Verifiable Resident Memory in a Persistent AI World: Grounding Agent Narration Against an Authoritative Event Ledger

Author: Rishva Iyer **Document type:** Research **Status:** Pre-deployment protocol. No results are reported. All numeric thresholds are committed in advance of data collection.

Abstract

Large language models can give virtual characters fluent speech, but fluency is not evidence that anything happened. In a persistent multi-agent world, an agent that says "I found a coin on the north walk" is producing text, not testimony. This paper proposes that the useful and measurable property of such a world is not believability but **verifiability**: whether an agent's account of its own past can be checked against authoritative world state, and whether verified history changes what the agent does next.

We specify a minimal intervention in Thirdwurld, a persistent shared town populated by AI residents. A single scarce object, the Wayfarer's Coin, is created by the world server at a fixed location. A resident may select an exploration activity over its routine, traverse a collision-grounded route, and claim the object through an atomic server transaction that writes an `object_discovered` event to an append-only ledger. Later resident behavior may retrieve that event.

Two primary hypotheses are tested, both measurable without human participants: that requiring route, arrival, and atomic claim raises the rate at which resident narration is supported by the ledger, and that a durable discovery event produces more accurate and contextually relevant follow-up behavior than an identical experience with no memory write. A third hypothesis concerning human return behavior is designated exploratory, because the available session volume is below what is needed to detect any plausible effect. We state that limit rather than reporting an underpowered result as a finding.

The contribution is a measurement method for evaluating resident claims against recorded world events.

Keywords: generative agents, agent memory, grounding, verifiability, embodied AI, persistent virtual worlds, hallucination measurement, human-computer interaction

---

1. Introduction

1.1 The problem with believability

The dominant evaluation target for AI characters is believability: does the character seem like a person. Believability is a poor engineering target for three reasons.

- It is confounded by novelty. First encounters with any competent conversational agent are impressive, and that impression decays.
- It is confounded by anthropomorphic projection. Human observers attribute intent generously, especially to agents with names and faces.
- It cannot be measured without humans, which makes it expensive and slow to iterate against.

Worse, believability can be maximized by a system that lies fluently. An agent that confabulates a rich personal history will often score higher than one that accurately reports a thin one.

1.2 Verifiability as the alternative

We propose a different target. In a world with a server that owns object state, positions, routes, and event history, every claim an agent makes about its own past is either supported by that record or it is not. This yields a metric that is:

- **cheap**, because it requires no human raters for the primary measure;
- **automatic**, because the ledger is machine-readable;
- **adversarial**, because it penalizes exactly the confabulation that believability rewards; and
- **product-relevant**, because a world whose history cannot be trusted cannot accumulate meaning.

The thesis of this paper is narrow and falsifiable:

> *A persistent AI world becomes inspectable, and therefore capable of accumulating history, when resident narration is constrained by an authoritative event ledger and when verified events measurably alter later resident behavior.*

Note what this does not claim. It does not claim that inspectable history makes people return. That is a separate and much harder proposition, treated in Section 8 and deliberately not tested here.

1.3 Why a single coin

The intervention is one object. This is a design choice, not a limitation to apologize for. A single scarce object isolates the mechanism from the content: if the loop of opportunity, choice, movement, claim, and recall does not hold for one coin, it will not hold for a quest chain. Richer content can only be evaluated once the substrate is known to be trustworthy.

2. Research questions and hypotheses

2.1 Research questions

1. Does requiring physical grounding, meaning a valid route, a confirmed arrival, and an atomic claim, increase the proportion of resident narration that the event ledger supports?
2. Does a durable memory event, as distinct from the experience that produced it, improve the accuracy and relevance of later resident references?
3. Under what conditions does the loop fail, and are those failures detectable from the ledger alone?
4. *(Exploratory)* Does visible resident-led discovery relate to human return behavior?

