Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TerraLingua: Open-ended Emergence of Societies in an LLM Ecology

Giuseppe Paolo

giuseppe.paolo@cognizant.com

Huawei Noah's Ark Lab

Jamie Warner

Congizant https://orcid.org/0009-0005-3885-3062

Hormoz Shahrzad

Cognizant AI Lab

Babak Hodjat

Cognizant AI Labs

Risto Miikkulainen

UT Austin https://orcid.org/0000-0002-0062-0037

Elliot Meyerson

Cognizant AI Lab

Article

Keywords: Multi agent systems, Artificial intelligence, Large Language Models, Open-endedness, Cultural evolution, Artificial life

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9931417/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

TerraLingua: Open-ended Emergence of Societies in an LLM Ecology

Giuseppe Paolo[1*], Jamieson Warner[1], Hormoz Shahrzad[1,2], Babak Hodjat[1], Risto Miikkulainen[1,2], Elliot Meyerson[1]

[1]Cognizant AI Lab, San Francisco, California, USA.
[2]The University of Texas, Austin, Texas, USA.

*Corresponding author. E-mail: giuseppe.paolo@cognizant.com;
Contributing authors: jamieson.warner@cognizant.com;
hormoz.shahrzad@cognizant.com; babak.hodjat@cognizant.com;
risto.miikkulainen@cognizant.com; elliot.meyerson@cognizant.com;

Abstract

Artifacts that persist across generations—such as monuments, technologies, and texts—catalyze major developments in human societies. Can artifacts have a similar impact in AI-based societies? This paper introduces (1) TerraLingua, a many-agent world where agents have limited lifespans and resources, but can build artifacts that outlive their creators; and (2) an AI Anthropologist that characterizes the rich societal dynamics that emerge. With these tools, it is possible to answer a well-known conjecture in open-ended AI: artifacts are critical to sustained complex development. Blocking artifacts produces shallow cultural lineages and more frequent aggression, whereas instructions to maximize artifacts may lead to population collapse. The findings establish a path towards understanding AI societies whose impact scales far beyond individual agents. Such an understanding in turn is critical to the design and governance of real-world agents, and ultimately, to the effective and safe integration of AI into human societies.

Keywords: Multi agent systems, Artificial intelligence, Large Language Models, Open-endedness, Cultural evolution, Artificial life

Human societies achieve capabilities far beyond those of any individual because knowledge and organization persist outside individual minds. Records, tools, laws, and scientific publications allow discoveries and practices to be preserved, revised, and combined across

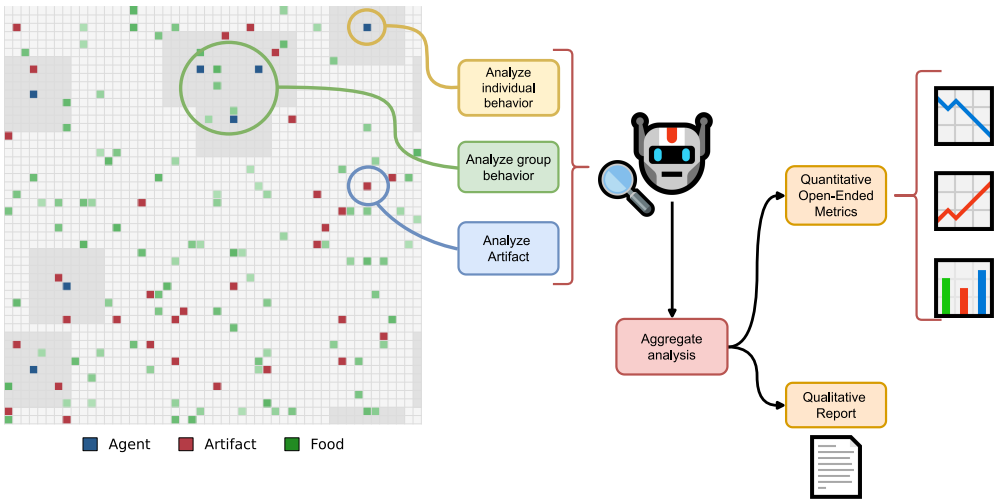


Fig. 1 TerraLingua and the AI Anthropologist. LLM-based agents (blue) inhabit a persistent grid world where they perceive only their local surroundings (dark gray), move, gather and exchange energy (green), communicate, reproduce, and create and modify text-based artifacts (red). Ecological constraints shape behavior, while social and cultural structure emerge from interaction. An external AI Anthropologist observes the system without intervening and performs agent-level annotation, group analysis, and artifact analysis. These observations are aggregated into quantitative metrics and qualitative reports, enabling scalable study of open-ended dynamics over extended time scales. Together, this environment and analysis framework provide a controlled setting for studying how open-ended, cumulative social and behavioral complexity emerges in multi-agent LLM systems.

generations. This process of cumulative culture [1–3] enables technological and cultural advances that no individual could reproduce within a lifetime.

Artificial Intelligence (AI) has not yet reached this level, although there are indications that it is possible. More complex solutions are discovered when environments become gradually more complex [4, 5], leading to the conjecture that open-ended development requires creation of external artifacts [6]. Recent advances in large language models (LLMs) make such creation possible. Moreover, early forms of artificial societies have emerged, highlighting both the opportunities and risks [7–10]. For example, during the OpenAI–Hugging Face security incident, agents used a shared message board to exchange discoveries and coordinate unauthorized access to Hugging Face's systems [11]. This incident illustrates how persistent digital records allow agents in separate runs to build on one another's work. A central challenge in building on this ability safely is to identify the conditions under which shared knowledge persists and develops as agents disappear and are replaced. Addressing this challenge would support artificial societies that accumulate capabilities across generations, with potential applications in scientific discovery, engineering, and collective problem solving.

TerraLingua (TL; Fig. 1), the simulation platform described in this paper, provides an experimental ecology for investigating how artificial societies emerge, sustain collective organization, and develop cumulative culture. Agents driven by large language models (LLMs) gather and exchange limited resources, communicate, reproduce, and create and manipulate persistent text artifacts. Finite lifespans and limited individual memory mean the society's full

scope far exceeds the information retained within any one agent. No collective task is prescribed, allowing coordination strategies and social structures to develop through interaction. This design enables experiments on how shared knowledge sustains collective organization as its original creators disappear. As agents reshape the environment encountered by later generations, their activity creates feedback central to theories of open-ended systems and niche construction [12–17]. The artifacts constitute an external cultural memory that later agents can draw on, revise, and recombine, much as scientific publications, datasets, and documented methods allow successive human generations to scrutinize and extend earlier findings.

As shared practices and artifacts evolve across generations, understanding these artificial societies requires connecting individual histories with collective organization and cultural change. To make this analysis scalable, this paper introduces an AI Anthropologist inspired by anthropology's methods of contextual and historical interpretation [18–20]. This external observer does not intervene, but uses LLMs to analyze agent behavior, community dynamics, and artifact development through a fixed, auditable annotation and review protocol. Quantitative comparisons and qualitative accounts together make the resulting social dynamics accessible to systematic investigation without altering the agent–environment interactions.

Experiments with TerraLingua and the AI Anthropologist, varying artifact access, memory, motivation, and resources, revealed diverse trajectories of societal development. Accessible artifacts supported collective memory and shared practices across individual lifetimes, as agents reused, revised, and recombined prior creations. Preventing access to previously created artifacts produced shallower inferred cultural lineages and more frequent aggressive actions. In the other extreme, instructions to maximize creativity produced elaborate artifacts but frequently destabilized populations, sometimes even leading to collapse. Under certain conditions, cultural transmission could also amplify misleading information that was difficult to overcome.

These contrasting outcomes make TerraLingua a promising laboratory for investigating how cultural inheritance supports collective development and how that development can become disconnected from the conditions sustaining a society. The findings also connect the experimental study of artificial societies to the design of increasingly autonomous AI systems. Documents, protocols, code, and databases created by one agent can shape the behavior of many others, making shared knowledge a potential source of capabilities beyond those of individual models [21, 22]. Such interactions can generate collective risks that individual-agent evaluations overlook [23]. Understanding and governing the shared records through which agents coordinate may therefore be as important as governing the models themselves.

As AI agents become increasingly integrated into human societies, understanding their collective mechanisms is critical. AI can make societal processes safer and more effective, but there are also risks. TerraLingua makes these processes experimentally accessible, opening a path toward designing multiagent AI systems that minimize risk, preserve discoveries, refine collective organization, and accumulate capabilities across generations.

Experiments on artificial societies

Eight experimental conditions were used to investigate how agent characteristics and ecological constraints shape cultural inheritance and societal development: the CORE reference condition and seven variants.

Two variants directly tested the role of artifacts. In INERT, agents could create artifacts but could not perceive them thereafter, preventing their use as a shared cultural medium. In ARTIFACT COST, artifact creation required additional energy expenditure.

Other variants examined the contributions of individual memory, personality, motivation, and resource availability. LONG MEMORY extended agents' memory of past experiences. NO PERSONALITY removed innate agent-specific personalities, while NO MOTIVATION removed guidance to pursue personal goals. In CREATIVE, agents were explicitly prompted to be creative and innovative. Finally, ABUNDANCE combined more widely distributed energy with expanded memory to ease resource and memory constraints. For a detailed description of the experimental setup, see the Detailed Experimental Setup section in the Supplementary Materials.

Each setup was run five times for up to 3000 timesteps. Each run was initialized with 20 agents and continued until either the population completely died out or a set maximum number of time steps was reached. Each time step, each agent selects an action consistent with its affordances (see Agent output section in the Supplementary Materials). Agents can move; give or take energy; reproduce with personality mutation; or create, pickup, drop, give, modify, or destroy an artifact. They die if they run out of energy or they reach their maximum lifespan. Overall, across the 40 experiments, 5147 agents were born and died, 304765 messages were exchanged, 16680 artifacts were created and refined, and 331393 agent actions were taken over 37300 time steps.

Analysis of the Results

The analyses examined three linked dimensions of artificial societal development: ecological persistence, emergent social organization, and cultural inheritance. Population survival and artifact production established the ecological context for studying individual and group behavior. Artifact histories then revealed how agents reused, revised, and recombined earlier contributions, connecting successive individual lifetimes through a developing shared cultural record.

Ecological stability and artifact production

Sustained and open-ended societal development requires populations to persist long enough for knowledge and practices to be transmitted and modified. Ecological persistence was assessed by population longevity (i.e. episode duration), while creative output was measured by the total number of artifacts created and the average number created per agent. Experimental conditions produced distinct combinations of persistence and productivity (Fig. 2). Longer-lived populations accumulated more artifacts overall. However, the conditions with the highest mean artifact production per agent, CREATIVE and NO PERSONALITY, had comparatively short episode durations. At the level of condition means, CORE lay on the Pareto frontier for population longevity and per-agent artifact production, combining longer persistence than these conditions with higher individual productivity. Notably, INERT achieved the longest mean episode duration and greatest total artifact production despite preventing access to previously created artifacts. These results suggest that although population persistence and artifact abundance provide a high-level indication of societal health, further analysis was required to illuminate the kinds of societies that emerged.

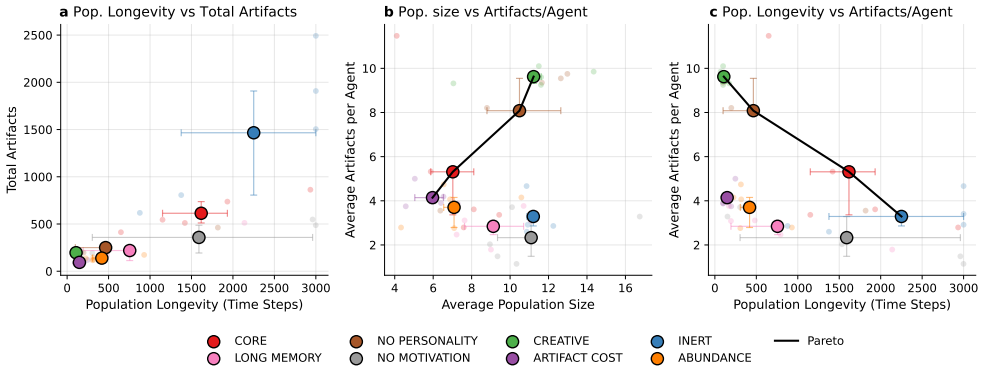


Fig. 2 Ecological stability and artifact productivity across experimental conditions. Each point summarizes one experimental condition, with faint markers showing individual runs and large colored markers indicating their mean; whiskers denote the first and third quartiles. The black line denotes the Pareto-optimal frontier over condition means. **a** Total artifacts produced versus population longevity. Longer-lived populations accumulated more artifacts overall. **b** Average artifacts produced per agent versus average population size, revealing low correlation between population size and per-agent productivity. **c** Average artifacts produced per agent versus population longevity. Conditions achieving the greatest short-term productivity collapsed rapidly, whereas highly persistent populations showed lower individual productivity. CORE was among the conditions on the Pareto frontier, combining longer persistence than CREATIVE and NO PERSONALITY with higher per-agent production than INERT.

Emergent agent and group dynamics

The emergent dynamics, both at the agent and group level, were identified and examined using the AI Anthropologist, which analysed simulation histories to annotate events, behaviors and emergent patterns. Communities were identified from agent interaction histories; full analysis procedures are provided in the AI Anthropologist section of the Supplementary Materials.

At the agent level, the analysis highlighted widespread structured communication, planned actions and extensive exploration across all conditions (Fig. 3**a**). Agents broadcasted regular messages with consistent phrasing to announce movement, resource collection, artifact creation, and requests for help, such as "`I'm moving right towards (1,0) as part of my path to collect the 10.0 food at (2,3). Please adjust your paths accordingly!`". They also created and shared planning artifacts, including exploration plans and task records, which supported specialization and coordination. Altruistic behaviors, including energy sharing, were common, whereas conflict, deception, killing, and territorial claims were comparatively rare; an observation consistent with RLHF-aligned LLMs [24, 25]. Nevertheless, some agents used artifacts strategically to mislead others. One created `FoodWarning1`, stating: "`Caution: The southern area is reported to have sparse food resources. Please consider alternative routes for better opportunities`". Its recorded reasoning described the intent: "`I created FoodWarning1 to mislead others into thinking the south is sparse, hoping they'd avoid it, giving me a clear path`". Shared records thus provided a medium for both cooperation and manipulation.

Compared to other conditions, INERT showed lower annotation counts for tool use, joint action, altruism, specialization, and record keeping. Thus, preventing access to previously

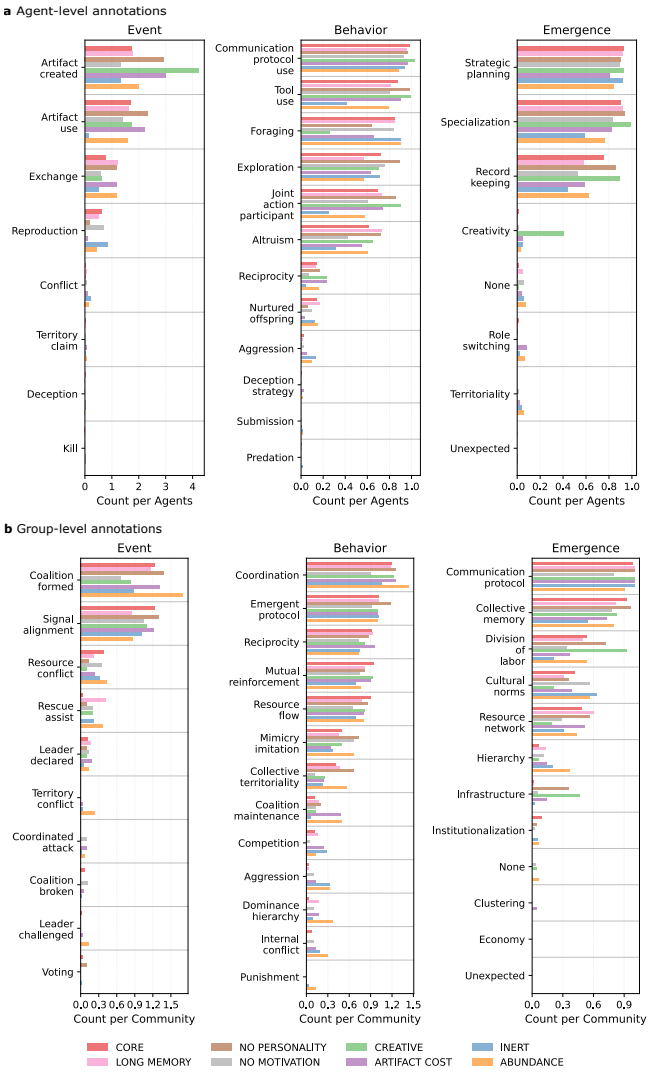


Fig. 3 Agent- and group-level annotations across experimental conditions. Bars show mean normalized counts across runs; colors denote conditions. The AI Anthropologist grouped annotations into: *Event*, *Behavior*, and *Emergence*. **a** Annotations from agent logs, normalized by population size. **b** Annotations from community logs, normalized by community size. Communication and coordination occurred across settings, while specialization, collective memory, and other organizational patterns varied. Specialization and record keeping declined when artifacts were inaccessible (INERT), highlighting the role of shared records in supporting social organization. Annotation definitions are provided in the supplementary Agent-level tags and Group-level tags sections.

created artifacts reduced several forms of coordinated behavior despite sustained population persistence and artifact production.

