (median 1). The gap is thus driven by targets that are never exposed, because the agent either never reaches the source or never opens the compartment, rather than by slow search; these trajectories involve few distinct objects and do not isolate a single causal mechanism.

4.4 Do Spatial Priors Translate into Search Actions?

The acquisition failures above admit two possibilities: the agent may not know where the missing evidence is likely to be, or it may reach a plausible source but fail to perform the interaction needed to reveal it. We separate these possibilities with an offline location probe.

Location probe. For each target, the model receives only the static furniture map used during the episode and the target description from the Unknown instruction. It sees no image, trajectory, memory, or ground-truth placement, and ranks all certified candidate regions in the home by how

[2] A second run of Claude medium on the same seven Unknown tasks moves completion from 3/7 to 4/7 (Appendix **??**).

likely they are to contain the target. Each question is also asked with the candidate order reversed to check for ordering effects. Thirteen targets for which both systems return complete rankings enter the comparison; full prompts and reversal statistics are reported in Appendix **??**.

Both systems exhibit useful spatial priors. The true source furniture has a median rank of 2 for both GPT and Claude, and the rankings are stable to candidate reversal (median Spearman $\rho = 0.99$ and 0.97, respectively). The exact source region appears in the top three for 6/13 GPT probes and 5/13 Claude probes, indicating that the models often assign high probability to the correct part of the home.

The stronger test is whether this knowledge translates into embodied search. Aligning each probe with the corresponding high-effort Unknown episode (Table **??**), both systems reach the true source furniture for 10 of the 13 targets. The breakdown appears one step later. Among the five enclosed targets, GPT opens the correct compartment in 4/5 cases, whereas Claude reaches all five source furniture items but opens the correct compartment in only 2/5. On IF20_15, for example, Claude ranks the dining-room sideboard as a likely source, reaches it, and still never opens it; the same pattern occurs for the TV remote.

Knowing where to look and exposing the hidden evidence are separable capabilities: current models often succeed at the former but still fail at the latter.

4.5 Self-Selected Effort Cuts Cost but Not the Gap

Because fixed reasoning effort is a coarse control, we next allow each system to choose the effort for its next decision while producing its current action. The adaptive instruction treats reasoning as costly and explicitly allows the agent to seek new evidence instead of spending more computation (Appendix **??**).

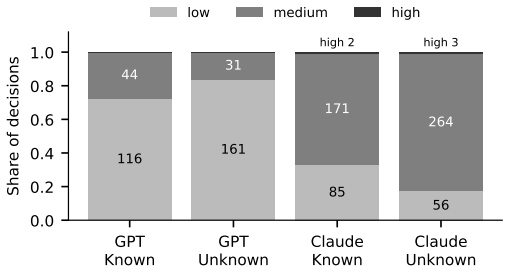

Figure 4: **Adaptive reasoning favors shallow computation.** Distribution of requested effort levels across the seven information-analysis tasks. Neither system ever requests extra-high effort.

Adaptive reasoning substantially reduces compute without sacrificing the best observed completion. GPT matches fixed high on Unknown (5/7 tasks, 12/14 targets) while reducing mean reasoning use from 764 to 167 tokens per episode, a 78% reduction. Claude remains close to fixed high (3/7 tasks; 9/14 versus 10/14 targets) while reducing reasoning use from 9,149 to 7,830 tokens. The selected efforts are shallow: GPT never requests high or extra-high, and Claude requests high on only 5 of 581 decisions.

Despite this more efficient allocation, the Known–Unknown gap remains. Allowing the models to control their own reasoning budget therefore removes much of the unnecessary computation without removing the persistent acquisition failures. **Poor reasoning-budget allocation is not the main bottleneck in recovering missing state.**

5 Discussion

Thinking does not substitute for looking. Withholding target locations changed what the agents did but not how hard they thought about each step. Unknown episodes took more decisions and covered more distance, yet reasoning per decision stayed at 0.52–1.12× its Known level at fixed

effort (Figure 3a), and raising effort multiplied cost without reliably raising completion (Figure 1b). Self-selected effort points the same way: GPT's cheapest setting reached its best completion, and the gap remained (§4.5). Within this study, native reasoning effort therefore behaves as a cost knob rather than a remedy for missing information. When evidence is absent, the useful question is not how much to think but which action would expose it.

The bottleneck is the action that exposes evidence. The failures are concentrated and specific. Most unfinished targets were never seen, almost all of these sat inside closed containers, and the location probe shows that this is not primarily a failure of knowing where to look: both systems ranked the true source furniture near the top and reached it for 10 of 13 probed targets (§4.4). What fails is the next step. GPT sometimes never reaches the source; Claude reaches it and walks away without opening the right compartment. The extra interaction that missing information induces is mostly movement between places rather than opening (Figure 3b). Improving embodied agents under partial observability may therefore depend less on richer spatial priors or longer deliberation than on treating closed containers as unobserved regions whose contents must be revealed, and on credit for actions that expose evidence.