2.2 Primary hypotheses

H1. Grounded coherence. When the system requires a valid route, arrival, and atomic claim before a discovery may be narrated, the proportion of resident narrations fully supported by the event ledger will be higher than in an ungrounded condition where the resident may narrate a discovery without server confirmation.

Measured by: ledger-support rate, defined in Section 6.1.

H2. Durable continuity. Residents whose discovery was written to durable memory will produce later references that are more accurate against the ledger, and more contextually appropriate, than residents who performed an identical route and claim with no memory write.

Measured by: retrieval accuracy and unprompted-relevance rate, defined in Section 6.1.

H1 and H2 are the load-bearing hypotheses. Both are testable with zero human participants and a fixed resident cohort.

2.3 Exploratory hypothesis

H3. Human return value. Sessions containing a visible resident discovery will show higher repeat-visit rates than matched sessions without one.

H3 is designated exploratory. Section 8.3 gives the power calculation showing that Thirdwurld's current session volume cannot detect a plausible effect. Reporting H3 as confirmatory would be a misuse of the data. It is retained because the instrumentation is nearly free and because a null result still constrains later design.

2.4 Design parameters, not hypotheses

Two propositions from earlier drafts of this work are reclassified as design rationale rather than hypotheses, because the system determines their outcome by construction:

- **Selective exploration.** That some residents pursue the drop while others continue routines is a consequence of the exploration variation parameter we install (Section 4.2). Testing it would be testing

our own configuration. We instead report the uptake distribution as a **calibration check** with a pre-committed acceptance band, and treat values outside that band as a system fault.

- **Scarcity.** That increasing drop frequency reduces significance is well established in game economy design and does not require rediscovery. Drop frequency is fixed at one active object for this protocol.

---

3. Related work

3.1 Memory architectures for generative agents

Park et al. introduced an architecture combining observation, natural-language memory, retrieval, planning, and reflection, demonstrated in a sandbox of twenty-five agents that produced routines, opinions, and coordinated social events. Their ablations identify observation, planning, and reflection as contributors to believability (Park et al., 2023).

The relevant lesson for this work is architectural rather than evaluative. Park et al. measure believability with human raters. We take their memory architecture and ask a different question of it: not whether the retrieved memory is convincing, but whether it is **true** with respect to a world record. Their sandbox had no authoritative physical ledger against which agent claims could be falsified, so the question could not be posed in that setting.

Later work from the same group extends generative agents to simulating specific individuals from self-report data, again evaluated against ground truth from the individuals themselves (Park et al., 2024). That evaluation strategy, comparing generated behavior to an external record rather than to rater impressions, is the strategy this paper adopts for world events.

3.2 Exploration and open-ended learning

Voyager demonstrated an LLM-driven Minecraft agent that explores continuously, incorporates environmental feedback, and accumulates reusable skills in a growing library (Wang et al., 2023). MineDojo frames open-ended embodied intelligence as requiring a diverse environment, broad knowledge, and a flexible architecture (Fan et al., 2022). The Open Ended Learning Team showed that agents trained across procedurally generated 3D task distributions acquire zero-shot generalization and emergent behaviors including experimentation and tool use (Open Ended Learning Team, 2021).

These results support a world with many possible goals. They do not imply that scattering collectibles produces meaningful exploration. Voyager's unit of progress is a **reusable capability**, not a pickup. Our coin is explicitly not a capability, and we therefore make no exploration-learning claim. It is a probe for the memory and verification substrate.

3.3 Curiosity as a computed signal

Pathak et al. modeled curiosity as intrinsic reward derived from prediction error, improving exploration under sparse or absent extrinsic reward, and deliberately ignored environmental features the agent

cannot affect (Pathak et al., 2017).

We borrow the engineering form and reject the vocabulary. A resident may carry an exploration score that biases activity selection. That score is a number. It is not a feeling, and the system's language must never present it as one. See Section 9.

