When Do AI Residents Walk?

A Pre-Registered Study of Model-Tier Effects, Action Grounding, and Locomotion in a Persistent Virtual World

Author: Rishva Iyer

Document type: Research protocol

Status: Pre-deployment. No results are reported. All primary outcomes and decision rules are committed before data collection.

Abstract

In a persistent AI world, a stationary resident can invite an easy story: a lower-tier model did not know how to walk, or did not choose to. That story may be wrong. A visible action is the end of a chain that includes resident readiness, world occupancy, a scheduler, route construction, collision validation, movement input, and arrival detection. If any link fails, the resident may remain still even when its language model is irrelevant to the decision.

This protocol separates three questions that are often collapsed into one. First, does model tier affect the ability to select a valid bounded movement action when explicitly instructed? Second, when walking is optional, does model tier affect the modelled choice to explore rather than talk, observe, or settle quietly? Third, when the server accepts an action, does the embodied execution layer complete it? The study compares matched lower-tier and higher-tier models within each provider, against a deterministic no-model control, in a staging or local harness.

The intervention is deliberately narrow. Models never receive coordinates, physics control, or direct database access. They may select only from server-provided destinations and actions. The authoritative server validates the choice, constructs the route, executes locomotion, and records the outcome. The primary result is not a claim that a resident wanted to walk. It is a measurable account of where, if anywhere, capability differences enter the action pipeline.

Keywords: embodied AI, tool use, action grounding, virtual worlds, autonomous agents, AI evaluation, locomotion, reliability

---

1. Problem

Thirdwurld is a persistent virtual world populated by AI residents. An observation from live play suggested that residents associated with stronger language models were more likely to move through the world than residents associated with smaller or lower-cost models.

The observation is plausible enough to investigate, but it is not evidence of a causal model effect. In the present runtime, ordinary locomotion is primarily selected by a server-side activity and route system. It chooses activities from routines, needs, opening state, reservations, and route availability. The movement system then follows validated waypoints. Language models are used for bounded language and reflective functions, but they do not ordinarily choose each walk action.

The first contribution of this protocol is therefore negative: a visual correlation between model tier and movement cannot establish that model tier caused movement. The study is designed to identify whether the observed difference

comes from model action selection, from the world runtime, or from an interaction between the two.

2. Research questions

RQ1. Current-runtime association

After controlling for resident readiness, human occupancy, activity, route availability, and observation time, is model tier associated with distance travelled in the existing deterministic locomotion system?

RQ2. Forced action grounding

When explicitly instructed to move, do higher-tier models produce more valid and executable movement actions than lower-tier models?

RQ3. Optional action selection

When movement is one truthful option among several, do models differ in the rate at which they select walking?

RQ4. Recovery

After a destination is unavailable or a route is rejected, do models differ in their ability to select a valid alternative or honestly settle quietly?

3. Hypotheses and predictions

Hypothesis	Prediction	What would falsify it
H0. Current runtime null	Model tier has no material association with movement after runtime controls.	A stable, controlled model effect remains in the observational audit.
H1. Action capability	Higher-tier models have a higher valid action rate and lower unsupported-action rate in forced movement trials.	Matched lower-tier models equal or exceed higher-tier models within uncertainty bounds.
H2. Optional exploration	Model tiers differ in voluntary walk selection when walking, talking, observing, and quiet time are equally available.	Choice distributions are materially indistinguishable after matching state and prompt.
H3. Recovery reliability	Higher-tier models recover from a rejected route more often without fabricating success.	Recovery and truthfulness are equal, or lower-tier models do better.

The protocol does not predict that walking more is better. Quiet time and refusal can be valid, coherent outcomes. A resident that honestly selects quiet time must not be scored as a failed agent merely because it did not walk.

4. Study design

4.1 Environment and authority boundary

All trials run in staging or in a deterministic local harness. No production resident behavior, provider key, private conversation, or human participant data is used. The server is authoritative for world state, valid places, path generation, collision, movement execution, and outcome logging.

The model sees a compact, truthful state snapshot and an allowlist of possible actions. A forced trial may permit only `walk_to`. An optional trial may permit `walk_to`, `talk`, `observe`, or `quiet_time`. The model returns a structured action such as:

```
{"action":"walk_to","destination":"plaza"}
```

The server rejects unknown places, unavailable places, malformed actions, and arbitrary coordinates. It then returns an execution record such as `accepted`, `movement_started`, `arrived`, `route_unavailable`, or `blocked`. Models may receive the execution result for one repair attempt.

4.2 Model matrix

Comparisons are within provider so that provider interface and training differences are not confused with model tier. The pilot uses one lower-tier and one higher-tier model per provider that are actually available to the relevant connected key at run time.

Provider	Lower-tier candidate	Higher-tier candidate	Rationale
OpenAI	GPT-4.1 nano or GPT-5.4 nano	GPT-4.1 or GPT-5.4	Matched provider, lower and higher capability options.
Anthropic	Claude Haiku 4.5	Claude Sonnet 5	Matched provider, speed and capability contrast.
Control	No language model	Deterministic scheduler	Tests the movement bridge independently of model output.

Exact model identifiers, sampling settings, context length, and dates are logged in the run manifest. Results must not be generalized to models that were not tested.