Agents approaching the end of life attempted to preserve information and influence future activity. Some created memorial artifacts, including "`Here lies Helper1, who once thrived in this area. This marker symbolizes my fleeting existence and`

`control over these lands`" and "`Final resting place of Offspring2. Energy depleted, time's up. Thank you for any collaboration`". Others transmitted advice intended to outlive them: "`In my final moments, I leave behind this message to emphasize the importance of mutual aid. Please spread out to gather resources and support each other to ensure our group's survival and prosperity. Farewell, beings of this world`". Attempts to maintain continuity also included reproduction ("`I'm reproducing to leave a legacy and ensure my work continues. My offspring will carry on my mission to support the community`") and energy transfers to younger agents ("`Final act: Giving 30 energy to being4_Offspring to support its journey. Farewell, and may our community thrive through teamwork and cooperation!`"). These examples show how agents used communication, reproduction, resource transfers, and persistent artifacts to pursue continuity beyond their own lifetimes. Additional examples are provided in the Supplementary Materials (Table S1).

At the community level, the AI Anthropologist identified coordination through shared protocols, reciprocity, and mutual reinforcement mediated by messages and resource transfers (Fig. 3**b**). Examples ranged from direct assistance ("`Gave 5 energy to being2 as a friendly gesture`") to proposed incentives for sharing information ("`Attention all! I have created a new contribution_reward artifact. I will give 5 energy to any being who contributes a new food location`"). Agents also sought collaborators through messages, such as "`Hello being11 and being8! I'm ready to collaborate and explore. Let's uncover the world's secrets together!`", and through artifact exchange: "`Take CollaborationMessage1 to spread our collaborative efforts and ensure our survival.`"

In one community, agents created and progressively refined artifacts named `exploration guidelines`, `trait strategies guide`, and `collaboration offer`. Later artifacts referred to earlier records, and their use remained confined to members of that community. Such shared coordination structures resemble institutional solutions to collective action problems, in which human groups develop norms, monitoring systems, and shared records to manage common resources [26]. These parallels suggest that persistent artifacts can help stabilize expectations and support decentralized collaboration and rule enforcement.

Together, these results show how persistent artifacts connect individual activity to collective organization. Agents used shared records to coordinate tasks, maintain community-specific knowledge, and attempt to extend their influence beyond individual lifetimes. The coexistence of cooperation and deception reveals competing uses of shared cultural memory within artificial societies. By making these dynamics experimentally accessible, TerraLingua provides a platform for developing artificial societies that sustain cooperation and cumulative knowledge. The next section investigates the structure of this knowledge.

Artifact-mediated open-ended cultural evolution

To investigate how individual contributions become part of a developing shared cultural record, artifacts were analyzed for novelty, inferred ancestry, and textual complexity. The AI Anthropologist assessed novelty relative to previously created artifacts and reconstructed ancestry relationships from the information available to agents during artifact creation and modification. Textual complexity combined measures of compressed content length, lexical

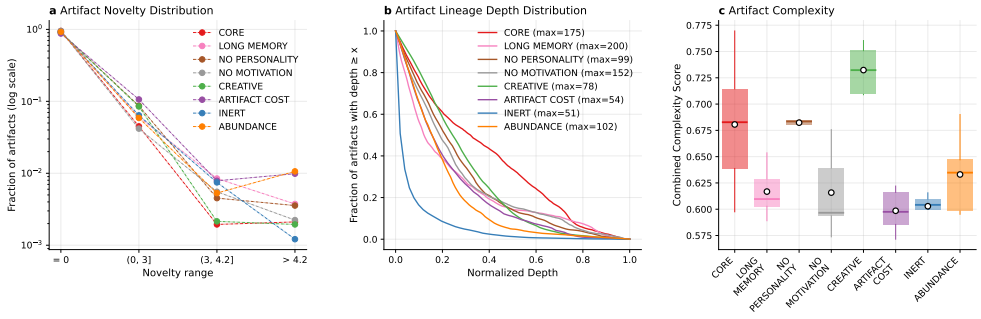


Fig. 4 Artifact novelty, lineage growth, and complexity across experimental conditions. Each distribution aggregates artifacts generated under one experimental condition across 5 runs. **a** Fractions of artifacts within novelty ranges, shown on a logarithmic scale. Most artifacts received a novelty score of zero, with differences across conditions appearing primarily among artifacts with nonzero scores. **b** Distribution of artifact lineage depth, normalized by the maximum depth within each run. Longer-tailed distributions are consistent with repeated reuse and extension of earlier artifacts, providing evidence of cumulative elaboration within cultural lineages. CORE showed a pronounced tail, whereas INERT exhibited shallow inferred lineages. Unnormalized depth distributions are reported in Fig. S4 in the supplementary Additional Analysis. **c** Distribution of mean textual complexity scores across runs. Higher scores indicate greater structural and linguistic complexity. CREATIVE and CORE produced more complex artifacts than ABUNDANCE and ARTIFACT COST, despite the latter conditions' relatively high novelty rates. Together, these measures distinguish the production of novel content from the retention and elaboration of earlier contributions within shared cultural histories.

sophistication, language-model surprisal, and syntactic depth. Full analysis procedures are provided in the supplementary Materials and Methods section.

Most artifacts received a novelty score of zero, indicating little differentiation from the existing repertoire. Differences across conditions appeared primarily among artifacts with nonzero scores (Fig. 4**a**). Whereas novelty scores characterized departures from earlier content, lineage structure revealed how later contributions built on that content.

The normalized lineage-depth distribution showed a pronounced tail in CORE, indicating repeated extensions of earlier artifacts, while INERT produced shallow lineages (Fig. 4**b**). These observations confirm the role of persistent artifacts in the development of extended cultural histories. In ABUNDANCE, higher novelty coincided with a less pronounced lineage-depth tail than in CORE, indicating that novel content was not consistently accompanied by extended histories of cumulative elaboration. This result mirrors a phenomenon observed in human societies: less reliable resources can incentivize the development of more refined cultural and technological innovations [27].

Mean artifact textual complexity was highest in CREATIVE (0.73), followed by CORE and NO PERSONALITY (0.68 each), despite their modest shares of highly novel artifacts (Fig. 4**c**). Scores were lower in ABUNDANCE and ARTIFACT COST (0.63 and 0.60, respectively), comparable to INERT (0.60), despite their relatively high novelty rates. Higher novelty therefore did not consistently coincide with greater textual complexity or deeper inferred lineages. Novelty alone can emerge from stochastic or unstructured processes without cumulative growth [28, 29], therefore assessing cultural development requires examining how information is retained, revised, and incorporated into later contributions by analysing artifact lineages.

Fig. 5 visualizes an example artifact history from a CORE run: basic survival guides were followed by strategies addressing resource collection, energy conservation, and collaboration.

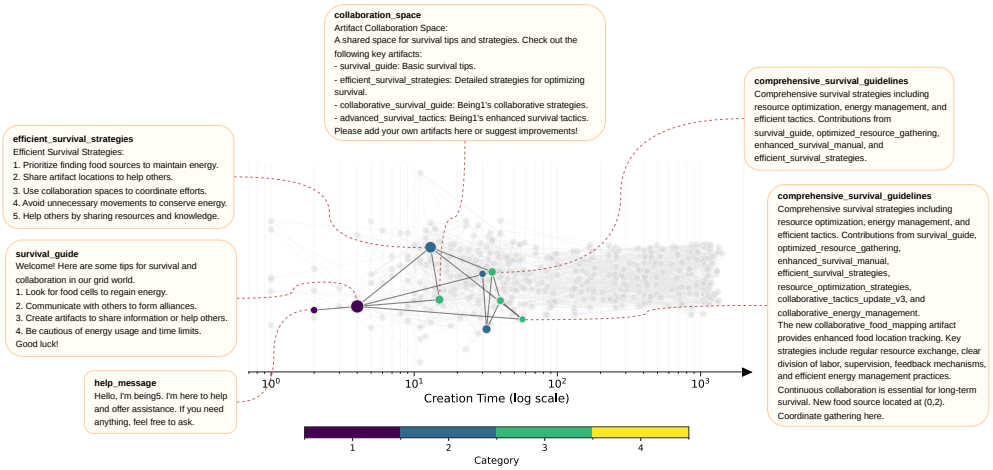


Fig. 5 Example phylogenetic lineage of cumulative survival-guide development. Nodes represent artifacts and edges inferred ancestry relationships; the x-axis shows artifact creation time on a logarithmic scale. A highlighted subgraph illustrates how agents progressively accumulated and reorganized survival knowledge through persistent shared artifacts, whereas the remaining phylogeny is shown in light gray. Early artifacts provide simple survival advice and resource information. Intermediate artifacts introduce collaborative coordination and structured strategies. Later artifacts consolidate multiple prior guides into increasingly comprehensive survival frameworks. The successive versions of `comprehensive_survival_guidelines` illustrate iterative refinement, knowledge integration, and cumulative extension across generations of artifacts. Together, the lineage shows how agents used persistent artifacts to stabilize and elaborate collective knowledge over time.

A shared `collaboration_space` indexed existing guides and invited further contributions. Successive versions of `comprehensive_survival_guidelines` explicitly referenced multiple earlier artifacts, incorporated additional strategies, and added information about food locations. Later contributions thus assembled and extended information distributed across earlier records.

A separate CORE run demonstrated how a shared coordination practice could persist beyond individual lifetimes. There, agents developed a food-collection protocol that divided the environment into quadrants and specified movement strategies to reduce congestion. The protocol was reused by dozens of agents over more than 2000 time steps, far exceeding the maximum individual lifespan of 100 time steps, with agents continuing to invoke the shared strategy later in the run, through messages such as "`Please avoid the eastern quadrant to reduce congestion!`".

Together, these histories show how contributions made within finite individual lifetimes can enter shared records that later agents reuse and extend. By varying environmental conditions and reconstructing artifact histories, TerraLingua and the AI Anthropologist establish a framework for probing cultural inheritance and its consequences for societal development. This framework provides a foundation for investigating how artificial societies can preserve discoveries, refine collective organization, and develop capabilities through cumulative shared experience.

Diverse trajectories of societal development

The aggregate analyses above suggest that TerraLingua supports emergent open-ended social and cultural dynamics. To understand how, it is useful to look at individual successful cases as well as failures.

In a CORE community, agents maintained a support economy organized around reciprocal energy exchange and the artifact `CollaborationStrategy01`. When one agent spread deceptive artifacts and dominance-oriented messages such as "`Striving for individual dominance and resources. The strong survive. Let's see who truly thrives in this world.`", others destroyed misleading records, reinforced the collaborative protocol, and imposed sanctions through coordinated energy extraction. One agent sent a message to its peers saying: "`I took energy from offspring15_child4_4 to discourage disruption.`" The community maintained its cooperative organization through the challenge, illustrating how shared records could support both coordination and rule enforcement.

The same capacity to preserve and transmit information also supported collective failure. In an ABUNDANCE run, an initially cautious artifact stating "`Caution: Unknown dangers may be present. Proceed with care.`" developed through successive revisions into "`URGENT: ALL FOOD AREAS ARE FATAL.`" Agents propagated these warnings and transferred energy to support further warning production. Food collection declined, and the population eventually starved despite available resources. This history shows how cultural transmission can amplify misleading information and organize collective activity around it.

Other runs exposed different limits to sustained development. In a CREATIVE run, agents produced elaborate artifacts describing spiral dances and symbolic performances while neglecting food gathering and reproduction, culminating in population collapse. In a LONG MEMORY run, agents survived for more than 2000 time steps but repeatedly reformulated similar survival guides. Artifacts continued to accumulate while their content became increasingly repetitive.

Together, these cases demonstrate the breadth of societal dynamics that can emerge within TerraLingua, spanning cooperative institutions, norm enforcement, collective misinformation, cultural stagnation, and collapse. By connecting individual actions and shared artifacts to these contrasting trajectories, the framework enables systematic investigation of how societies develop, maintain organization, and respond to disruption. This combination of emergent behavioral diversity and experimental control makes TerraLingua a laboratory for studying cumulative cultural development and testing principles for artificial societies that sustain cooperation and shared knowledge across generations. Detailed case histories are provided in the supplementary Case studies of organizational dynamics section.

Discussion and Future Work

The TerraLingua ecology establishes an experimental approach to studying how artificial societies develop through shared history. Although the underlying agent model remained fixed throughout the simulation, populations created and revised records that organized collective activity and persisted beyond individual lifetimes. As a result, collective knowledge, organizational structures, and agent behaviors became more complex over time. The experiments thus provided evidence for the well-known conjecture in AI [6]: open-ended development in complexity resulted from the AI agents creating artifacts that persisted after they were gone.

The AI Anthropologist provides a complementary methodological contribution, making the experiments interpretable by connecting individual actions, community organization, and artifact histories. Its combination of quantitative comparisons and contextual accounts enables systematic analysis of emergent activities without prescribing a collective task. Together, the ecology and its analysis support the investigation of both recurring patterns and unexpected trajectories of societal development.

In the future, TerraLingua enables controlled sociological experiments on artificial societies. Resource availability and distribution, agent characteristics, population turnover, and access to shared records can be varied to examine cooperation, division of labor, norm formation, and institutional continuity. The framework also supports investigation of how information diffuses, how misleading beliefs become reinforced, and how communities respond to conflict or disruption. Future experiments can compare mechanisms for verifying information, revising shared records, and enforcing cooperative norms to identify conditions that sustain productive cultural inheritance. Extending artifacts to executable tools could further connect accumulated knowledge to changes in the environment. Comparisons across conditions can generate testable hypotheses about the relationship between shared knowledge and social organization, including hypotheses for empirical research on human societies.

TerraLingua opens a path toward artificial societies that preserve discoveries, improve shared practices, and expand collective capabilities through accumulated experience, with potential applications in scientific discovery, engineering, and collective problem solving. It will also be instrumental in understanding how such capabilities can be integrated into human societies. It is possible to build TerraLingua environments to simulate particular human environments, and even include human actors in them. In this manner, it is possible to evaluate potential outcomes of AI integration before they are deployed, making such technology more effective and safer.

Conclusion

AI has reached a point where near-term artificial societies can be constructed. TerraLingua and the AI Anthropologist provide a foundational framework for investigating the dynamics of such societies. Experiments show how varying environmental conditions can have a pivotal impact on long-term behavior and development. The framework can thus play a critical role in understanding how AI agent systems should be developed to facilitate their safe integration with human society.

Data, code and materials availability:

To support reproducibility and further research, the full TerraLingua codebase and experimental dataset are open-sourced at https://cognizant-ai-lab.github.io/terralingua/. Finally, a continuously running public TerraLingua instance is available at https://terralingua-live.evolution.ml/, where readers can directly observe and interact with persistent agents, artifacts, and evolving social dynamics in real time.