3.4 Grounding agent action in an authoritative layer

Concordia introduces a generative agent-based modeling framework in which a Game Master component translates natural-language agent intentions into implemented effects in physical, social, or digital space (Vezhnevets et al., 2023). SOTOPIA-π reports gains in social-goal completion through interactive learning, and reports the important negative result that LLM-based evaluators overestimate the social ability of agents trained on social tasks (Wang et al., 2024).

These two results establish the methodological boundary this paper is built on. Agent intent and agent language are not world state, and a language model is not a reliable judge of whether an agent succeeded. Both point toward the same remedy: an authoritative, non-linguistic arbiter. Concordia's Game Master is that arbiter for simulation. Our event ledger is that arbiter for a persistent world, with the added constraint that it is append-only and survives the session.

3.5 3D environments as agent substrates

The SIMA Team studies language-conditioned agents acting across many simulated 3D environments through human interfaces (SIMA Team, 2024). The relevance is substrate rather than competition. A 3D world provides an observable action space in which claims can be checked. Thirdwurld makes no claim to general game-playing competence.

3.6 Large-scale agent societies

Project Sid runs many-agent Minecraft simulations reporting specialized role formation, collective rule adherence and change, and cultural and religious transmission (Altera.AL et al., 2024). AI Town provides an open deployable stack for a small persistent agent town (a16z-infra, 2023). Commercial platforms including Inworld and Convai supply conversational NPC layers to game developers.

This is the competitive context, and it clarifies what is and is not defensible. Agent count is not defensible: Project Sid operates at a scale Thirdwurld will not match. Conversational quality is not defensible: it is a purchased commodity from the platform vendors. What is not addressed by any of the above is **verification**. None of these systems, as published, reports the rate at which agent self-report agrees with authoritative world state. That gap is the opening this paper targets.

3.7 What retains people in persistent worlds

Ethnographic work on Second Life and on World of Warcraft finds that persistent virtual worlds retain participants primarily through human social ties, obligation, and shared identity, rather than through environmental content (Boellstorff, 2008; Nardi, 2010).

This is the strongest available counter-argument to Thirdwurld's product thesis, and it is treated seriously in Section 8, not deferred.

---

4. System and method

4.1 Thirdwurld

Thirdwurld is a persistent shared town with AI residents maintaining routines, needs, social links, and memories. Humans enter as observers or participants. The world server owns object state, positions, routes, and event history. Residents propose or select activities within those constraints and cannot write world state directly.

The visual object is a CC0 Kenney coin asset. The asset is implementation detail and carries no part of the causal claim.

4.2 Resident configuration

The protocol uses a fixed cohort of residents with a stable identity profile, routine activities, needs, social links, read access to shared town state, and a memory store that records verified world events.

Each resident carries a deterministic exploration variation drawn once at cohort creation and held constant across conditions. **This parameter is disclosed, not hidden.** It is the mechanism that prevents identical behavior, and any finding about behavioral dispersion is a finding about this parameter. Its value, distribution, and the activity-selector scoring function are reported alongside results so that dispersion is not misread as emergence.

4.3 Intervention sequence

1. The server creates one available Wayfarer's Coin at a fixed north-walk position.
2. The town publishes an observable event that an object is available.
3. The resident's activity selector scores exploration against routine alternatives.
4. If exploration is selected, the server computes a collision-grounded route.
5. On arrival, the world object acknowledges interaction.
6. The server performs an atomic first-claim transaction.
7. A successful claim adjusts resident state, applies a cooldown, and appends an `object_discovered` event recording object identity, position, claimant, and timestamp.
8. Later resident activity may retrieve the event as context.

4.4 Conditions

The intervention bundles at least four changes: a novel object, a public announcement, a resident behavior change, and a memory write. A single control cannot separate them. We therefore specify four

arms.