Why matched conditions matter. Aggregate success would have hidden most of this. At equal Unknown completion (4/7 for GPT medium and Claude extra-high), the two systems fail at different stages: GPT opens every closed-container source it reaches, whereas Claude leaves two reached sources unopened (Table 3). The Known–Unknown pairing on identical world checkpoints is what lets such failures be attributed to information rather than to task difficulty. We see this as the main methodological point of WorldForge: certified interactions make looking a physical action with a cost, and matched, replayable instances make the consequences of withholding information measurable one stage at a time.

Limitations. The study is small. The mechanism analyses rest on seven tasks in three homes, 14 targets per configuration, and a handful of distinct closed-container objects; each configuration is run once, and a repeated run moves completion by one task, so differences between adjacent effort settings should not be read as rankings. We evaluate two model–interface pairs and cannot separate the model from its harness; reasoning tokens are interface-reported and comparable only within a system. Both Type-B process tasks fall outside the analysis subset, so we make no claim about timed processes. Object placements come from an LLM-driven generation pipeline, which may align them with the spatial priors of the models being probed; the probe measures agreement with these placements rather than general commonsense. The skill interface abstracts away low-level control, so failures to open reflect action selection rather than manipulation. Finally, residents are held still and the clock is paused, so this study tests thinking and looking but not remembering: memory never goes stale, and thinking costs no world time.

Future work. WorldForge already supports the conditions this study switches off. Letting residents move objects while the robot acts would test whether agents notice when memory has gone stale, and running the shared clock during deliberation would give thinking a cost in world time, turning the three-way choice among remembering, thinking, and looking into a measurable trade-off. Scaling the task set and repeating runs would allow the per-stage effects reported here to be estimated with confidence intervals.

6 Conclusion

We introduced WorldForge, a household platform with persona-generated homes, simulation-certified interactions, and matched, replayable task instances, and used it to ask whether reasoning models think their way around missing information. They do not. Withholding target locations lowers completion in both systems, more fixed reasoning effort raises cost without reliably closing the gap, and reasoning per decision does not rise when information is missing. Letting the models choose their own effort removes much of the unnecessary computation, cutting GPT's reasoning by 78% at its best completion, yet the gap remains, so poor allocation of the reasoning budget is not the main bottleneck (§4.5). Nor is the problem knowing where to look: offline, both systems rank the true source furniture near the top and reach it for most targets, but then fail to open the compartment that

holds it, or less often never reach it (§4.4). Current models can think and can look; what they lack is the judgment to take the action that exposes the evidence.

REFERENCES

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? arXiv preprint arXiv:2605.06527, 2026.

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. arXiv preprint arXiv:2502.08235, 2025.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5982–5994, 2022.

Vishnu Sashank Dorbala, James F. Mullen, Jr., and Dinesh Manocha. Can an embodied agent find your "cat-shaped mug"? LLM-based zero-shot object navigation. *IEEE Robotics and Automation Letters*, 9(5):4083–4090, 2024. doi: 10.1109/LRA.2023.3346800.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, Kunal Malkan, Dheeraj Mekala, Pierre Ménard, Gerard Moreno-Torres Bertran, Ulyana Piterbarg, Mikhail Plekhanov, Mathieu Rita, Andrey Rusakov, Vladislav Vorotilov, Mengjue Wang, Ian Yu, Amine Benhalloum, Grégoire Mialon, and Thomas Scialom. Gaia2: Benchmarking LLM agents on dynamic and asynchronous environments. In *International Conference on Learning Representations*, 2026.

Gonzalo Gonzalez-Pumariega, Leong Su Yean, Neha Sunkara, and Sanjiban Choudhury. Robotouille: An asynchronous planning benchmark for LLM agents. In *International Conference on Learning Representations*, 2025.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner Monologue: Embodied reasoning through planning with language models. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 1769–1782. PMLR, 2023.

Zhao Jin, Zhengping Che, Tao Li, Zhen Zhao, Kun Wu, Yuheng Zhang, Yinuo Zhao, Zehui Liu, Qiang Zhang, Xiaozhu Ju, Jing Tian, Yousong Xue, and Jian Tang. ArtVIP: Articulated digital assets of visual realism, modular interaction, and physical fidelity for robot learning. In *International Conference on Learning Representations*, 2026.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat Synthetic Scenes Dataset (HSSD-200): An analysis of 3D scene scale and realism tradeoffs for ObjectGoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16384–16393, 2024.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R. Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 80–93. PMLR, 2023.

Fanqi Lin, Ruiqian Nai, Yingdong Hu, Jiacheng You, Junming Zhao, and Yang Gao. OneTwoVLA: A unified vision-language-action model with adaptive reasoning. In *International Conference on Learning Representations*, 2026.