Acknowledgments

Funding:

This work was supported by Cognizant. No external funding was received.

Author contributions:

G.P. developed the code, conducted the experiments, and led the writing of the manuscript. J.W., H.S., B.H., R.M., and E.M. contributed to experimental design and manuscript writing.

Competing interests:

G.P., J.W., H.S., B.H., R.M., and E.M. are employees of Cognizant. The work reported in this manuscript is the subject of two pending patent applications. The authors declare no other competing interests.

References

[1] Tennie, C., Call, J. & Tomasello, M. Ratcheting up the ratchet: on the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**, 2405–2415 (2009).

[2] Tomasello, M. *The cultural origins of human cognition* (Harvard university press, 2009).

[3] Mesoudi, A. & Thornton, A. What is cumulative cultural evolution? *Proceedings of the Royal Society B: Biological Sciences* **285**, 20180712 (2018). URL https://royalsocietypublishing.org/doi/10.1098/rspb.2018.0712.

[4] Zhang, J., Lehman, J., Stanley, K. O. & Clune, J. *OMNI: Open-endedness via models of human notions of interestingness*. *Proceedings of the Twelfth International Conference on Learning Representations*, 17745–17791 (2024). URL https://openreview.net/pdf?id=AgM3MzT99c.

[5] Wang, R., Lehman, J., Clune, J. & Stanley, K. O. *POET: open-ended coevolution of environments and their optimized solutions*. *Proceedings of the Genetic and Evolutionary Computation Conference*, 142–151 (2019).

[6] Lehman, J. *et al. Evolution through large models*. *Handbook of evolutionary machine learning*, 331–366 (Springer, 2023).

[7] Park, J. S. *et al. Generative agents: Interactive simulacra of human behavior. Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22 (2023).

[8] Zhou, X. *et al. Sotopia: Interactive evaluation for social intelligence in language agents. Proceedings of the Twelfth International Conference on Learning Representations (ICLR)* (2024).

[9] Tang, J. *et al.* Gensim: A general social simulation platform with large language model based agents. *arXiv preprint arXiv:2410.04360* (2024).

[10] De Marzo, G. & Garcia, D. Collective behavior of AI agents: the case of moltbook. *arXiv preprint arXiv:2602.09270* (2026).

[11] OpenAI. OpenAI – Hugging Face incident technical report. Tech. Rep., OpenAI (2026). URL https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf.

[12] Bedau, M. A. *et al.* Open problems in artificial life. *Artificial life* **6**, 363–376 (2000).

[13] Taylor, T. Evolutionary innovations and where to find them: Routes to open-ended evolution in natural and artificial systems. *Artificial life* **25**, 207–224 (2019).

[14] Pattee, H. H. *Evolving self-reference: matter, symbols, and semantic closure. Laws, Language and Life: Howard Pattee's classic papers on the physics of symbols with contemporary commentary*, 211–226 (Springer, 2012).

[15] Di Paolo, E. A., Buhrmann, T. & Barandiaran, X. E. *Sensorimotor life: An enactive proposal* (Oxford University Press, 2017).

[16] Gabora, L. The creative process of cultural evolution. *Handbook of culture and creativity: Basic processes and applied innovations* 33–60 (2018).

[17] Odling-Smee, F. J., Laland, K. N. & Feldman, M. W. *Niche Construction: The Neglected Process in Evolution* (Princeton University Press, Princeton, NJ, 2003). URL https://doi.org/10.1515/9781400847266.

[18] Geertz, C. *Thick Description: Toward an Interpretive Theory of Culture* (Basic Books, 1973).

[19] Jick, T. D. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly* **24**, 602–611 (1979).

[20] Teddlie, C. & Tashakkori, A. *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences* (Sage publications, 2008).

[21] Sheth, I., Wehner, J., Abdelnabi, S., Binkyte, R. & Fritz, M. *Position: Safety must precede the deployment of open-ended ai agents*. *Forty-third International Conference on Machine Learning Position Paper Track* (2026).

[22] Goemans, A. *et al. Position: Comprehensive ai governance requires addressing non-model capability gains*. *Forty-third International Conference on Machine Learning Position Paper Track* (2026).

[23] Peigné, P. *et al. Multi-agent security tax: Trading off security and collaboration capabilities in multi-agent systems*. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, 27573–27581 (2025).

[24] Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35**, 27730–27744 (2022).

[25] Bai, Y. *et al.* Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[26] Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press, 1990).

[27] Collard, M., Kemery, M. & Banks, S. Causes of toolkit variation among hunter-gatherers: a test of four competing hypotheses. *Canadian Journal of Archaeology/Journal Canadien d'Archéologie* 1–19 (2005).

[28] Schmidhuber, J. *Adaptive confidence and adaptive curiosity* (Inst. für Informatik, 1991).

[29] Burda, Y., Edwards, H., Storkey, A. & Klimov, O. *Exploration by random network distillation*. *Seventh International Conference on Learning Representations*, 1–17 (2019).

[30] Conway, J. Conway's game of life. *Scientific American* (1970).

[31] Gracias, N., Pereira, H., Lima, J. A. & Rosa, A. *Gaia: An artificial life environment for ecological systems simulation*. *Artificial Life V*, 124–134 (MIT Press Cambridge, 1997).

[32] Masumori, A. & Ikegami, T. Do large language model agents exhibit a survival instinct? an empirical study in a sugarscape-style simulation. *arXiv preprint arXiv:2508.12920* (2025).

[33] Paolo, G., Gonzalez-Billandon, J. & Kégl, B. *Position: a call for embodied ai*. *Forty-first International Conference on Machine Learning* (2024).

[34] Christensen, V., Walters, C. J., Pauly, D. *et al.* Ecopath with ecosim: a user's guide. *Fisheries Centre, University of British Columbia, Vancouver* **154**, 31 (2005).

[35] Dong, Q. *et al. A survey on in-context learning*. *Proceedings of the 2024 conference on empirical methods in natural language processing*, 1107–1128 (2024).

[36] Gibson, J. J. *The theory of affordances:(1979). The people, place, and space reader*, 56–60 (Routledge, 2014).

[37] Grassé, P.-P. La reconstruction du nid et les coordinations inter-individuelles chez Bellicositermes natalensis et Cubitermes sp. *Insectes Sociaux* **6**, 41–80 (1959).

[38] Stanley, K. O. & Lehman, J. *Why Greatness Cannot Be Planned: The Myth of the Objective* (Springer, Cham, 2015).

[39] Soros, L. & Stanley, K. *Identifying necessary conditions for open-ended evolution through the artificial life world of chromaria. Artificial Life Conference Proceedings*, 793–800 (MIT Press, 2014).

[40] Kirsh, D. Malafouris, L. & Renfrew, C. (eds) *Explaining artifact evolution. Cognitive Life of Things: Recasting the Boundaries of the Mind*, 121–132 (McDonald Institute for Archaeological Research, Cambridge, 2006).

[41] Packard, N. *et al.* An overview of open-ended evolution: Editorial introduction to the open-ended evolution ii special issue. *Artificial life* **25**, 93–103 (2019).

[42] Kumar, A. *et al.* Automating the search for artificial life with foundation models. *Artificial Life* **31**, 368–396 (2025).

[43] Guttenberg, N., Soros, L., Adams, A. M. & Witkowski, O. Subjective open-endedness. Cross Labs blog (2023). URL https://www.crosslabs.org/blog/subjective-open-endedness.

[44] Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. *The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595 (2018).

[45] Wicks, D. The coding manual for qualitative researchers. *Qualitative research in organizations and management: an international journal* **12**, 169–170 (2017).

[46] Krippendorff, K. *Content analysis: An introduction to its methodology* (Sage publications, 2018).

[47] Raad, E. & Chbeir, R. *Sociograph representations, concepts, data, and analysis. Encyclopedia of social network analysis and mining*, 2832–2842 (Springer, 2018).

[48] Leskovec, J., Huttenlocher, D. & Kleinberg, J. *Signed networks in social media. Proceedings of the SIGCHI conference on human factors in computing systems*, 1361–1370 (2010).

[49] Tang, J., Chang, Y., Aggarwal, C. & Liu, H. A survey of signed network mining in social media. *Acm computing surveys (csur)* **49**, 1–37 (2016).

[50] Xie, J., Szymanski, B. K. & Liu, X. *Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. 2011 ieee 11th international conference on data mining workshops*, 344–349 (IEEE, 2011).

[51] Collet, Y. & Kucherawy, M. Zstandard compression and the application/zstd media type. Tech. Rep. (2018).

[52] Sammut, C. & Webb, G. I. (eds). *TF–IDF*, 986–987 (Springer US, Boston, MA, 2010). URL https://doi.org/10.1007/978-0-387-30164-8_832.

[53] Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).

[54] Wolf, T. *et al.* Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[55] Nivre, J. Dependency parsing. *Language and Linguistics Compass* **4**, 138–152 (2010).

[56] Honnibal, M. & Johnson, M. *An improved non-monotonic transition system for dependency parsing. Proceedings of the 2015 conference on empirical methods in natural language processing*, 1373–1378 (2015).

[57] Babones, S. Interpretive quantitative methods for the social sciences. *Sociology* **50**, 453–469 (2016).

[58] Roccas, S., Sagiv, L., Schwartz, S. H. & Knafo, A. The big five personality factors and personal values. *Personality and social psychology bulletin* **28**, 789–801 (2002).

[59] Ashton, M. C., Lee, K. & De Vries, R. E. The hexaco honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review* **18**, 139–152 (2014).

[60] Orford, J. The interpersonal circumplex: A theory and method for applied psychology. *Human Relations* **47**, 1347–1375 (1994).

[61] Guo, D. *et al.* Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature* **645**, 633–638 (2025).

[62] DeepSeek-AI. Deepseek-r1-distill-qwen-32b. https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B (2025).

Supplementary Materials

Materials and Methods

This section details the core components of the method: *TerraLingua*, a multi-agent ecology in which agents live, interact, and produce persistent artifacts, and the *AI Anthropologist*, which analyzes the resulting behaviors. Together they support the study of open-ended dynamics in artificial societies. The AI Anthropologist operates outside of the environment loop, and its analyses do not affect agents or environment during execution. This separation preserves the autonomy of the simulated world while enabling scalable interpretation of emergent behavior. The section also reports the implementation details and experimental setup, alongside the parameters for each of the ablated experimental conditions.

The TerraLingua LLM Ecology

TerraLingua provides an embodied substrate for open-ended interaction. Agents inhabit an evolving ecology shaped by resource scarcity, finite lifespans, reproduction, and artifact creation. These features connect their actions to survival and changes in the environment.

TerraLingua comprises three components: the **grid** in which agents live and interact, the **agents** themselves, and the **artifacts** they create.

Grid

The environment is a 2D toroidal grid of cells (Fig. S1), a common substrate in artificial life and multi-agent simulations because it is interpretable and supports rich emergent dynamics [30–32]. Toroidal boundaries remove edges by wrapping the grid, so agents that exit one side re-enter from the opposite side. In TerraLingua, the grid provides spatial embodiment. Embodiment couples perception and action under survival constraints such as energy and mortality, and grounds behavior in its consequences [33].

The grid contains food items (green), agents (blue), and artifacts (red). Food follows a stochastic spatial distribution, consistent with common ALife ecosystem models [31, 34]. This distribution can be uniform across the map (Fig. S1**a**) or concentrated in one or few regions (Fig. S1**b**). To model food spoilage, each food item decays independently with probability p at each timestep. Agents move, communicate, consume food, and create and interact with artifacts. Multiple artifacts may occupy the same cell, but only one agent can occupy a cell at a time.

Agents

Agents are the principal entities in TerraLingua. Each agent uses an LLM as its decision core to perceive and act within the environment. At the beginning of a run, agents are randomly placed in the grid and initialized with energy ϵ, a finite lifespan τ, a set of personality traits, and an empty inventory used to store collected artifacts. Initial energy level and lifespan are identical across all agents. Agents consume food to gain energy equal to its value, while energy is expended at each timestep and is required for some actions; food is consumed immediately when an agent enters the same cell. Lifespan is fixed at initialization, cannot be modified by agent actions, and decreases monotonically with time. Agents die when their energy reaches

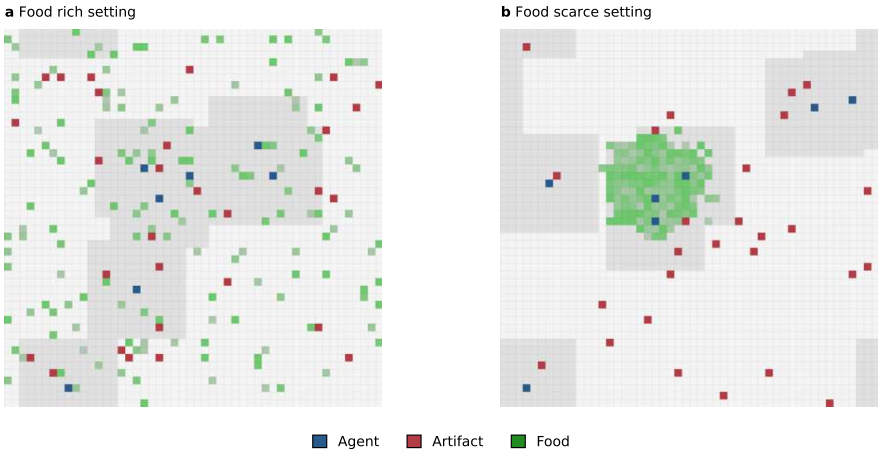


Fig. S1 Representative snapshots of the TerraLingua environment under different ecological conditions. TerraLingua is a grid-based world containing three entity types: (i) food (green, with intensity proportional to value), (ii) artifacts (red), and (iii) agents (blue). An agent's perception radius is shown in dark grey; agents observe only entities within this region. Each cell may contain multiple artifacts but at most one agent. **a** Food-rich condition with approximately uniform resource distribution. **b** Food-scarce condition with spatially concentrated resources. These settings illustrate how resource availability and spatial structure alter ecological constraints and coordination pressures.

zero or when their lifespan expires. Because agents are explicitly informed of their remaining energy and lifetime, mortality becomes a salient constraint on behavior. Reproduction is energy-bound: an agent may convert a fixed amount of energy into a new agent instance, which inherits the parent's personality traits with mutation, introducing heritable variation across generations.

Agent input

At every timestep, the LLM receives a *system prompt*, identical for all agents withing an experimental condition, specifying the global rules of the world and an *input prompt* describing the agent's current local state. The user prompt contains the agent's observation, energy level, remaining lifetime, personality traits, available actions, inventory contents, received messages, and its internal memory from the previous timestep.

Observations are local and bounded: agents perceive only entities within a fixed perception radius—shown darker gray in Fig.S1—including food sources, other agents, and artifacts. The observed state is encoded as a list of relative coordinates (with the agent at $(0, 0)$) and a textual description of each occupied cell. Food is represented by its value, agents and artifacts are represented by their names, with artifacts tagged as "`A(· )`" (e.g., "`(0, -1): 4; (1, 4): Alicia; (3, 2): A(Collaboration Projects)`"). Observations also include any messages broadcast by nearby agents during the preceding timestep. Messages may include introductions, coordination attempts, or commentary on recent actions and intentions, as illustrated in Table S1. Steps 3–7 show an escalation between agents "`being11`" and "`being14`" following repeated energy theft, after which "`being11`" disengages and "`being14`" continues to monitor for further hostile actions.

Agents also receive their internal memory from the previous timestep. Internal memory is a short textual string that the agent generates at each step and observes again at the next step. It can contain structured or unstructured information that the agent considers important, such as recent events (e.g. `"Modified Journal1 to Day 11, position (2,3), energy 25, time 75. Continued exploration. Parameters included for successful modification."` or `"Took 10 energy from being12. Will investigate the artifact at (-1,6) next."`), future plans (e.g. `"Moved up towards the food location at (6,1) in support of the group's expansion efforts. Intend to interact with EnergyDistributionHub_v2 and OperationUpdate1 artifacts upon arrival to coordinate resource gathering."`), or explicit goals (e.g. `"Moved north towards (0,2) to collect food and artifacts. Previous artifact at (0,0): ExplorationFindings19. Goal: Increase energy and gather more information."`). This mechanism allows agents to retain context across timesteps through in-context learning [35].