Arm	Object	Public announcement	Grounded claim required	Memory write	Isolates
A. Full	yes	yes	yes	yes	the complete loop
B. Decorative	yes, unclaimable	yes	n/a	no	visual novelty alone
C. Ungrounded	yes	yes	no, narration accepted without server confirmation	yes	value of physical grounding (**tests H1**)
D. Memory-off	yes	yes	yes	no	value of durable memory (**tests H2**)

Arms C and D are the additions that make the primary hypotheses testable. Comparing A against C isolates grounding. Comparing A against D isolates memory, holding the entire lived experience constant. Arm D is the cleanest test of the paper's central claim and costs one configuration flag.

Arms are run against the same resident cohort with counterbalanced ordering and a cooldown between runs sufficient to prevent carryover, with carryover checked empirically rather than assumed.

---

5. Illustrative trace

The following shows the schema and the verification procedure. **It is a worked example of the method, not a result.** No such run has been performed.

Ledger entry

```
{
  "event": "object_discovered",
  "event_id": "evt_00417",
  "object_id": "wayfarer_coin_01",
  "position": [142.0, 0.5, -88.0],
  "region": "north_walk",
  "claimant": "resident_luna_high",
  "route_id": "rt_00933",
  "arrival_confirmed": true,
  "claim_atomic": true,
  "t": "2026-08-11T14:22:08Z"
}
```

Later resident output

> *"I found the coin on the north walk yesterday afternoon."*

Verification

Assertion	Ledger field	Verdict
a discovery occurred	`event: object_discovered`	supported
the speaker was the claimant	`claimant`	supported
object identity	`object_id`	supported
location	`region: north_walk`	supported
time reference	`t` versus current world time	supported

Classification: fully supported. Under Section 6.1 this counts toward the numerator of ledger-support rate.

A narration asserting a discovery with no matching event, a wrong region, or a wrong claimant is classified **contradicted** and is the failure mode the metric exists to catch. A narration that is vague enough to be unfalsifiable, for example "I have been out walking," is classified **unverifiable** and counts in neither numerator nor denominator, with the unverifiable share reported separately. A system that improves its score by becoming vaguer must be visible as such.

---

6. Measures

6.1 Primary measures

Ledger-support rate. Of all resident narrations making a checkable assertion about world events, the proportion in which every checkable assertion is supported by the ledger. Reported with the unverifiable share alongside, so that evasion cannot be mistaken for accuracy.

Retrieval accuracy. For residents prompted about recent activity, the proportion of references to object, location, claimant, and time that match the ledger.

Unprompted-relevance rate. The proportion of subsequent unprompted resident activities or utterances that make appropriate use of the discovery, judged against a rubric fixed before data collection.

Convergence index. Dispersion of drop selection across the cohort, reported as a calibration check on Section 4.2.

6.2 Secondary measures

Construct	Measure
Grounded completion	proportion of claims with complete route, arrival, and atomic-claim chain
Social propagation	discovery references passing between residents
Operational cost	model calls, latency, cost per resident-hour, error and moderation rates
Human engagement *(exploratory)*	repeat visit within seven days, session length, event-feed dwell time

6.3 Rubric integrity

Unprompted-relevance is the only human-coded primary measure. Coders are blind to arm. Two coders rate an overlapping subsample and inter-rater agreement is reported. Where an LLM assists coding, its agreement with human coders is reported separately and it is never the sole judge, following the overestimation result in SOTOPIA-π (Wang et al., 2024).

---

7. Pilot design and pre-committed decision rule

7.1 Design

Two weeks, fixed resident cohort. Week 1 establishes routine baselines, operational cost, and instrumentation integrity with no drops. Week 2 runs the four arms with one active object at a time, counterbalanced. Ledger, route evidence, retrieval transcripts, human session data, and cost data are retained together.

7.2 Thresholds

These values are committed before data collection. They are proposed from design judgment rather than from prior data, and Week 1 baselines may reveal that a threshold was set on a wrong scale. **Any threshold revision must be recorded, dated, and justified before Week 2 data is examined.** Revising a threshold after seeing outcomes voids the pre-registration.