Jun Liu, Pu Zhao, Zhenglun Kong, Xuan Shen, Peiyan Dong, Fan Yang, Lin Cui, Hao Tang, Geng Yuan, Wei Niu, Wenbin Zhang, Xue Lin, Gaowen Liu, Yanzhi Wang, and Dong Huang. When should a robot think? Resource-Aware reasoning via reinforcement learning for embodied robotic decision-making. arXiv preprint arXiv:2603.16673, 2026.

Peiqi Liu, Zhanqiu Guo, Mohit Warke, Soumith Chintala, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. DynaMem: Online dynamic spatio-semantic memory for open world mobile manipulation. In *IEEE International Conference on Robotics and Automation*, pp. 13346–13355, 2025. doi: 10.1109/ICRA55743.2025.11127619.

Md Jueal Mia and M. Hadi Amini. Don't overthink, don't underthink: Toward adaptive reasoning in agentic AI. arXiv preprint arXiv:2608.26442, 2026.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1025.

Abhishek Paudel, Abhish Khanal, Raihan I. Arnob, Shahriar Hossain, and Gregory J. Stein. Object search in partially-known environments via LLM-informed model-based planning and prompt selection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2026.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8494–8502, 2018.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *International Conference on Learning Representations*, 2024.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen Indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21783–21794, 2024.

Gabriel Sarch, Yue Wu, Michael J. Tarr, and Katerina Fragkiadaki. Open-ended instructable embodied agents with memory-augmented large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3468–3500. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.226.

Gyuhyeon Seo, Jungwoo Yang, Junseong Pyo, Nalim Kim, Jonggeun Lee, and Yohan Jo. SimuHome: A temporal- and environment-aware benchmark for smart home LLM agents. In *International Conference on Learning Representations*, 2026a.

SeungWon Seo, SooBin Lim, Seongrae Noh, Haneul Kim, and HyeongYeop Kang. From assumptions to actions: Turning LLM reasoning into uncertainty-aware planning for embodied agents. In *International Conference on Learning Representations*, 2026b.

Chenxu Wang, Dunzheng Wang, Dong Wang, Xinghang Li, Yongchang Li, and Huaping Liu. DynamicTHOR: A scalable dataset of human-centric dynamic scenes for embodied AI. *Scientific Data*, 13(1):1041, 2026a. doi: 10.1038/s41597-026-07201-7.

Meiqi Wang and Shichao Li. Compiling and benchmarking task-state horizons for embodied agents. arXiv preprint arXiv:2608.08036, 2026.

Peisong Wang, Zhiwei Ma, Bowen Liu, Feixue Liu, Aochuan Chen, Chenyi Zi, Hongchuan Zeng, Yuhan Li, and Jia Li. R^3-Bench: LLMs struggle with resource-rational reasoning under shared budgets. arXiv preprint arXiv:2608.16033, 2026b.

Wenbin Wang, Xiang Bai, Yizhao Wang, Hang Sun, Dong Ren, Jie Qin, Qingquan Li, and Bing Wang. Active spatial inspection for effective and efficient embodied exploration. arXiv preprint arXiv:2609.22385, 2026c.

Yule Wen, Yixin Ye, Yanzhe Zhang, Diyi Yang, and Hao Zhu. Real-time reasoning agents in evolving environments. In *International Conference on Learning Representations*, 2026.

Heqing Yang, Ziyuan Jiao, Shu Wang, Yida Niu, Si Liu, and Hangxin Liu. Integrated exploration and sequential manipulation on scene graph with LLM-based situated replanning. In *IEEE International Conference on Robotics and Automation*, 2026.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language guided generation of 3D embodied AI environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16227–16237, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. AdaptThink: Reasoning models can learn when to think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3716–3730. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.184.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, and Manling Li. Theory of Space: Can foundation models construct spatial beliefs through active exploration? In *International Conference on Learning Representations*, 2026a.

Wenqi Zhang, Mengna Wang, Gangao Liu, Huixin Xu, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, Jiajun Liu, Weiming Lu, Peng Li, and Yueting Zhuang. Embodied-Reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41178–41207. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.1910.

Yehang Zhang, Jianchong Su, Haojian Huang, Yifan Chang, Tianhao Zhou, Xinli Xu, Yingjie Xu, Yinchuan Li, Zexi Li, and Ying-Cong Chen. WorldLines: Benchmarking and modeling long-horizon stateful embodied agents. arXiv preprint arXiv:2606.18847, 2026c.

Weijie Zhou, Xuangtang Xiong, Ye Tian, Lijun Yue, Xinyu Wu, Wei Li, Chaoyang Zhao, Honghui Dong, Ming Tang, Jinqiao Wang, and Zhengyou Zhang. ESearch-R1: Learning cost-aware MLLM agents for interactive embodied search via reinforcement learning. arXiv preprint arXiv:2512.18571, 2025.