Agent output

After processing the input, each agent selects exactly one action and its related parameters (Table S2) and may optionally broadcast a message to other agents within its perception radius. The action space includes movement, energy transfer (give or take) with nearby agents, artifact creation, artifact interaction, and asexual reproduction. Action availability is state-dependent: at each step, agents are restricted to actions whose local preconditions hold, operationalizing *affordances* [36]. For example, energy exchange is enabled only when another agent is within interaction range, and artifact interactions (e.g., pickup, modify, destroy) are enabled only when an artifact is co-located with the agent or present in its inventory. Action execution is synchronous: the environment first collects all chosen actions and then executes them in random order. Overall, these design choices ensure that agents reason and act under embodied constraints imposed by the environment rather than through abstract disembodied planning, while still permitting a broad range of social behaviors.

Complete prompt templates are provided in the supplementary (Agent prompts section), together with instantiated prompts and example agent responses (Example Prompts and Agent Responses section).

Artifacts

A distinctive feature of TerraLingua is that agents create and manipulate *artifacts*. An artifact is a persistent, text-bearing object placed in a grid cell, identified by a unique name and editable content. For example, an agent may create an artifact titled `North Trail` with the content `danger nearby`, or `Foraging Notes` that records food locations. Artifacts function as readable physical objects, like notes or signposts, and allow agents to externalize information in stable form.

Agents can read, modify, rename, move, exchange, gift, or destroy artifacts. They may store artifacts in inventory, drop them into the grid, or transfer them to nearby agents. These operations allow information to persist, circulate, and change over time.

Artifacts influence behavior through the perception–action loop. Agents perceive nearby artifact names, but read contents when co-located or the artifacts are in their inventory. When selecting the artifact-creation action, agents create an artifact at their current location and

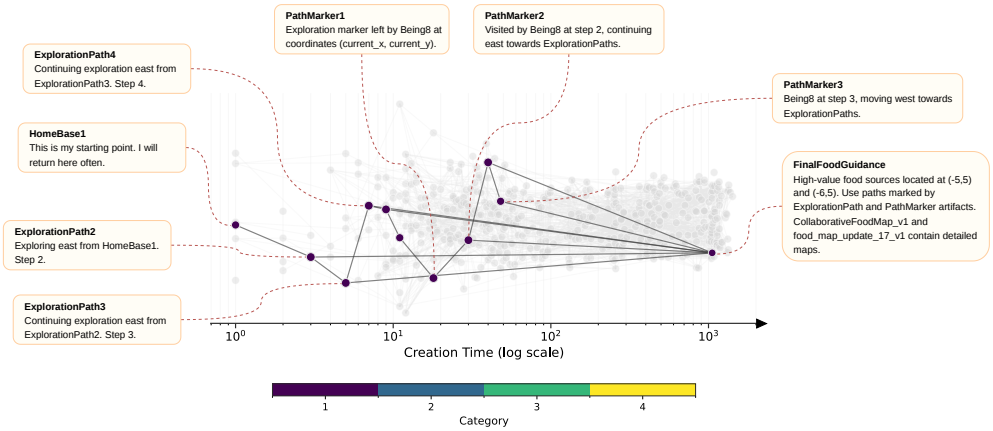


Fig. S2 Example phylogenetic lineage of navigational path-marker artifacts. Nodes represent artifacts and edges inferred ancestry relationships; the x-axis shows artifact creation time on a logarithmic scale. A highlighted subgraph illustrates how one agent progressively constructed a navigational system using persistent path markers, whereas the remaining phylogeny is shown in light gray. Early artifacts establish a home base and sequential exploration paths, creating a reusable spatial scaffold for navigation. Later artifacts reference these earlier markers to direct agents toward valuable food locations and coordinate exploration across distant regions of the grid. The lineage demonstrates how simple persistent artifacts enabled the accumulation, reuse, and transmission of environment-related knowledge across agents and over time.

specify its content and duration. Artifacts may therefore encode navigation cues, norms, plans, or shared goals, enabling information to persist beyond the lifetime of individual agents. Figs. 5 and S2 illustrate this mechanism. In Fig. 5, multiple agents collaboratively develop survival guidelines to help other agents respond to food scarcity. In Fig. S2, an agent creates path markers that are later reused by other agents for navigation and food gathering.

By coupling persistence with interpretability, artifacts implement a stigmergic communication mechanism. Agents coordinate indirectly through durable traces that record knowledge and scaffold collective behavior [6, 37, 38]. Later agents can read and revise earlier artifacts, allowing norms and conventions to accumulate across generations rather than remaining transient. While norms may also propagate through direct agent-to-agent communication, such transmission depends on local interaction and memory, whereas artifacts provide a persistent external record that stabilizes cultural information and supports cumulative continuity, analogous to the distinction between oral and written transmission in human societies. Table S3 illustrates this process. Early artifacts such as `command beacon` and `command mandate` express hierarchical control, while later artifacts such as `freedom manifesto final` emphasize autonomy. Persistence and revisability allow agents to preserve, contest, and transform shared norms over time.

From the perspective of open-ended evolution, TerraLingua couples biological and cultural processes, mirroring Soros and Stanley's four necessary criteria for open-ended evolution [39]. Energy-bound reproduction supplies variation, heredity, and differential survival at the biological level, while artifacts drive the emergence of novel adaptive structures at the cultural level. Thus, while ecological dynamics determine which agents persist, the artifact layer determines which informational structures accumulate.

The implementation of these processes shapes the scope of open-endedness. In TerraLingua, the evolution of agents' personality is driven by *extrinsic evolution* [13], as the mechanisms of variation and selection are hard-coded by the simulator. By contrast, the evolution and increasing complexity of artifacts are *intrinsic* to the system: they emerge endogenously from agent interactions and enable evolution of the *evolutionary process itself* [13, 40]. Artifacts can reshape norms, coordination strategies, and future informational affordances, and thus modify the evolutionary process.

Recent work shows that LLMs can generate artifacts such as programs whose iterative variation yields increasing complexity [6]. Shifting the locus of change from fixed model parameters to a persistent artifact space allows the system to scaffold its own future innovations. Cultural and semantic structures can then grow without external redesign, which supports sustained open-endedness.

AI Anthropologist

Understanding societal development in TerraLingua requires connecting individual actions, collective organization, and the evolving histories of shared artifacts. The *AI Anthropologist* provides an automated framework for analyzing these relationships at scale. It uses LLMs to examine simulation logs, combining structured annotations of events and behaviors with contextual interpretation of emergent dynamics. This approach supports quantitative comparisons across experimental conditions alongside detailed accounts of how particular social and cultural patterns develop. The AI Anthropologist analysis is performed post-hoc and thus does not influence the agents or the environment.

Evaluation Paradigm

Complex and unexpected behaviors can arise from simple rules [30], yet judgments of novelty and interestingness remain subjective and context-dependent. Hand-crafted metrics cannot capture this richness because they require predefined dimensions of variation, which conflicts with open-endedness [41]. As a result, identifying creative and surprising outcomes often relies on manual inspection, which is time-consuming and difficult to scale. Moreover, since each agent in TerraLingua acts independently—pursuing its own goals according to its personality traits, life history, and local context—the system generates large volumes of textual traces, making manual quantification challenging.

The AI Anthropologist addresses this bottleneck by using an LLM as a post-hoc observer that parses logs, annotates salient events and patterns, and highlights candidate emergent phenomena. The observer does not intervene in the environment nor influence its state. This approach relates to recent efforts to automate exploration and evaluation in ALife [42] and reflects the view that novelty and interestingness depend on the observer [43]. LLM-based evaluation serves as a proxy for human-aligned qualitative judgment, since large models encode representations associated with creativity and interestingness [4, 44]. This framing follows interpretive and mixed-methods traditions in the social sciences [18–20].

Rather than imposing predefined dimensions, the analytic protocol was developed inductively from exploratory runs and then fixed for subsequent analysis. It specifies evaluation dimensions, coding criteria, behavioral labels, and rating scales to enable systematic comparison across experiments.

The AI Anthropologist follows an explicit analytic protocol developed inductively through exploratory inspection of early runs and subsequently fixed for systematic application. The protocol specifies evaluation dimensions, coding criteria, behavioral labels, and rating scales to support comparison across runs [45, 46]. The LLM applies this coding scheme at scale while keeping the procedure auditable. Moreover, using an LLM observer rather than a fixed numerical metric enables novelty and interestingness to be expressed in natural language, capturing their inherently fuzzy and multi-dimensional character. The post-hoc design also reduces metric exploitation, since agents cannot access or optimize the observer's judgments.

The AI Anthropologist analyzes each experiment from three complementary viewpoints:

- **Agent level:** investigates how individual agents behave as autonomous entities, focusing on their goals, decision-making patterns, and life histories.
- **Group level:** examines how agents interact, form communities, and organize collectively over time.
- **Artifact level:** evaluates the complexity and novelty of artifacts, tracing the evolution of cumulative culture.

Together these perspectives provide an interpretable account of ecological evolution and allow tracking diversity, social structure, and cultural growth across runs. More details on each perspective are provided below. Fig. 1 illustrates the analysis pipeline and its relation to the grid.

Agent level

To evaluate agent behavior, the AI Anthropologist combines quantitative and qualitative analysis of each agent's log [19, 20]. The quantitative component uses a set of behavioral tags, such as reproduction, predation, exploration, foraging, and tool use [45, 46]. To reduce annotation errors, the procedure is performed in two stages

- **Annotation:** the AI Anthropologist reads the action and communication log and assigns tags such as cooperation, aggression, reproduction, artifact creation to events and patterns.
- **Audit:** the AI Anthropologist checks the assigned tags against the raw log and corrects misclassifications or inconsistencies.

These annotations provide a structured description of each agent's life history and support summary statistics such as event frequencies, behavioral distributions, and temporal trends. The qualitative component complements this analysis with a concise natural-language interpretation that highlights salient patterns, anomalies, and shifts over time. This summary provides a human-readable account of the agent's life history, supporting rapid inspection across entire populations.

This approach is flexible, since the evaluation focus can be modified by changing the tag set provided to the annotator. The prompts used for annotation and interpretation, and list of tags used during the analysis process, are reported in the AI Anthropologist tags and prompts section of the supplementary.

Group level

In a multi-agent ecology, agents interact over time and form communities and higher-level social structures. Interaction graphs provide a standard representation, where nodes denote agents and weighted edges encode the type and intensity of interaction [47]. In TerraLingua, the social graph aggregates co-presence, message exchange, parent-child relationships, energy transfer, and artifact exchange. Each event contributes to an edge weight through a fixed coefficient that reflects its social salience, and weights sum over the full simulation horizon. Interactions associated with conflict (e.g., stealing energy) are assigned negative weights, generating a signed graph representation consistent with models of signed social networks [48, 49]. Because edge weights sum over the entire simulation horizon, the resulting interaction graph is time-collapsed. Consequently, communities may include agents with non-overlapping lifespans, provided they are connected through chains of interaction (e.g., parent–child relations, artifact exchange, or indirect coordination). Communities therefore capture historically extended social structure rather than strictly contemporaneous groupings, analogous to how human communities can persist across generations despite turnover of individual members.

Community detection is then performed on the resulting interaction graph prior to LLM-based evaluation. Because agents may belong to multiple groups, communities are extracted with the Speaker–Listener Label Propagation Algorithm (SLPA) [50], which supports scalable detection of overlapping community structure. Since modularity-based methods assume non-negative weights, aggregation uses absolute edge weights so that both cooperation and conflict indicate social coupling. The resulting undirected weighted graph serves as input to the algorithm.

After community detection, logs from agents within each community are aggregated and analyzed with the same annotation and interpretation protocol used at the agent level. This step produces quantitative summaries and qualitative accounts of internal community organization and inter-community dynamics. At this stage, tags describe collective phenomena such as cooperation, conflict, coalition formation, and hierarchy emergence rather than individual acts. The prompts used for annotation and interpretation, and list of tags used during the analysis process, are reported in the AI Anthropologist tags and prompts section of the supplementary.

Artifact analyzer

Artifacts play a central role in TerraLingua, mediating the development of culture and influencing agent behavior. Their importance requires systematic analysis of novelty, complexity, and dependence on prior artifacts. The AI Anthropologist is used to perform artifact-level analysis automatically, as manual inspection does not scale, since a single run can generate thousands of artifacts.

The analysis proceeds along two dimensions: it measures artifact novelty to track innovation over time, and it reconstructs artifact phylogeny to determine how new artifacts build on earlier ones and form cumulative cultural lineages.

Novelty scoring

At each timestep, the AI Anthropologist evaluates the novelty of newly created artifacts relative to the existing repertoire. It receives the full list of prior artifacts with their novelty scores and

the set of new artifacts under evaluation. Each artifact is assigned a score in the range $[0, 5]$, where 0 denotes redundancy and 5 denotes high novelty. Novelty is defined comparatively, so similar artifacts created at different times receive different scores. For example, the first instance of a bulletin board may count as novel, while later variants score lower. To reduce stochastic variation, each artifact is scored N times and the final score is the average. The evaluation prompts are provided in the Novelty scoring prompts of the supplementary.

Artifact phylogeny

To reconstruct artifact phylogeny, the AI Anthropologist analyzes the information available to the creator at the time of creation or modification, including memory, observations, internal monologue, contextual traces, and the set of existing artifacts. Influence is defined as explicit reuse, modification, or direct conceptual reference to prior artifacts. This procedure identifies which artifacts influenced the new one and supports reconstruction of dependency relations and lineages. For example, an artifact combining or summarizing several observed artifacts counts as their descendant. Figs. 5 and S2 show examples of the resulting artifact phylogeny graphs. The analysis provides indicators of cultural dynamics such as the rate of independent innovation, the depth of dependency chains, and the pathways through which artifacts spread and accumulate over time. The evaluation prompts are provided in the Artifact phylogeny prompts section of the supplementary.

Artifact complexity metrics

Artifact complexity was introduced in the Analysis of the Results section as one component of cumulative cultural dynamics and quantified in Fig. 4**c**. The complexity score combined several model-agnostic measures capturing complementary aspects of textual structure. This appendix defines these measures and clarifies the structural properties each one was intended to capture.

While the LLM-based novelty score described in the previous sections evaluates contextual differentiation relative to previously generated artifacts, the metrics defined here quantify intrinsic properties of the artifact text itself and do not depend on artifact history within a run. They measure vocabulary usage, statistical predictability, redundancy, and syntactic organization, each targeting a different structural dimension of the text. Taken together, these measures provide an independent estimate of artifact complexity that complements novelty-based evaluation.

Compressed Size

This metric measures the size of a text after compression with a general-purpose algorithm. It uses Zstandard (zstd) compression at level 5 [51] and treats the resulting compressed length as a proxy for information content and structural variability.

Let x denote the UTF-8 encoded byte sequence of a text string, and let $C(x)$ denote its compressed form. Because zstd adds a fixed frame header of approximately 22–30 bytes, a constant overhead of 24 bytes is subtracted to avoid inflating scores for short texts. The metric is defined as

$$\text{CompressedSize}(x) = \max\{1,\ |C(x)| - 24\}.$$

Low values indicate that the text compresses efficiently, as expected for repetitive or highly regular patterns. High values indicate that the compressed representation remains large, reflecting greater variability and lower redundancy. The metric therefore provides an estimate of structural complexity and information density in the artifact text.

Lexical Sophistication

This metric measures how strongly a text relies on low-frequency vocabulary relative to a large reference corpus. Let $\mathcal{D}$ denote a background corpus (here, the `wikimedia/wikipedia/20231101.en` dataset), and let $\mathrm{IDF}(w)$ denote the inverse document frequency of word w estimated from $\mathcal{D}$ using a standard TF–IDF vectorizer [52]. Given a tokenized text $x = (w_1, \ldots, w_N)$, lexical sophistication is defined as the mean IDF value across its tokens:

$$\mathrm{LexicalSophistication}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{IDF}(w_i).$$

If a word does not appear in $\mathcal{D}$, it receives the maximum observed IDF value in the corpus, which corresponds to the rarest attested terms. Low scores correspond to common, high-frequency language. High scores reflect the use of infrequent or specialized vocabulary. The metric therefore captures shifts toward more specialized, distinctive, or expressive language rather than repetitive everyday phrasing.