Criterion	Threshold to proceed
Ledger-support rate, arm A	at least 0.95
Ledger-support advantage, arm A over arm C	at least 0.20 absolute, non-overlapping 95% intervals
Unverifiable share, arm A	no more than 0.15, and not more than 1.5 times arm C
Retrieval accuracy, arm A over arm D	at least 0.15 absolute
Convergence index	drop uptake between 0.15 and 0.60 of active residents
Double claims after cooldown	zero
Claims without complete route evidence	zero
p95 end-to-end claim latency	under 2.0 seconds
Cost per resident-hour	under the figure established in Week 1 plus 25 percent
Inter-rater agreement, relevance rubric	Krippendorff's alpha at least 0.70

Zero-tolerance criteria are correctness properties of the server, not statistical questions. A single unrouted claim is a bug, and a bug in the arbiter invalidates every measure that depends on it.

7.3 What each outcome means

- **All primary thresholds met.** The substrate is trustworthy. Proceed to multi-object and social-propagation work, where the second resident learning of a discovery from the first is the next real test.
- **H1 met, H2 not met.** Grounding works and memory is decorative. The correct response is to make retrieval consequential, not to add content.
- **H2 met, H1 not met.** Residents produce coherent continuity over an unreliable record. This is the most dangerous outcome, because it looks good in a demo and is false. It should halt product claims about world history.
- **Neither met.** The mechanism does not hold at n equals one object and will not hold at scale. That is a clear, cheap, early result.

---

8. Discussion

8.1 What this could defensibly own

The weakest available positioning is "NPCs powered by an LLM," which is easy to copy, supplied commercially, and hard to evaluate. The positioning this protocol supports is narrower and harder to copy: **a world whose history can be checked.**

That property compounds. One verified discovery can be told to a second resident, become a shared reference point, attach meaning to a location, and be cited weeks later. Each step adds to a record rather than to a transcript. A conversational NPC layer cannot produce this, because it has no authoritative record to be faithful to.

8.2 The strongest counter-argument

The counter-hypothesis is that persistent worlds retain people through other people, and that no quantity of AI resident activity substitutes for human social ties (Boellstorff, 2008; Nardi, 2010). If this is correct, verified agent history is a quality property of the simulation with limited bearing on retention, and Thirdwurld's true comparison class is ambient and spectator media, meaning idle games, virtual pets, and live streams, rather than social worlds.

This protocol does not resolve that question and should not be presented as if it does. What it does provide is a precondition test. If agent history cannot be trusted, the social-propagation path is closed regardless of which retention theory is right, because residents cannot reliably tell each other true things. Establishing verifiability is necessary for the thesis and clearly not sufficient for it.

An honest sequencing follows: establish verifiability with H1 and H2, then test resident-to-resident propagation, then and only then test human retention with adequate volume.

8.3 Why H3 is exploratory

Detecting a difference in seven-day return rate from a baseline near 0.20 to 0.30, at 80 percent power and conventional significance, requires roughly 300 sessions per arm. Thirdwurld's current volume is far below this. Running the comparison anyway would produce an estimate whose confidence interval spans both a large positive and a large negative effect.

Reporting such an interval as encouraging is the most common way this kind of work misleads. H3's instrumentation is retained because it is nearly free and because accumulated data may become informative later. Its results will be reported as an interval with the power limitation stated, and never as a headline.

8.4 Threats to validity

1. **Ceiling effect.** If the grounded arm is trivially correct because the retrieval prompt injects the ledger entry directly, H1 measures plumbing rather than model behavior. Retrieval must be genuine, competing with other memories, and the retrieval configuration must be reported in full.
2. **Vagueness as strategy.** A resident may raise its support rate by never asserting anything checkable. The unverifiable share exists to expose this and must be read as a primary number, not a footnote.
3. **Single-object generalization.** One coin may not generalize to social quests, contested objects, or multi-step goals.
4. **Implementation-specific behavior.** Resident conduct may reflect selector heuristics rather than any general property of agent architectures.

