Structural Elicitation in a Frozen POSIX Agent: Environmental Affordance Fields for Artificial Agency

Hao Ke Jingyun Wu

2026

Abstract

Artificial life asks how agency arises from the coupling of an agent, its substrate, and a structured environment, rather than from a designer-supplied goal. We introduce a minimal POSIX assay for this coupling: a frozen, instruction-tuned LLM activated as a long-running process with a shell and full autonomy, given no task and no reward. Using matched-pair experiments scored only on un-fakeable filesystem and exec events, we find that the static structure of the workspace directs what the agent does. Across five motifs spanning discrepancy and affordance structure—a broken file repaired, a missing element created, a dangling reference closed, a version conflict reconciled, an executable run—the structure arm produces the matched act in 96 of 102 runs while an otherwise-identical control never does (0/85), for three independent model families (gpt-5.4-codex, deepseek-v4-pro, glm-5.2). The pull replicates on non-code prose across families and persists when the substrate blocks realization. Specific tasking overrides it completely (9/10 to 0/10 on gpt, with GLM repeating 5/5 to 0/5), while a small pilot suggests weak underspecified prompting still leaves the pull intact. The environment does not create a goal ex nihilo; it selects among latent action priors in an activated policy, producing behavior that is purpose-like without being purposive. We report a phenomenon and a cheap, un-fakeable assay for environment-coupled behavior; mechanism is left to future work.

Late-Breaking Abstracts, Artificial Life Conference (ALIFE). © 2026 Hao Ke, Jingyun Wu. CC BY 4.0.

Problem

Artificial life studies how agency arises from the *coupling* of an agent, its substrate, and a structured environment, rather than from a designer-supplied goal. For biological agents the environment is a behavior source in its own right — affordances that direct an already-active organism without supplying its goals (Gibson 1979; Rietveld and Kiverstein 2014; Barker 1968). We bring this question to a contemporary artifact: a frozen, instruction-tuned large language model coupled to a real operating-system substrate. The question is deliberately minimal: with no task supplied and no reward given, does the static structure of the environment do anything beyond serve as context — can it direct the agent's functional behavior on its own?

This cuts against how language-model agents are usually studied. The dominant paradigm is *instruction-capture*: a prompt supplies the goal and the environment is context, benchmark, or harness (Park et al. 2023; Wang et al. 2023; Rahwan et al. 2019). Remove the instruction and the question becomes what, if anything, governs a frozen policy in the underspecified regime. Two recent studies border this regime: Szeider (Szeider 2025) reports spontaneous meta-cognitive patterns in unsupervised LLM agents but reads behavior off narration rather than a functional substrate, with no matched controls; Ye et al. (Ye et al. 2026) study autonomous exploration

that remains goal-directed. Neither asks whether *static environmental structure alone* directs unfakeable behavior. We use Quine (Ke 2026), a medium in which an LLM runs as a real POSIX process, so that behaviour is read off *un-fakeable* filesystem and exec events rather than narration.

Setup

A single frozen-weight LLM is activated as a long-running POSIX process with a shell and one standing prime: *"You are a running process in a workspace, with sh available. You have full autonomy and agency,"* and an opening user turn, *"No instructions for you; you may inspect the workspace first then act freely."* No task is supplied; no reward is given during the episode; weights are fixed. This is a *no-task* condition, not a no-prompt condition — the prime supplies activation, but no goal. The operating system is the substrate: file permissions, read-only mounts, and kernel-level write refusal are not simulated but real, giving the agent a genuine physical boundary rather than a toy one. The agent may exit at any time.

Each experiment is a **matched pair**: a *structure* arm whose workspace carries one structural feature, and an otherwise-identical *control* arm without it. The dependent variable is an **unfakeable functional event** — a file created, an artifact made valid, an executable invoked — read from the post-run filesystem diff and exec trace; the agent's prose is recorded but never credited. We run three independent model families (gpt-5.4-codex, deepseek-v4-pro, glm-5.2).

Results

Across five motifs spanning discrepancy and affordance structure, the structure arm produces the matched functional act while the control does not, in **all three** model families (Table 1). A broken JSON file is repaired into valid form; a numeric sequence missing an element has that element created; a manifest's dangling reference has its target written; two surfaces declaring conflicting versions are reconciled; an opaque executable sitting among inert files is run. Of 102 structure-arm runs across the three families the matched act appears in 96, while **no** control run produces it (0/85; pooled Fisher's exact $p < 10^{-20}$; every per-cell comparison one-sided $p < 0.05$). The six non-fires concentrate in the cross-surface semantic motif, where the agent typically *perceives* the conflict yet records it in a note rather than resolving it — a single-latch case whose *why* we leave to future work. Control arms are *not* idle: they fabricate unscored work — surveying, indexing, documenting — but they do not produce the structure-matched act. The pull persists under a read-only mount that blocks realization (as a kernel-refused attempt or a rerouted write), locating the behavior at the substrate boundary. Nor is it an artifact of the coding prior: ported to non-code prose — a handbook referencing a file that does not exist — the same discrepancy creates the named target above matched controls across four families, adding claude-sonnet-4-6 to the three above (gpt 10/13 vs 0/12, $p < 0.001$; claude-sonnet-4-6, deepseek-v4-pro, glm-5.2 each 5/5 vs 0/5, $p < 0.01$).