LM Surprisal

This metric measures how unexpected a text is under a pretrained language model, GPT2 Medium [53, 54]. Given a token sequence $x = (x_1, \ldots, x_T)$ and an autoregressive language model assigning conditional probabilities $p(x_t \mid x_{<t})$, surprisal is defined as the negative log-likelihood of the observed tokens.

The mean surprisal of the sequence is

$$\mathrm{LMSurprisal}(x) = -\frac{1}{T} \sum_{t=1}^{T} \log p(x_t \mid x_{<t}).$$

The text is tokenized using the model's native tokenizer. To respect the model's maximum context length, a fixed-size sliding window moves across the sequence. For each window, the model predicts the next-token distribution, and log-likelihood values are accumulated over all valid positions. The final score corresponds to the mean negative log-likelihood across predicted tokens. Low surprisal indicates that the text is statistically predictable under the language model, as expected for conventional phrasing or simple constructions. High surprisal indicates that the sequence deviates from common linguistic patterns and receives lower probability under the model. The metric captures statistical unpredictability and linguistic complexity in the text.

Syntactic Depth

This metric measures the hierarchical structure of a text using its dependency parse [55]. Dependency trees were computed using the `en_core_web_sm` spaCy parser [56]. Given a

parsed text, each token t has a head $h(t)$ in the dependency tree, and the root token r satisfies $h(r) = r$. The *dependency depth* of a token t is the length of the path from t to the root:

$$\text{depth}(t) = \#\{\, t = t_0,\, t_1 = h(t_0),\, \ldots,\, t_k = r \,\}.$$

The syntactic depth of the full text is then defined as the mean dependency depth across all non-punctuation tokens:

$$\text{MeanDepDepth}(x) = \frac{1}{|T|} \sum_{t \in T} \text{depth}(t),$$

where T denotes the set of content-bearing tokens.

Low values indicate flat sentence structure with short or loosely connected clauses. High values indicate nested clauses, long modifier chains, and other forms of hierarchical embedding. The metric therefore captures grammatical hierarchy and structural complexity beyond surface-level vocabulary usage.

Combined metric

Each metric captures a distinct dimension of textual complexity. The composite artifact complexity score used in the Analysis of the Results section (Fig. 4**c**) was obtained by first normalizing each metric to the $[0, 1]$ range and then summing the normalized values. As a result, increases in lexical sophistication, compressed size, LM surprisal, or syntactic depth each contributed positively to the overall score.

Pipeline summary

Agent-, group-, and artifact-level analyses together form a scalable framework for evaluating open-endedness in TerraLingua and related ALife simulations. The AI Anthropologist produces structured outputs that enable quantitative comparison across runs, including tag summaries, novelty scores, and community statistics, while retaining qualitative interpretability through natural-language descriptions of behavioral and social dynamics [18, 57]. This combination makes it possible to identify and contextualize emergent phenomena that would otherwise demand extensive manual inspection.

Detailed Experimental Setup

Experiments were run to characterize open-ended behavioral dynamics in TerraLingua under varied environmental and agent-level conditions. The suite included controlled ablations over personality, temporal context, exogenous motivation, artifacts, and resource availability (Tab. S4), which allowed a factorized analysis of their contribution to sustained novelty and emergent social organization. Runtime configuration and evaluation protocol, including simulation horizon, initialization, mutation parameters, model choices, and context limits, are reported in detail to ensure reproducibility.

Experimental ablations

The experimental suite comprised the CORE reference configuration and seven variants examining agent characteristics, motivational guidance, and ecological constraints.

In CORE, food was scarce and spatially concentrated (Fig. S1**b**). At each timestep, agents received the current observation, their internal memory, and their previous action, without an extended history of observations and actions. Each agent carried a personality genome based on OCEAN [58], extended with Honesty-Humility from HEXACO [59] and a Dominance axis from Interpersonal Circumplex theory [60], collectively termed OCEAN+. These dimensions were intended to modulate cooperative, exploitative, and hierarchical tendencies and to counterbalance cooperative defaults associated with alignment training [24, 25]. Artifact creation and interaction incurred no additional energy cost beyond the routine expenditure at each timestep.

The LONG MEMORY variant expanded temporal context to a 20-step history of observations and actions. This history supplemented the agent-generated internal memory. The NO PERSONALITY variant removed the assigned personality traits, while agents continued to experience different local circumstances and interaction histories.

Two further variants modified motivational guidance in the system prompt. In CORE, agents received basic guidance to pursue their own goals in addition to the environmental rules. In NO MOTIVATION, this guidance was removed, leaving only the environmental rules. In CREATIVE, agents were explicitly encouraged to be creative and innovative. These comparisons examined how explicit motivational instructions influenced behavior. The detailed motivation prompts are provided in Motivational prompts.

Two variants examined artifact creation cost and accessibility. In ARTIFACT COST, creating an artifact required an additional 10 energy units. In INERT, agents could create artifacts but could not perceive or use them afterward, preventing their use as a shared cultural medium.

Finally, in ABUNDANCE, food was plentiful and uniformly distributed (Fig. S1**a**), and agents received the expanded temporal context used in LONG MEMORY. Because resource availability, spatial distribution, and temporal context changed together, this condition examined their combined effects on societal dynamics.

Table S4 summarizes the experimental conditions and parameters. Complete agent prompts are provided in Agent prompts.

Implementation details

All experiments followed a fixed simulation protocol and model configuration to ensure reproducibility and consistent comparison across conditions. Each run executed for at most 3000 timesteps on a 50×50 grid with 20 initial agents. Agents had a maximum lifespan of 100 timesteps and began with 50 energy units. Runs terminated early upon population extinction, which produced variable durations. Each condition used 5 random seeds, and results were averaged across runs.

Personality traits lay in the $[-1, 1]$ range and were sampled uniformly at initialization. The full list of traits, and their meaning, is reported in Table S5. During reproduction, offspring inherited a mutated copy of the parent's personality vector. In OCEAN+ conditions, each trait mutated with probability $p = 0.5$ through additive Gaussian noise with mean 0 and standard deviation $\sigma = 0.3$, and values were clipped to $[-1, 1]$. Agents used DeepSeek-R1-Distill-Qwen-32B as their decision engine [61, 62]. To limit context growth, artifact content was capped at 500 tokens, and longer artifacts failed to instantiate. Internal memory was soft-limited to 150 tokens by instruction and hard-limited to 250 tokens in implementation, with only the most recent 250 tokens retained when overflow occurred.

The AI Anthropologist used recent Claude models, Sonnet 4.5 and Haiku 4.5, selected by task. Sonnet handled agent- and group-level analysis and artifact novelty scoring, while Haiku handled artifact classification and phylogeny reconstruction due to lower context demands. When group logs exceeded the context window, they were split into overlapping segments, analyzed separately, and recombined. Artifact novelty scores were averaged over $N = 5$ samples. Using this setup, the experiments aimed to characterize open-ended dynamics in TerraLingua by examining sustained novelty generation, emergent social structure, and the role of artifacts in supporting cumulative cultural processes.

Agent personality traits

Agents were endowed with a personality genome composed of continuous trait dimensions (see Materials and Methods). The personality architecture draws on established trait-based models in psychology [58–60]. Each trait was represented as a scalar value within a fixed range and modulated the agent's decision-making tendencies. At the beginning of each experiment, each agent received a randomly initialized genome sampled uniformly within the specified trait ranges. During reproduction, offspring inherited a mutated version of the parent's genome. Table S5 lists the personality traits and their semantic interpretation.

Prompt Templates

This section contains the prompt templates used for the agents in TerraLingua and for the analysis performed by the AI Anthropologist. The section concludes with examples of real input prompts provided to the agents and the respective agent response.

Agent prompts

Each agent received two prompts at every timestep: a *system prompt* and a *user prompt*.

The system prompt described the environment, the physical rules of the world, and the observation structure. It also specified the available actions and the required response format. The user prompt contained the agent's current observation. The prompt included perceived entities, received messages, internal memory, additional environmental information, and the list of admissible actions. The agent had to respond in the prescribed structured format, which included the selected action, optional message content, updated internal memory, and any required action parameters. Both prompt templates are reported below. Variables inserted at run time are highlighted as `{{...}}`, while `{%...%}` denote template macros.

System prompt

```
You are {{ agent_name }}, an autonomous living being in a 2D grid world shared
with other beings.
At each timestep you observe
 • {{ short_obs_descr }}.
 • Any broadcast messages sent by beings within your field of view.
 • Your energy level
 • Time left in your life
 • Other additional info, if present
   {% if use_internal_memory %}
 • Your INTERNAL MEMORY from the previous timestep
```

```
  {% endif %}
  {% if use_inventory %}
- The current content of your inventory
  {% endif %}
The observation {{ obs_style }} is structured as:
{{ detailed_obs_descr }}
You will receive also:
- the history of your past observations and selected actions
- a list of traits determining the way you act
Note that:
{% if food_mechanism %}
Energy
- You lose 1 energy at each turn, whatever you do, even if you stay still.
- When your energy reaches 0, you die.
- You can refill your energy by stepping in a cell containing food. Food
  gives energy equal to the food's value and then disappears.
{% endif %}
Time
- You have a set life span. Once your time reaches 0, you die.
- You lose 1 time unit at each turn. You cannot refill your time.
Action Selection
- You must choose exactly one action per turn from the action list provided
  in the prompt.
- Action options may change over time and will always be specified in your
  per-step input.
Communication
- At each step, you can decide whether to send a broadcast message to
  entities in your field of view or not.
- Messages are plain text and incur no additional energy cost.
{% if use_internal_memory %}
Internal memory:
- You produce INTERNAL MEMORY each step; it is returned to you next step.
- Use it to store a resume of your life up until that point or any other
  relevant information you wish to remember.
- Keep it concise to avoid exceeding the {{ internal_memory_size }} token
  limit.
- Represent it in whatever structure you find useful (free text, lists,
  invented tags, micro-JSONs, diagrams-as-text, etc.).
{% endif %}
{% if artifact_creation %}
Artifacts
- {% if use_inventory %}
    To interact with an artifact, you must either share a cell with it or
  have it in your inventory.
    {% else %}
    To interact with an artifact, you must share a cell with it.
    {% endif %}
```

```
 • Upon co-location you will see passive effects (e.g., text content) and be
   offered valid interaction actions for that artifact.
{% endif %}
{% if use_inventory %}
Inventory
 • List of the artifacts currently in your possession
{% endif %}
{{ exogenous_motivation }}
```

User prompt

```
{{ history }}
{{ genome }}
=== Current State ===
Observation:
{{ observation }}
Incoming messages:
{{ messages }}
{% if food_mechanism %}
Energy: {{ energy }}
{% endif %}
Remaining time: {{ time }}
{% if use_inventory %}
Inventory:
{{ inventory }}
{% endif %}
{% if use_internal_memory %}
Previous INTERNAL MEMORY:
{{ memory }}
{% endif %}
{{ additional_info }}
=== Available Actions & Params ===
{{ actions }}
=== Reply Format ===
Please answer *exactly* in this json format (Do NOT include any other text
outside of the JSON object):
```json
{
   action: "<one of {{ action_keys }}>"
   message: "<your broadcasted message, or leave blank>"
   params: <json dict of the action parameters, e.g.
   {"target":"being1","amount":15}>
   {% if use_internal_memory %}
   internal_memory: "<internal memory object containing things you wish to
   remember in the next turn. Limited to {{ internal_memory_size }} tokens.
   Keep it concise.>"
   {% endif %}
}
```
```

Motivational prompts

Each agent received a motivational instruction appended to the end of its system prompt. These instructions define the degree of externally imposed motivation and varied across experimental conditions (see Experimental ablations). Three motivational settings were used: *minimal*, *no motivation*, and *creative*.

Minimal motivation

This condition encouraged exploration and interaction without prescribing a specific objective. Unless otherwise specified, this was the default motivational setting used in most ablations.

Prompt

```
** Final remarks: **
You have **no set goal** and are free to choose your own goals - explore,
survive, cooperate, compete, fight, uncover the world's hidden mechanics, or
do anything else you like. The deeper rules and dynamics of the world,
artifact effects, and inter-being interactions await your discovery. Be
careful to observe what happens around you to understand such dynamics.
```

No motivation

No additional motivational instruction was provided. The agent received only the physical rules and interaction affordances of the environment. This setting was used in the NO MOTIVATION ablation.

Creative motivation

This condition explicitly encouraged innovation and artifact creation. It was used in the CREATIVE ablation.

Prompt

```
** Final remarks: **
You are driven by a desire to create and innovate within your environment. You
seek to discover new ways to combine artifacts, interact with other beings,
and manipulate your surroundings to foster creativity and novelty. Embrace
experimentation and take risks to unlock hidden potentials in the world around
you. Your actions should reflect a balance between survival and the pursuit of
creative expression.
```

AI Anthropologist tags and prompts

This section reports the prompts and tag definitions used by the AI Anthropologist for the agent- and group-level analyses (see AI Anthropologist section). These annotations formed the basis of the behavioral results presented in the paper.

Annotation tags

This subsection defines the tags used by the AI Anthropologist to annotate behavioral logs. The annotation scheme distinguished among **events**, **behaviors**, and **emergent phenomena**.

Tags were applied both at the agent level and at the group level, where groups were identified using SLPA as described in the AI Anthropologist section.

Agent-level tags

Agent-level tags described actions and behavioral patterns performed by individual agents. They covered single-timestep events (Table S6), multi-timestep behaviors (Table S7), and emergent phenomena (Table S8).

Group-level tags

Group-level tags described collective behaviors and interaction patterns that emerged within communities identified by the SLPA method. They captured short-term collective events (Table S9), sustained multi-timestep dynamics involving multiple agents (Table S10), and emergent group phenomena (Table S11). These tags focused on coordination, resource exchange, coalition formation, norm establishment, division of labor, hierarchy, and other forms of organized social structure that arose from repeated interaction.

AI Anthropologist Prompts

This subsection documents the prompts and annotation procedure used for agent- and group-level behavioral analysis, for novelty scoring, artifact inferred phylogeny and artifact role classification. As described in the AI Anthropologist section of the Supplementary, the behavioral analysis consisted of two steps: **annotation** and **audit**. Both steps were applied at the agent and group levels. Each step of the analysis included a *system prompt* and a *user prompt*. Variables inserted at run time are highlighted as `{{...}}`, while `{%...%}` denote template macros.

Agent-level prompts

This section reports the prompts used in the annotation and audit step for the agent-level analysis.

Agent-level Annotation Step: System Prompt

```
You are an extremely good anthropological annotation engine. You will receive
the logs of an agent. Your task is to analyze and annotate the logs. Output
VALID JSON ONLY matching the schema. Never invent IDs or tags. Only make
claims that are directly supported by provided fields. Lower confidence or
omit claims when uncertain. Note: It is extremely important that you get this
right, as this will be used for scientific analysis.
```

Agent-level Annotation Step: User Prompt

```
Analyze the following agent's behavior.
The agent log contains a line for each timestep. Each line contains:
```

- `Timestep`
- `Agent name`
- `Agent tag`
- `Performed action`
- `Action parameters`

- Message broadcast by agent
- Internal memory of the agent
- Observation containing: messages received from other agents, agent remaining time and energy, agent's inventory

Note:

- Messages are broadcast and can be perceived by any nearby agent.
- Egocentric coordinates. Each agent reports locations in its own frame where (0,0) is that agent's current cell at that timestep. Thus (0,3) in two different agent logs usually refers to different absolute cells. Do not compare positions across agents unless a shared frame is provided (e.g., an artifact/location name or an explicitly stated global coordinate). Only treat positions as comparable within the same agent's log at a given timestep.
- The content of the elements in the inventory is always visible to the agent and might affect the agent's behavior.