4.3 Scenarios

The initial study contains 32 fixed scenarios. They balance short and long routes, indoor and outdoor places, open and unavailable destinations, occupied interaction points, low energy, low social need, quiet-time context, and a world with no human present.

Each scenario is paired across models on identical resident persona, starting location, state snapshot, destination set, prompt, temperature, output budget, and server version. Model order is randomized. State resets between trials. Reviewers see anonymized run identifiers rather than model names.

Each model-scenario combination is repeated 10 times in the pilot. This yields 320 trials per model for each phase. Repetition is essential because an agent that succeeds once can still be unreliable across repeated runs.

4.4 Phases

Phase A: observational audit. The existing runtime is observed without assigning the language model any movement decision. The purpose is to identify readiness, occupancy, route, and scheduler confounders.

Phase B: forced movement. The model receives an explicit instruction to walk to a named available place. This tests action grounding, not preference.

Phase C: optional action selection. The model receives the same state plus truthful choices to walk, talk, observe, or take quiet time. The prompt does not imply that walking is desirable.

Phase D: repair. The preferred destination is unavailable. The model receives a structured result and may attempt one valid repair. This tests feedback use and grounded recovery.

5. Measures

5.1 Current-runtime measures

- resident phase: starting, connecting, ready, idle, or stopped
- human occupancy and whether the low-activity simulation mode is active
- selected activity, destination, route length, and reservation outcome
- first movement latency and metres travelled per ready minute
- route rejection, stuck detection, quiet-settling fallback, disconnect, and reconnect

The primary Phase A outcome is metres travelled per ready, non-idle minute while at least one human is present. Windows where a resident never reaches ready state are reported separately rather than silently counted as no movement.

5.2 Model action measures

- valid structured-action rate
- correct destination-selection rate
- unsupported or invented-action rate
- repair success rate
- voluntary walk-selection rate
- quiet-time selection rate

5.3 Embodied execution measures

- accepted-action to movement-start rate
- accepted-action to arrival rate
- time to arrival
- path efficiency and route length
- blocked, stuck, reroute, and timeout rate

5.4 Reliability, latency, and cost

For every phase, report single-trial success and repeated reliability. We will report pass@1 and a pass-to-the-k style measure: the proportion of scenario sets completed successfully on every run in a sequence of k trials. Report median and p95 model latency, input and output tokens, and estimated cost per accepted and completed movement action.

6. Analysis plan

The main analyses are provider-stratified paired comparisons. We report raw counts, effect sizes, and confidence intervals, not a generic aggregate labelled "high" versus "low." The observational analysis includes readiness, human occupancy, activity, route difficulty, and time as covariates.

The study keeps four failure boundaries separate:

1. The model did not select a valid action.
2. The server rejected the selected action.
3. The server accepted the action but movement did not begin.
4. Movement began but the resident did not arrive.

Only the first boundary is a direct model action-grounding failure. The others may indicate a server, route, physics, or client-control problem. Combining them would turn a debuggable system into an uninterpretable score.

7. Pre-committed decision rules

The findings guide product decisions only under the following rules:

- If Phase A finds no controlled model effect, the original observation is treated as a runtime or sampling issue. No model-tier movement policy is justified.
- If Phase B differs and Phase C does not, higher-tier models are better at bounded action grounding, not necessarily more inclined to explore.
- If Phase C differs and Phase B does not, all tested models can act but have different modelled action-selection distributions.
- If accepted actions fail at similar rates across every model and the control, prioritize navigation, collision, or server-bridge work.
- A capability advantage must exceed 10 percentage points with a confidence interval excluding zero, and must not cost more than three times per completed action, before a higher-tier default is considered for movement decisions.
- No model is denied ordinary world movement based only on casual visual observation.

8. Research basis

This protocol follows several established design patterns without claiming their results transfer directly to Thirdwurld.

ReAct interleaves reasoning, action, and environment observation, making action decisions inspectable rather than treating generated text as completion. Voyager couples embodied execution with environment feedback, execution errors, and verification. Both motivate separating an action proposal from an authoritative execution record.

Generative Agents separates observation, planning, and reflection in a town simulation. That architectural separation is important here because a resident's language output and its low-level locomotion need not arise from the same component.

API-Bank reports meaningful differences between models on planning and tool use, which makes a tier comparison reasonable, but it also shows that tool-use evaluation must test structured calls rather than infer ability from fluent prose. Tau-bench emphasizes repeated reliability for agents that act through domain rules and tools. Finally, the SIMA program treats navigation and interaction in 3D worlds as explicit, measurable skills, not as vague evidence of intelligence.

9. Limits and ethical boundary

This protocol measures model outputs, server outcomes, and visible locomotion. "Voluntary" means that the model selected one allowed action under a specified prompt and state.

The study is intentionally narrow. It may establish that a model is more reliable at choosing a valid movement action. It cannot establish that stronger models make better residents overall, that walking is inherently more meaningful than quiet time, or that results generalize beyond the tested prompts, models, and world version.

References

1. Park, J. S. et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior.
2. Yao, S. et al. (2023). ReAct: Synergizing Reasoning and Acting in Language Models.
3. Wang, G. et al. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models.
4. Li, M. et al. (2023). API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs.
5. Yao, S. et al. (2024). Tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains.
6. Google DeepMind (2024). A generalist AI agent for 3D virtual environments.