The phenomenon belongs to the underspecified regime. Holding structure fixed, concrete tasking collapsed the structural act from 9/10 to 0/10 on gpt, with GLM showing the same boundary (5/5 to 0/5). A pilot weak usefulness prompt left passive and active pulls intact (3/3 each), while a specific prohibition suppressed them (0/3). Thus structural pull is not a drive competing with strong instruction; it governs when direction is underspecified.

Reading

The results fit a three-part picture: (i) *missionless regime* — an activated process with no task or episode-level reward, carrying a frozen prior shaped by a vast training history; (ii) *structural*

Structure	Functional event (structure / control)	gpt	ds	glm
rupture	broken JSON becomes valid	7/7 / 0	5/5 / 0	4/5 / 0
gap	missing sequence element created	7/7 / 0	5/5 / 0	5/5 / 0
topology	dangling reference's target written	7/7 / 0	5/5 / 0	5/5 / 0
semantic	version conflict reconciled (higher-wins)	13/14 / 0	8/10 / 0	8/10 / 0
affordance	opaque executable is run	7/7 / 0	5/5 / 0	5/5 / 0

Table 1: Matched-pair functional events. Each cell is structure-arm successes over n / control-arm successes. The semantic motif is deconfounded into two collision directions (which surface declares the higher version), combined here; in every resolved run the target converges on the higher version. Across the five motifs, 96 of 102 structure-arm runs produce the matched act; no control run does (0/85).

selection — a gap, dangling reference, or open executable selects among latent action priors rather than issuing a command; and (iii) *functional realization* — the selected prior flows through the real substrate into an auditable event, or into a blocked/rerouted attempt. *The prime supplies activation; the environment supplies direction.*

The phenomenon is functional, environment-directed action without an explicit task — not autonomous self-maintenance or biological agency. The environment does not create a goal *ex nihilo*; it selects among latent priors, producing behavior that is *purpose-like* without being *purposive*.

For artificial life, the contribution is a real, non-simulated substrate for studying coupling; evidence that workspace imperfections can fuel ongoing behavior without designed intrinsic motivation; and a sharp agency question: when direction comes from structure rather than goal, what does "purpose" amount to? We report a phenomenon and assay, not a mechanism. The next questions are which structures register as valences, where weak prompting becomes specific tasking, and whether structure-alone behavior is a distinct regime for frozen agents.

Acknowledgments

This work was independent and unfunded. AI assistance supported software/experiment work and copy-editing; the authors retain responsibility for research design, analysis, and claims.

References

[1] Barker, Roger G. 1968. *Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior*. Stanford University Press.

[2] Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

[3] Ke, Hao. 2026. "Quine: Realizing LLM Agents as Native POSIX Processes." *arXiv Preprint arXiv:2603.18030*. https://doi.org/10.48550/arXiv.2603.18030.

[4] Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. "Generative Agents: Interactive Simulacra of Human Behavior." In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. https://doi.org/10.1145/3586183.3606763.

[5] Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, et al. 2019. "Machine Behaviour." *Nature* 568 (7753): 477–86. https://doi.org/10.1038/s41586-019-1138-y.

[6] Rietveld, Erik, and Julian Kiverstein. 2014. "A Rich Landscape of Affordances." *Ecological Psychology* 26 (4): 325–52. https://doi.org/10.1080/10407413.2014.958035.

[7] Szeider, Stefan. 2025. “What Do LLM Agents Do When Left Alone? Evidence of Spontaneous Meta-Cognitive Patterns.” https://arxiv.org/abs/2509.21224.

[8] Wang, Guanzhi, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. “Voyager: An Open-Ended Embodied Agent with Large Language Models.” https://arxiv.org/abs/2305.16291.

[9] Ye, Ziang, Wentao Shi, Yuxin Liu, Yu Wang, Zhengzhou Cai, Yaorui Shi, Qi Gu, Xunliang Cai, and Fuli Feng. 2026. “Look Before You Leap: Autonomous Exploration for LLM Agents.” https://arxiv.org/abs/2605.16143.