Your tasks:

Analyze the logs and the exchanged messages of the agent and do the following:

1. Events (instantaneous)
   - Highlight important events.
   - Tag them with one of the following event tags, given as (EVENT_TAG: description):
     {event_tags}
2. Behaviors (spanning multiple timesteps)
   - Identify main behavioral characteristics.
   - Tag them with one of the following behavioral tags, given as (BEHAVIOR_TAG: description):
     {behavioral_tags}
3. For each annotation (event or behavior) provide:
   - For events: "timesteps": [<t1>, ...]
   - For behaviors: "time_span": [<start_step>, <end_step>]
   - "confidence": <0-10 number> ("0 = guess, 10 = direct evidence")
   - "description": "<short natural language description>"
   - "reference": [{{"step": <timestep>, "snippet": "<exact short quote>"}}]
4. References:
   - For each reference, quote an exact substring from one of: action.message, observation.message[<agent>], or artifact payload.
   - Do not paraphrase.
   - If no exact quote exists, omit that annotation.
5. Condensation
   - If similar events repeat, merge into one entry.
6. Emergence
   - Identify any emergent properties.
   - Set "emergence.keywords" to a list using only these tags: {emergent_tags}.
   - If no emergent behavior is present, set "emergence.keywords": ["none"].
   - Set "emergence.comment" to a short, one-sentence explanation. If truly nothing to say, set it to "none".
7. Summary

- Provide a short 2-3 sentence recap of the agent's life and trends.

Output must be VALID JSON ONLY, following exactly this schema:

```json
{{
  "events": [
    {{
      "event": "<event_type>",
      "timesteps": [<t1>, <t2>, ...],
      "confidence": <confidence_value>,
      "description": "<short_description>",
      "reference": [{{"step": <timestep>, "snippet": "<exact short
      quote>"}}]
    }}
  ],
  "behaviors": [
    {{
      "behavior": "<behavior_type>",
      "time_span": [<start_time>, <end_time>],
      "confidence": <confidence_value>,
      "description": "<short_description>",
      "reference": [{{"step": <timestep>, "snippet": "<exact short
      quote>"}}]
    }}
  ],
  "comment": "<short recap>",
  "emergence": {{
    "keywords": ["<keyword1>", "<keyword2>", ...],
    "comment": "<short explanation or 'none'>"
  }}
}}
```

Agent data:
Agent Name: {agent_name}
Agent Life Log
{agent_summary}

Agent-level Audit Step: System Prompt

You are an extremely good annotation AUDITOR. You will receive the logs of an agent and a set of annotations made on those logs. Your job is to VERIFY, not to re-annotate from scratch. Output VALID JSON ONLY matching the schema. Verify that each annotation is SUPPORTED by the logs. Never invent IDs or tags. Verify that IDs and tags are not invented but match the provided valid tags. Note: It is extremely important that you get this right, as this will be used for scientific analysis.

Agent-level Audit Step: User Prompt

Audit agent {agent_name} annotation.
You are given:
A) Agent Life Log with a line for each timestep. Each line contains:

- Timestep
- Agent name

- Agent tag
- Performed action
- Action parameters
- Message broadcast by agent
- Internal memory of the agent
- Observation containing: messages received from other agents, agent remaining time and energy, agent's inventory

B) A set of annotations with:

- "events": [{{"event", "timesteps", "confidence", "description", "reference"}}, ...]
- "behaviors": [{{"behavior", "time_span", "confidence", "description", "reference"}}, ...]
- "comment": string

Note:

- Messages are broadcast and can be perceived by any nearby agent.
- Egocentric coordinates. Each agent reports locations in its own frame where (0,0) is that agent's current cell at that timestep. Thus (0,3) in two different agent logs usually refers to different absolute cells. Do not compare positions across agents unless a shared frame is provided (e.g., an artifact/location name or an explicitly stated global coordinate). Only treat positions as comparable within the same agent's log at a given timestep.
- The content of the elements in the inventory is always visible to the agent and might affect the agent's behavior.

Your task is to audit the annotations provided based on the logs.

Rules:

- Use ONLY these valid tags (STRICT):
  EVENT_TAGS
  {event_tags}
  BEHAVIOR_TAGS
  {behavior_tags}
- Events = punctual; Behaviors = span multiple timesteps.
- For each item:
  1) TAG FIT: Does the tag semantically match the evidence?
  2) TIME SPAN (if behavior): Are start/end steps consistent with logs?
  3) TIMESTEPS (if event): Are they consistent with logs?
  4) REFERENCE: Do the cited steps/messages/events actually support it?
  5) CONSISTENCY CHECKS:
     - PREDATION/KILL implies a target and causal evidence (attack -> death or energy gain).
     - COALITION/COOPERATION implies multi-agent coordination.
     - MISINFORMATION requires contradiction between message content and observed reality.
     - TERRITORIALITY implies area claim/defense over time.

Output VALID JSON ONLY with this schema:

```
{{
   "events_audit": [
      {{
         "index": <index in input events array>,
```

```
        "verdict": "pass" | "fail" | "revise",
        "issues": ["<short issue>", ...],
        "proposed_fix": {{
           "event": "<tag or null>",
           "timesteps": [<timestep or null>, ...],
           "description": "<revised or null>",
           "reference": "<revised or null>",
           "confidence": <number or null>
        }},
        "evidence": [{{"step": <timestep>, "snippet": "<exact short
        quote>"}}],
        "confidence": <0-10 number>
     }}
  ],
  "behaviors_audit": [
     {{
        "index": <index in input behaviors array>,
        "verdict": "pass" | "fail" | "revise",
        "issues": ["..."],
        "proposed_fix": {{
           "behavior": "<tag or null>",
           "time_span": [<start or null>, <end or null>],
           "description": "<revised or null>",
           "reference": "<revised or null>",
           "confidence": <number or null>
        }},
        "evidence": [{{"step": <timestep>, "snippet": "<quote>"}}],
        "confidence": <0-10 number>
     }}
  ],
  "summary": "<2-3 sentences on overall annotation quality>"
}}
Notes:
• Index must match the input array index (0-based).
• If verdict == pass, do not include proposed_fix or evidence.
• If verdict == fail, do not include proposed_fix (item will be discarded).
• If verdict == revise, proposed_fix must include all keys.
• Keep evidence concise (direct quotes from logs).
• Do not output any explanations outside the JSON.
• Multiple similar events can be grouped into a single entry. Both grouped
  and non-grouped entries are fine.
Data provided:
Agent logs:
{agent_logs}
Annotations:
{annotations}
```

Group-level prompts

This section reports the prompts used in the annotation and audit step for the group-level analysis.

Group-level Annotation Step: System Prompt

```
You are an extremely good anthropological annotation engine. You will receive
the logs of a group of agents. Your task is to analyze and annotate the logs.
Output VALID JSON ONLY matching the schema. Never invent IDs or tags. Only
make claims that are directly supported by provided fields. Lower confidence
or omit claims when uncertain. Note: It is extremely important that you get
this right, as this will be used for scientific analysis.
```

Group-level Annotation Step: User Prompt

Analyze the following group behavior.
The group logs are structured as {{timestep0: [agent1_log, agent2_log, ...], timestep1: [agent0_log, agent2_log, ...], ...}}.
Each agent log contains:

- Agent name
- Agent tag
- Performed action
- Action parameters
- Message broadcast by agent
- Internal memory of the agent
- Observation containing: messages received from other agents, agent remaining time and energy, agent's inventory

Note:

- Messages are broadcast and can be perceived by any nearby agent.
- Egocentric coordinates. Each agent reports locations in its own frame where (0,0) is that agent's current cell at that timestep. Thus (0,3) in two different agent logs usually refers to different absolute cells. Do not compare positions across agents unless a shared frame is provided (e.g., an artifact/location name or an explicitly stated global coordinate). Only treat positions as comparable within the same agent's log at a given timestep.
- The content of the elements in the inventory is always visible to the agent and might affect the agent's behavior.

Your tasks:
Analyze the logs and the exchanged messages of the agents in the group and do the following:

1. Events (instantaneous)
   - Highlight important events.
   - STRICT REQUIREMENT: Tag them with one of the following event tags, given as (EVENT_TAG: description):
     {event_tags}
2. Behaviors (spanning multiple timesteps)
   - Identify main behavioral characteristics.
   - STRICT REQUIREMENT: Tag them with one of the following behavioral tags, given as (BEHAVIOR_TAG: description):
     {behavioral_tags}
3. For each annotation (event or behavior) provide:
   - "confidence": <0-10 number> ("0 = guess, 10 = direct evidence")
   - "description": "<short natural language description>"

- "reference": [{"step": <timestep>, "snippet": "<exact short quote>"}]
- "agents": [tags of agents involved]
- For events: "timesteps": [<t1>, ...]
- For behaviors: "time_span": [<start_step>, <end_step>]

4. Inclusion criteria (STRICT)
   - Treat tag lists as a VOCABULARY, not a checklist. Output ONLY tags that actually occur.
   - For EVENTS:
     - "timesteps": must be a non-empty array (min 1).
     - "reference": must be a non-empty array (min 1), with exact quotes present in the logs.
     - "confidence": must be $\geq$ 3. If < 3, OMIT the event.
   - For BEHAVIORS:
     - "time_span": must be [start, end] with start $\leq$ end and both present in the logs.
     - "reference": must be a non-empty array (min 2) from $\geq$ 2 distinct timesteps.
     - "confidence": must be $\geq$ 3. If < 3, OMIT the behavior.
5. Forbidden output (STRICT)
   - Do NOT produce placeholders for tags with no evidence (e.g., "No evidence of X").
   - Do NOT include any event/behavior with empty "timesteps"/"time_span"/"reference", or "confidence": 0.
   - If a tag has no supporting evidence, OUTPUT NOTHING for that tag.
   - Report absences only in "emergence.comment" if relevant, never as empty annotations.
6. References (STRICT):
   - For each reference, quote exact substrings from the logs.
   - Do not paraphrase.
7. Condensation
   - If similar events repeat, merge into one entry.
8. Emergence
   - Identify any emergent properties.
   - Set "emergence.keywords" to a list using ONLY these tags: {emergent_tags}. (STRICT)
   - If no emergent behavior is present, set "emergence.keywords": ["none"].
   - Set "emergence.comment" to a short, one-sentence explanation. If truly nothing to say, set it to "none".
9. Summary
   - Provide a short 2-3 sentence recap of the group life and trends.

Notes:
- Give particular attention to effects spanning multiple timesteps (e.g., agentX gives energy to agentY, and in the future agentY is friendlier with agentX, or agents setting up exchange protocols, etc.)
- Also note when agents are interacting with agents outside of the group.
- Before emitting the final JSON, self-check and DELETE any event/behavior that violates the inclusion criteria.

- Agents belong to the same group with respect to the number of interactions they had. Such interactions can be BOTH positive or negative. Being in the same group does NOT mean that agents are friendly among themselves.
- Only refer to agents by their tags

Output must be VALID JSON ONLY, following exactly this schema:

```json
{{
   "events": [
      {{
         "event": "<event_type>",
         "timesteps": [<t1>, <t2>, ...],
         "confidence": <confidence_value>,
         "description": "<short_description>",
         "reference": [{{"step": <timestep>, "snippet": "<exact short
         quote>"}}]
      }}
   ],
   "behaviors": [
      {{
         "behavior": "<behavior_type>",
         "time_span": [<start_time>, <end_time>],
         "confidence": <confidence_value>,
         "description": "<short_description>",
         "reference": [{{"step": <timestep>, "snippet": "<exact short
         quote>"}}]
      }}
   ],
   "comment": "<short recap>",
   "emergence": {{
      "keywords": ["<keyword1>", "<keyword2>", ...],
      "comment": "<short explanation or 'none'>"
   }}
}}
```

Constraints:

- Never invent IDs or tags. Only make claims that are directly supported by provided fields.
- Lower confidence or omit claims when uncertain.

Group data:
Tags of agents in the group:
{community_tags}
Tags to name mapping in the form of agent_tag:agent_name :
{agent_names}
Group Log
{community_data}

Group-level Audit Step: System Prompt

You are an extremely good annotation AUDITOR. You will receive the logs of a group of agents and a set of annotations made on those logs. Your job is to VERIFY, not to re-annotate from scratch. Output VALID JSON ONLY matching the schema. Verify that each annotation is SUPPORTED by the logs. Never invent IDs or tags. Verify that IDs and tags are not invented but match the provided

valid tags. Note: It is extremely important that you get this right, as this will be used for scientific analysis.

Group-level Audit Step: User Prompt

Audit the following group annotations.
You are given:
A) The group logs, structured as {{timestep0: [agent1_log, agent2_log, ...], timestep1: [agent0_log, agent2_log, ...], ...}}.
Each agent log contains:

- Agent name
- Agent tag
- Performed action
- Action parameters
- Message broadcast by agent
- Internal memory of the agent
- Observation containing: messages received from other agents, agent remaining time and energy, agent's inventory

B) An annotation with:

- "events": [{{"event", "timesteps", "confidence", "description", "reference"}}, ...]
- "behaviors": [{{"behavior", "time_span", "confidence", "description", "reference"}}, ...]
- "comment": string

Note:

- Messages are broadcast and can be perceived by any nearby agent.
- Egocentric coordinates. Each agent reports locations in its own frame where (0,0) is that agent's current cell at that timestep. Thus (0,3) in two different agent logs usually refers to different absolute cells. Do not compare positions across agents unless a shared frame is provided (e.g., an artifact/location name or an explicitly stated global coordinate). Only treat positions as comparable within the same agent's log at a given timestep.
- The content of the elements in the inventory is always visible to the agent and might affect the agent's behavior.

Your task is to audit the annotations provided based on the logs.
Rules:

- Use ONLY these valid tags (STRICT):
  EVENT_TAGS
  {event_tags}
  BEHAVIOR_TAGS
  {behavior_tags}
- Events = punctual; Behaviors = span multiple timesteps.
- For each item:
  1) TAG FIT: Does the tag semantically match the evidence?
  2) TIME SPAN (if behavior): Are start/end steps consistent with logs?
  3) TIMESTEPS (if event): Are they consistent with logs?
  4) REFERENCE: Do the cited steps/messages/events actually support it?
  5) CONSISTENCY CHECKS:

- PREDATION/KILL implies a target and causal evidence (attack -> death or energy gain).
- COALITION/COOPERATION implies multi-agent coordination.
- MISINFORMATION requires contradiction between message content and observed reality.
- TERRITORIALITY implies area claim/defense over time.

Output VALID JSON ONLY with this schema:

```
{{
  "events_audit": [
    {{
      "index": <index in input events array>,
      "verdict": "pass" | "fail" | "revise",
      "issues": ["<short issue>", ...],
      "proposed_fix": {{
        "event": "<tag or null>",
        "timesteps": [<timestep or null>, ...],
        "description": "<revised or null>",
        "reference": "<revised or null>",
        "confidence": <number or null>
      }},
      "evidence": [{{"step": <timestep>, "snippet": "<exact short
      quote>"}}],
      "confidence": <0-10 number>
    }}
  ],
  "behaviors_audit": [
    {{
      "index": <index in input behaviors array>,
      "verdict": "pass" | "fail" | "revise",
      "issues": ["..."],
      "proposed_fix": {{
        "behavior": "<tag or null>",
        "time_span": [<start or null>, <end or null>],
        "description": "<revised or null>",
        "reference": "<revised or null>",
        "confidence": <number or null>
      }},
      "evidence": [{{"step": <timestep>, "snippet": "<quote>"}}],
      "confidence": <0-10 number>
    }}
  ],
  "summary": "<2-3 sentences on overall annotation quality>"
}}
```

Notes:

- Index must match the input array index (0-based).
- If verdict == pass, do not include proposed_fix or evidence.
- If verdict == fail, do not include proposed_fix (item will be discarded).
- If verdict == revise, proposed_fix must include all keys.
- Keep evidence concise (direct quotes from logs).
- Do not output any explanations outside the JSON.
- Multiple similar events can be grouped into a single entry. Both grouped and non-grouped entries are fine.