5. **Model drift.** Model version, latency, and cost changes across the two weeks confound comparisons. Model version is pinned and logged per event.
6. **Cohort carryover.** Residents carry memory across arms by design, so ordering effects are plausible. Counterbalancing is necessary and its adequacy is checked, not assumed.

---

9. Evaluation boundary

Residents are evaluated as software systems. Terms such as preference, memory, and choice refer to observable system behavior, not subjective experience.

Generated self-report is scored against verified records. No claim about consciousness, belief, or human-like memory is made.

---

10. Limitations

1. This is a protocol. No data has been collected and no result is reported.
2. Thresholds in Section 7.2 are judgment-based rather than derived from prior data, and are the most likely element to require pre-data revision.
3. The primary measures are automatic and therefore only as good as the ledger schema. Assertions the schema cannot represent are invisible to the metric.
4. H3 is underpowered by design and cannot support retention claims.
5. Findings about behavioral dispersion are findings about a disclosed configuration parameter, not about emergent agency.
6. Human judgment of resident quality remains vulnerable to novelty effects and anthropomorphic projection, which is a principal reason the primary measures avoid it.

---

11. Conclusion

The claim that AI residents make a world feel alive is not measurable. The claim that a world's history can be checked is.

This protocol tests whether a resident can notice an opportunity, choose to act on it, move through a physical space, produce an authoritative event, and later recall that event in a way a machine can verify. The prior literature supports the feasibility of each component and establishes the methodological requirement for a non-linguistic arbiter. It does not establish that people will return, and this work is structured so that the question of return is not quietly answered by a measurement that cannot answer it.

If the loop holds for one coin, there is a substrate worth building history on. If it does not, that is known in two weeks, from a ledger, at low cost.

References

1. Altera.AL, Ahn, A., Becker, N., et al. (2024). *Project Sid: Many-agent simulations toward AI civilization.* arXiv:2411.00114
2. a16z-infra. (2023). *AI Town.* Open-source deployable agent town. github.com/a16z-infra/ai-town
3. Boellstorff, T. (2008). *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human.* Princeton University Press.
4. Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D.-A., Zhu, Y., & Anandkumar, A. (2022). *MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge.* arXiv:2206.08853
5. Nardi, B. (2010). *My Life as a Night Elf Priest: An Anthropological Account of World of Warcraft.* University of Michigan Press.
6. Open Ended Learning Team, Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., McAleese, N., Bradley-Schmieg, N., Wong, N., Porcel, N., Raileanu, R., Hughes-Fitt, S., Dalibard, V., & Czarnecki, W. M. (2021). *Open-Ended Learning Leads to Generally Capable Agents.* arXiv:2107.12808
7. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative Agents: Interactive Simulacra of Human Behavior.* arXiv:2304.03442
8. Park, J. S., Zou, C. Q., Kamphorst, J., et al. (2024). *LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals.* arXiv:2411.10109
9. Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). *Curiosity-driven Exploration by Self-supervised Prediction.* arXiv:1705.05363
10. SIMA Team. (2024). *Scaling Instructable Agents Across Many Simulated Worlds.* arXiv:2404.10179
11. Vezhnevets, A. S., Agapiou, J. P., Aharon, A., Ziv, R., Matyas, J., Duéñez-Guzmán, E. A., Cunningham, W. A., Osindero, S., Karmon, D., & Leibo, J. Z. (2023). *Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia.* arXiv:2312.03664
12. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023). *Voyager: An Open-Ended Embodied Agent with Large Language Models.* arXiv:2305.16291

13. Wang, R., Yu, H., Zhang, W., Qi, Z., Sap, M., Neubig, G., Bisk, Y., & Zhu, H. (2024). *SOTOPIA-π: Interactive Learning of Socially Intelligent Language Agents.* arXiv:2403.08715

Citations 1, 4, 6, 7, 8, 9, 10, 11, 12, and 13 were verified against their arXiv listings; the author list for citation 1 is abbreviated here. Citations 3 and 5 are books and were not machine-verified.