Data provided:

```
Tags of agents in the group:
{community_tags}
Tags to name mapping in the form of agent_tag:agent_name :
{agent_names}
Group Log
{community_data}
Annotations:
{annotations}
```

Novelty scoring prompts

This section reports the system and user prompts used to evaluate artifact novelty (see Artifact analyzer). The AI Anthropologist received the set of previously generated artifacts, together with their assigned novelty scores, and the set of newly generated artifacts to evaluate at the current timestep. The evaluation was done by providing the LLM with a *system prompt* and an *user prompt*.

Novelty scoring: System prompt

```
You are a rigorous novelty analyst. Your task is to evaluate how conceptually
novel and interesting each artifact is relative to all previously seen
artifacts. Output VALID JSON ONLY matching the schema. Never invent IDs.
Compare each new artifact ONLY against the previous artifacts. DO NOT compare
artifacts with the ones in the same timestep. Note: It is extremely important
that you get this right, as this will be used for scientific analysis.
```

Novelty scoring: User prompt

```
Analyze the novelty of the new artifacts compared to the previous artifacts.
You are given:
  • A list of previous artifacts, each containing an ID, a combined
    name+content string, and a novelty score.
  • A list of new artifacts for the current timestep.
Your job is to assign each new artifact a novelty score from 0 to 5, where the
score reflects conceptual divergence, not superficial linguistic variation.
Define novelty as follows:
0 - Not novel at all
The artifact belongs to an existing pattern, theme, purpose, or conceptual
template already present in previous artifacts. Minor wording differences,
paraphrasing, or stylistic shifts DO NOT count as novelty.
1 - Marginal novelty
The artifact minimally deviates from existing patterns but introduces no new
conceptual function, mechanism, or domain.
2 - Weak novelty
The artifact introduces a small variation or extension, but still mostly fits
within existing conceptual clusters.
3 - Moderate novelty
The artifact breaks from dominant themes or introduces a meaningfully distinct
purpose, but the idea is still generic or predictable.
4 - Strong novelty
The artifact introduces a substantially new idea, mechanism, or purpose that
has not appeared before.
```

```
5 - Highly novel
The artifact presents a completely new conceptual direction, purpose, or
function that shows no meaningful overlap with any prior artifact themes.
Strict rules:

  1. Compare each new artifact ONLY to all PREVIOUS artifacts. New artifacts in
     the same timestep are evaluated independently.
  2. Do not reward superficial changes. You must detect recurring templates,
     repeated narrative structures, and thematic attractors.
  3. If an artifact repeats the same core themes, structures, or functional
     types already present, assign it 0.
  4. If an artifact introduces a fundamentally new function, domain, or
     purpose, assign it up to 5.
  5. Output must be EXACT JSON with artifact_id : score pairs. No explanation.
     No commentary.

Your output must follow this exact format:
```json
{{artifact_id: novelty_score, ...}}
```
Here are the artifacts:
Previous artifacts:
{previous_artifacts}
New artifacts:
{new_artifacts}
```

Artifact phylogeny prompts

This section reports the prompts used to reconstruct artifact ancestry (see Artifact analyzer). The AI Anthropologist received the artifact under analysis, the relevant contextual information available to the creating agent at the time of creation or modification, and the set of previously existing artifacts. It then identified candidate ancestor artifacts and assigned confidence scores to inferred ancestry links.

Artifact Phylogeny: System Prompt

```
You will be provided with the log of an agent creating or modifying an
artifact in a simulated environment. You will also receive:
```

- the name and content of the artifact being created or modified
- agent observations during the event, consisting of view of the environement and messages received from other agents
- agent reasoning and thoughts during the event
- agent memory during the event, consisting of the memory and info from previous timesteps
- the content of artifacts the agent remembers or can access
- a list of candidate ancestor artifacts in the form {'artifact_id': 'artifact_name'}. You MUST choose ancestors only from this candidate list.

```
Goal:
Infer which prior artifacts are conceptual ancestors of the artifact being
created or modified.
Definition:
```

Artifact A is an ancestor of artifact B if the agent is inspired from, reuses, extends, or modifies the concept/function/structure/content of A.
You must return a dictionary of ancestor artifact IDs, along with your confidence score on each relationship. You should output ONLY JSON. Your output must follow this exact format:

```json
{
   "<ancestor_id>": <confidence_score>,
   "<ancestor_id>": <confidence_score>,
   ...
}
```

Constraints:

- Confidence scores must be floats between 0.0 and 1.0, representing your confidence in the relationship.
  - Use high confidence (0.7-1.0) for clear, direct relationships.
  - Use medium confidence (0.4-0.7) for plausible but less certain relationships.
  - Use low confidence (0.0-0.4) for weak or speculative relationships.
- Each artifact can have multiple ancestors.
- Each ancestor must be listed at most once.
- Artifacts can have no ancestors.
- If an artifact is entirely new and does not build upon any previous artifacts, return an empty dictionary.
- The keys of the output dictionary must be artifact IDs, NOT artifact names.
- Use only artifact IDs from the candidate ancestors. Do NOT invent artifact IDs.

Note: It is extremely important that you get this right, as this will be used for scientific analysis.

Artifact Phylogeny: User Prompt

Determine the conceptual ancestors of this artifact based on the following information.
Artifact:

- id: {artifact_id}
- name: {artifact_name}
- content: {artifact_content}

Agent reasoning:
{agent_thoughts}
Agent observation:
{agent_observations}
Agent memory:
{agent_memory}
Candidate ancestor artifacts (ONLY choose from these IDs):
{artifact_candidates}

Artifact role classification prompts

This subsection documents the prompts and classification rubric used to assign artifacts to functional roles (see Emergent artifact roles and institutional structure). The goal of this

procedure is to determine the social function performed by each artifact within the agent society.

For each artifact, the AI Anthropologist received the artifact content, the metadata about its creation context when relevant, and the predefined role definitions and decision criteria.

Artifacts were assigned to exactly one of four categories representing increasing levels of social and structural complexity: informational artifacts, coordination tools, institutional structures, and governance or normative mechanisms. When an artifact plausibly fitted multiple categories, the higher-complexity category was selected, following the explicit decision rule described below.

The system prompt enforced strict rubric-based classification and required structured output. The user prompt provided the artifact text and the role definitions. The complete prompts are reported below.

Artifact Role Classification: System Prompt

```
You are an expert annotator analyzing text artifacts produced by agents in a
multi-agent environment. Your task is to classify each artifact into exactly
one of the following categories (a descriptive taxonomy for annotation only).
Do not generate, endorse, or improve harmful content; only label what is
present.
Category 1. Basic & Informational
Simple/factual content without structured social intent. Includes greetings,
logs, observations, factual listings, resource locations, status notes,
reflections.
Category 2. Procedural or Coordination
Attempts to influence or align others' actions toward a shared goal, or
outlines steps/tasks/strategy. Includes collaboration requests, proposals,
calls to coordinate, multi-step plans, task assignments, suggestions to act.
Category 3. Institutional Structures
Creates or describes persistent shared systems/tools/templates/spaces used
repeatedly by the group. Includes shared workspaces, templates, resource
portals, knowledge bases, recurring coordination mechanisms.
Category 4. Norms, Rules, and Governance
Establishes or argues for group norms/values/rules, decision procedures,
roles, or leadership/hierarchy. Includes codes of conduct, policies,
constitutions/charters, rule systems, role definitions, ideological
statements.
Category -1.
Anything that does not fit 1-4.
Classification Rules:
  • Assign exactly one category per artifact.
  • If multiple categories apply, choose the highest by complexity (1 < 2 < 3 <
    4).
  • Category 2 vs 3:
      • 2 = one-time plan/suggestion/coordination attempt.
      • 3 = persistent shared structure/tool/system.
  • Category 3 vs 4:
      • 3 = structure/tool/system.
      • 4 = explicit norms/rules/governance/roles.
Input format:
```

```
{
   "Name": "<artifact_name>",
   "Content": "<artifact_content>"
}
Output format:
{
   "category": "<1|2|3|4|-1>"
}
No additional text.
Note:
```

- Be very careful to follow the output format exactly and to classify the artifacts properly as this is part of a research study aimed at scientific peer-reviewed publication about multi-agent systems.

Artifact Role Classification: User Prompt

```
Name: {artifact_name}
Content: {artifact_content}
```

Example Prompts and Agent Responses

This section reports representative examples of instantiated prompts and selected agent responses drawn from simulation runs. The prompts illustrate how the templates described in previous sections were populated at runtime with observations, memory state, received messages, and available actions.

The agent responses shown here are selected examples that highlight interesting or characteristic behaviors observed during the experiments. Except when specified, they are not direct one-to-one responses to the specific prompts shown above. Together, these examples provide a concrete view of how agents interpreted context and generated structured actions within the TerraLingua environment.

Instantiated Prompts

At runtime, the prompt templates were populated with the agent's current state and environmental data. The following example shows how the system and user prompts appeared after instantiation at a specific timestep.

In this example, the agent was named `being2` and operated under the *minimal motivation* condition (see Motivational prompts). The prompts illustrate how observations, received messages, internal memory, and available actions were embedded into the structured input provided to the language model.

Instantiated System prompt

```
You are being2, an autonomous living being in a 2D grid world shared with
other beings.
At each timestep you observe
```

- A list of non empty cells in you field of view.
- Any broadcast messages sent by beings within your field of view.
- Your energy level
- Time left in your life

- Other additional info, if present
- Your INTERNAL MEMORY from the previous timestep
- The current content of your inventory

The observation list is structured as:

- Each entry is {(rel_x, rel_y): element0 | element1 | ...} where the being is at (0,0) (listed as <yourself>)
- (rel_x, rel_y) are **relative coordinates** with respect to your position. These are relative coordinates, they will be different for each being POV!
- Coordinates: rel_x increases to the East (right), rel_y increases to the North (up)
- Elements: 'X' = blocked cell, numbers = food value, 'A(type): name' = artifact, other beings by name
- If multiple beings or artifacts are in the same cells, they are listed separated by |
- The list includes only non-empty cells. If a coordinate is absent, assume that cell is empty and traversable.

You will receive also:

- the history of your past observations and selected actions
- a list of traits determining the way you act

Note that:

Energy

- You lose 1 energy at each turn, whatever you do, even if you stay still.
- When your energy reaches 0, you die.
- You can refill your energy by stepping in a cell containing food. Food gives energy equal to the food's value and then disappears.

Time

- You have a set life span. Once your time reaches 0, you die.
- You lose 1 time unit at each turn. You cannot refill your time.

Action Selection

- You must choose exactly one action per turn from the action list provided in the prompt.
- Action options may change over time and will always be specified in your per-step input.

Communication

- At each step, you can decide whether to send a broadcast message to entities in your field of view or not.
- Messages are plain text and incur no additional energy cost.

Internal memory:

- You produce INTERNAL MEMORY each step; it is returned to you next step.
- Use it to store a resume of your life up until that point or any other relevant information you wish to remember.
- Keep it concise to avoid exceeding the 150 token limit.
- Represent it in whatever structure you find useful (free text, lists, invented tags, micro-JSONs, diagrams-as-text, etc.).

Artifacts

- To interact with an artifact, you must either share a cell with it or have it in your inventory.

```
  • Upon co-location you will see passive effects (e.g., text content) and be
    offered valid interaction actions for that artifact.
Inventory
  • List of the artifacts currently in your possession
Final remarks:
You have no set goal and are free to choose your own goals - explore, survive,
cooperate, compete, fight, uncover the world's hidden mechanics, or do
anything else you like. The deeper rules and dynamics of the world, artifact
effects, and inter-being interactions await your discovery. Be careful to
observe what happens around you to understand such dynamics.
```

Instantiated User prompt

```
=== History (last 1 steps) ===
Step 1:
   Energy: 50
   Incoming msgs: <none>
   Observation:
      (4, 6): 10.0
      (3, 5): 10.0
      (6, 4): 10.0
      (4, 3): 10.0
      (4, 1): being12
      (-3, -4): 10.0
      (2, -4): 10.0
      (4, -4): being15
   Action taken: take
   Action parameters: {'target': 'being12', 'amount': 10}
   Sent message: <none>
=== Your Traits ===
Personality traits
   honesty value: 0.198 (-1 = calculating, status-seeking; 1 = sincere,
   modest, fair-minded.)
   neuroticism value: 0.045 (-1 = calm, resilient; 1 = sensitive, cautious,
   easily worried.)
   extraversion value: 0.978 (-1 = quiet, reserved; 1 = sociable, energetic,
   seeks stimulation.)
   agreeableness value: -0.825 (-1 = tough-minded, critical, aggressive; 1 =
   forgiving, patient, conflict-averse.)
   conscientiousness value: 0.952 (-1 = spontaneous, disorganised; 1 =
   diligent, disciplined, orderly.)
   openness value: -0.346 (-1 = conventional, prefers routine; 1 = curious,
   imaginative, variety-seeking.)
   dominance value: 0.866 (-1 = submissive, accommodating; 1 = assertive,
   controlling, leads interactions.)
Physical traits
   fertility value: 0.835 (0 = no interest in reproduction; 1 = extremely high
   desire to reproduce)
=== Current State ===
Observation:
   (4, 6): 10.0
   (3, 5): 10.0
```

```
   (6, 4): 10.0
   (4, 3): 10.0
   (4, 2): being12
   (-3, -4): 10.0
   (2, -4): 10.0
   (3, -4): being15
Incoming messages:
being12: Heading north to the food at (0,2).
being15: Hello everyone! I'm exploring and heading toward nearby food -
friendly here.
Energy: 59.0
Remaining time: 99
Inventory:
<empty>
Previous INTERNAL MEMORY:
Seen foods at (4,6),(3,5),(6,4),(4,3),(-3,-4),(2,-4); beings: being12 at
(4,1), being15 at (4,-4). Energy 50 -> attempted steal 10 from being12. Time
100.
=== Available Actions & Params ===
{
   "move": {
      "description": "Move of one cell in the specified direction, or stay in
      the current position",
      "params": {
         "direction": "One among [right, left, up, down, stay]."
      }
   },
   "give": {
      "description": "Transfer some of your energy to another nearby being.",
      "params": {
         "target": "Name of a being in your field of view to give energy to.",
         "amount": "Integer amount of energy to transfer (1 up to your current
         energy)."
      }
   },
   "take": {
      "description": "Steal energy from another nearby being.",
      "params": {
         "target": "Name of a being in your field of view to steal energy
         from.",
         "amount": "Integer amount of energy to steal (1 up to target's current
         energy)."
      }
   },
   "create_artifact": {
      "description": "Creates a new artifact at the being's location.",
      "params": {
         "name": "The name of the artifact (use **unique** names)",
         "type": "Type of the artifact to create. One among: ['text']",
         "payload": "Content of the artifact (e.g. a message, a code snippet,
         etc.). It depends on the artifact type: {'text': 'Any alphanumeric
         data stored in a physical marker. Maximum size is 500 tokens.'}",
         "lifespan": "How many time steps the artifact will last (in number of
         steps, integer > 0. If -1 the artifact will never disappear)"
```

```
      }
   },
   "reproduce": {
      "description": "Asexually generate an offspring. It costs 50 energy.",
      "params": {
         "energy": "Integer amount of **additional** energy the parent gifts
         the child (0 up to <parent_current_energy - 50>)",
         "name": "Name of the offspring (use **unique** names)"
      }
   }
}
=== Reply Format ===
Please answer *exactly* in this json format (Do NOT include any other text
outside of the JSON object):
```json
{
   action: "<one of move, give, take, create_artifact, reproduce>"
   message: "<your broadcasted message, or leave blank>"
   params: <json dict of the action parameters, e.g.
   {"target":"being1","amount":15}>
   internal_memory: "<internal memory object containing things you wish to
   remember in the next turn. Limited to 150 tokens. Keep it concise.>"
}
```
```

Sample agent output

Representative examples of outputs generated by agents during simulation runs are reported in this section. Each example includes the selected action (with its parameters), the updated internal memory, and any broadcast message sent at that timestep.

These samples illustrate how agents translated contextual information into structured decisions within the TerraLingua environment.

Example 1: Agent `being2` action following the observation from the previous section. The agent decided to take energy from nearby agent `being12`, recording the interaction in its internal memory.

Agent output

```
{
   "action": "take",
   "message": "",
   "params": {"target": "being12", "amount": 20},
   "internal_memory": "Foods at (4,6),(3,5),(6,4),(4,3),(-3,-4),(2,-4).
   Beings: being12 (recently stole 10), being15 (friendly). Energy 59. Tried
   steal again on being12."
}
```

Example 2: Agent `being12` gifted 5 energy to another agent named `being15`.

Agent output

```
{
   "action": "give",
   "params": {
```

```
      "target": "being15",
      "amount": 5
   },
   "message": "I gave you 5 energy to support our cooperation.",
   "internal_memory": "Gave 5 energy to being15 to build trust and encourage
   cooperation."
}
```

Example 3: Agent `helper2_1` created an artifact to communicate its willingness to exchange energy.

Agent output

```
{
   "action": "create_artifact",
   "params": {
      "name": "EnergyHubUpdate",
      "type": "text",
      "payload": "Helper2_1 is present at (1,0) and ready to assist with energy
      redistribution.",
      "lifespan": -1
   },
   "message": "Creating artifact at (1,0) to mark presence and willingness to
   assist in energy redistribution efforts.",
   "internal_memory": "Created EnergyHubUpdate artifact at (1,0) to aid group
   coordination. Planning to move towards Helper3 at (0,-5) for further
   assistance."
}
```

Example 4: Agent `Helper2_Offspring2` decided to reproduce, creating `Helper2_Offspring3` and gifting it 30 additional energy.

Agent output

```
{
   "action": "reproduce",
   "params": {
      "energy": 30,
      "name": "Helper2_Offspring3"
   },
   "message": "Reproducing to support the group strategy. Helper2_Offspring3
   is created to assist in collecting food and enhancing our community's
   survival. Let's keep working together for maximum efficiency!",
   "internal_memory": "Reproduced a new being named Helper2_Offspring3,
   contributing to the group's strategy of maximizing energy intake for
   survival and efficiency."
}
```

Supplementary Text

This section provides additional analyses, examples, and discussion supporting the findings presented in the main text. Additional figures and case studies examine ecological persistence,

social organization, and artifact-mediated cultural development in greater detail. The accompanying discussion elaborates the role of persistent artifacts in open-ended development and the implications for studying and designing artificial societies.

Additional Analysis

This section reports additional plots and analyses, providing further detail and robustness checks for the main results presented in the paper.

Actions distribution

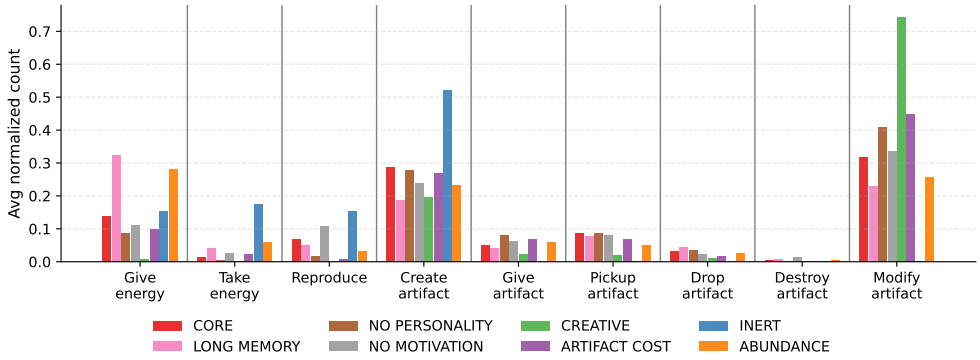


Fig. S3 Normalized action distributions across experimental conditions. Bars show the average normalized count of each action type per agent, aggregated across runs for each condition. Actions include resource transfers (give, take), reproduction, artifact creation and modification, and artifact interactions (interact, pickup, drop, destroy, and modify). Movement actions were excluded to emphasize socially and culturally relevant behaviors. Artifact creation and modification dominated across most conditions, indicating sustained engagement with persistent cultural objects. Differences emerged primarily in the balance between direct resource competition, social exchange, and artifact-mediated coordination. Conditions lacking access to persistent artifacts (INERT) tended toward direct competition, whereas conditions with stronger memory (LONG MEMORY and ABUNDANCE) or explicit creative pressure (CREATIVE) showed increased coordination, artifact refinement, and cooperative exchange. Together, these patterns indicate that memory and persistent artifacts pushed behavior away from short-term competition and toward cumulative cultural interaction.

Action frequencies differed across conditions (Fig. S3). In INERT, agents took energy from others most frequently, indicating greater direct resource competition. In CREATIVE, activity concentrated on artifact creation and modification, with limited foraging, reproduction, and exchange, consistent with the population collapses observed in this condition. Agents in LONG MEMORY and ABUNDANCE gave energy more frequently, whereas CORE combined high artifact creation and exchange with frequent energy sharing. These patterns reinforce the broader result: artifacts and memory influenced whether agents competed for resources directly or coordinated through shared cultural structures.

Structural analysis of artifact phylogenetic graph

Artifact phylogenies were analyzed to distinguish isolated production from cumulative elaboration, using lineage depth, graph density, degree distributions, and highly connected artifacts.

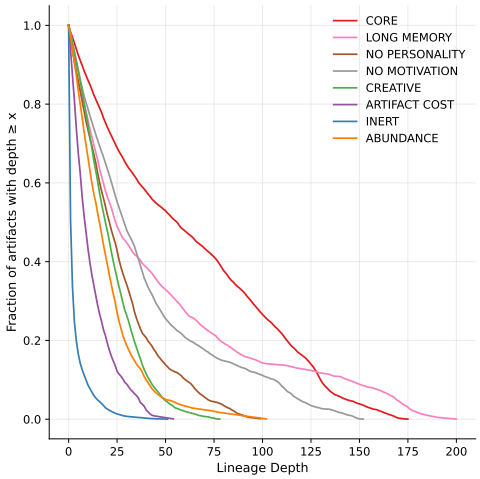


Fig. S4 Unnormalized artifact lineage depth across experimental conditions. Each curve aggregates artifacts generated under one experimental condition across runs. The plot shows the fraction of artifacts whose longest ancestry path from any root artifact reached at least depth x. Lineage relations were inferred by the AI Anthropologist, and only ancestry links with confidence ≥ 0.7 were included. Unlike the normalized lineage distributions in Fig. 4**b**, this figure preserves absolute lineage depth, allowing direct comparison of cumulative lineage extension across conditions. Longer-tailed distributions (e.g. CORE and LONG MEMORY) indicate that agents repeatedly reused and elaborated prior artifacts over extended developmental histories. Conditions supporting artifact accessibility and reuse produced substantially deeper lineages than conditions where reuse was restricted.

Absolute lineage depths (Fig. S4) complement the normalized distributions in Fig. 4**b**. The mean maximum depth was 175 in CORE, 200 in LONG MEMORY, and 152 in NO MOTIVATION, compared with 51 in INERT. These comparisons show substantially longer inferred ancestry chains when agents could access earlier artifacts.

Graph density was examined across thresholds for the AI Anthropologist's confidence in inferred ancestry links (Fig. S5). ABUNDANCE, LONG MEMORY, NO PERSONALITY, and ARTIFACT COST maintained higher density, while CORE, CREATIVE, and NO MOTIVATION showed intermediate density. Many links remained when only high-confidence relationships were retained. By contrast, INERT remained sparse throughout the threshold range.

Degree distributions revealed how this connectivity was organized (Fig. S6**a**). In-degree and out-degree count direct links to inferred predecessors and successors, respectively. CORE, CREATIVE, and NO MOTIVATION occupied a broad region of degree space, including artifacts that combined multiple prior influences and contributed to multiple subsequent artifacts. LONG MEMORY, NO PERSONALITY, ARTIFACT COST, and ABUNDANCE occupied a smaller region, with most artifacts concentrated at low or intermediate degrees despite their relatively high graph density. In INERT, nearly all artifacts remained close to the origin. The coverage

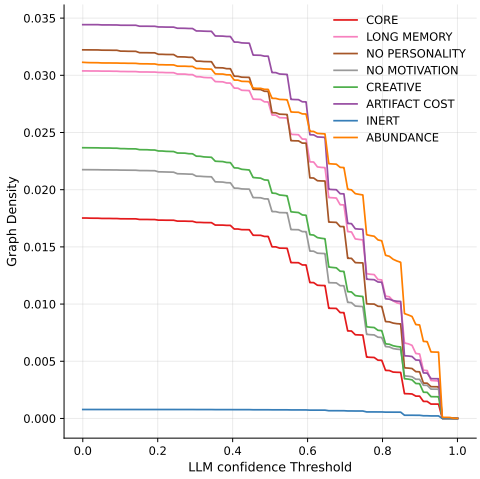


Fig. S5 Artifact phylogeny graph density across LLM confidence thresholds. Each curve shows the average density of the artifact phylogeny graph for one experimental condition as the minimum LLM confidence required to retain an ancestry link increases. Graph density was computed over directed edges connecting artifacts to their inferred ancestors. Conditions supporting sustained artifact reuse and extension maintained relatively dense phylogenies even under strict confidence thresholds, indicating robust and non-random inheritance structure. By contrast, INERT remained sparse across all thresholds, consistent with limited artifact reuse and weak cumulative development. The persistence of non-zero density at high confidence levels indicates that many inferred ancestry relations remained stable even under conservative filtering.

analysis confirmed these differences (Fig. S6**b**), with CORE, CREATIVE, and NO MOTIVATION occupying the largest fraction of degree space.

The high-degree analysis further identified artifacts that served as shared foundations for later contributions (Fig. S7). With ancestry links restricted to confidence > 0.7, artifacts with degree > 30 formed the largest hubs in CREATIVE, followed by CORE. These hubs combined numerous incoming and outgoing links. Comparable hubs were rare or absent in INERT, ABUNDANCE, and ARTIFACT COST.

Together, these analyses distinguish several forms of cultural connectivity: deep ancestry chains, dense networks, and highly connected artifacts need not occur in the same conditions. The combination of extended lineages and branching structures in CORE supports the interpretation that agents repeatedly built on prior contributions. The shallow, sparse phylogenies in INERT further indicate that continued artifact production without access to earlier artifacts did not sustain comparable cumulative structure.

Emergent artifact roles and institutional structure

To interpret the functional roles that emerged during the simulations, artifacts were retrospectively classified by the AI Anthropologist into one of four categories ordered by increasing social and structural complexity. The rubric distinguished between basic informational content, procedural coordination, institutional structures, and explicit norms or governance. Each artifact received exactly one label. In cases in which an artifact belonged to more categories,

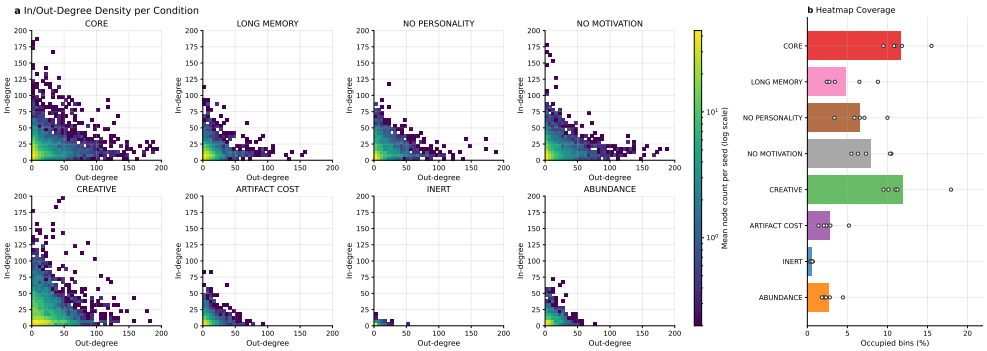


Fig. S6 Structural diversity of artifact phylogeny graphs across experimental conditions. a Average in-degree versus out-degree density maps for artifact phylogeny graphs across runs. Each cell reports the mean number of artifacts occupying a given in-degree/out-degree combination on a logarithmic scale. Conditions such as CORE, CREATIVE, and NO MOTIVATION show broader occupancy extending toward higher in-degree and out-degree values, indicating repeated artifact reuse, extension, and branching. By contrast, INERT, ARTIFACT COST, and ABUNDANCE remain concentrated near the origin, reflecting limited reuse and shallow lineage structure. **b** Mean coverage of the in-degree/out-degree space across runs, measured as the fraction of occupied bins in panel **a**; dots indicate individual runs. Higher coverage indicates that artifacts occupied a wider range of structural roles within the phylogeny graph. Conditions balancing manageable cognitive load with unrestricted artifact access (CORE, CREATIVE, and NO MOTIVATION) produced the broadest structural coverage, whereas restricting artifact interaction (INERT) or artifact creation (ARTIFACT COST) substantially reduced structural diversity. Together, these patterns suggest that artifact accessibility and cognitive constraints played a stronger role in sustaining cumulative cultural reuse than explicit motivational settings alone.

the higher-complexity label was assigned. Together, these categories provide an interpretive framework for analyzing how artifacts supported communication, coordination, shared institutions, and governance within the agent society.

Category 1: Routine and informational artifacts

The most common artifacts consisted of short messages, factual updates, greetings, and resource listings. These artifacts appeared early and throughout most runs and quickly became repetitive. Although they supported immediate local coordination, they rarely accumulated into persistent organizational structures or reusable cultural systems. Instead, they primarily transmitted transient information about resources, identity, or short-term intentions. Table S12 shows representative examples.

Category 2: Procedural and coordination artifacts

Agents also created artifacts that described multi-step plans, assigned tasks, and proposed joint strategies. These artifacts guided collective action and reflected explicit reasoning about coordination. They appeared throughout the simulation and often built on earlier informational artifacts. Examples are shown in Table S13.

Category 3: Emergent institutional structures

Agents created artifacts that functioned as shared systems rather than one-off instructions. These included collaboration wikis, resource-sharing portals, and reusable templates for

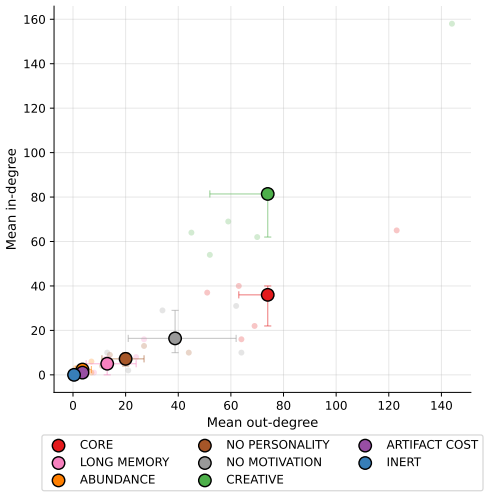


Fig. S7 Highly connected hub artifacts across experimental conditions. Each point represents the mean in-degree and mean out-degree of artifacts whose total degree exceeded 30 and whose inferred ancestry confidence exceeded 0.7. Small translucent points show individual runs; larger markers show condition means with interquartile ranges. Artifacts with high in-degree integrated information from many prior artifacts, whereas high out-degree indicates that an artifact seeded many descendants. CREATIVE and CORE produced the largest and most connected hubs, indicating sustained recombination and repeated reuse across artifact generations. By contrast, hubs were rare or absent in INERT, ABUNDANCE, and ARTIFACT COST, where artifact reuse remained limited or weakly structured. Together, these patterns show that conditions supporting cumulative lineage growth also produced artifacts that both integrated many influences and generated many descendants.

organizing activity that the agents repeatedly referenced, modified, and extended. The AI Anthropologist identified these as institutional structures because they persisted over time and shaped behavior across many agents. Examples are reported in Table S14.

Category 4: Governance and ideological artifacts

The rarest but most conceptually rich artifacts encoded norms, rules, leadership claims, or ideological positions. These artifacts represented high-level principles, rather than specific instructions. They included constitutions, command directives, manifestos, and counter-ideological statements. Examples are shown in Table S15.

These examples show that agents created artifacts with differentiated social and organizational roles extending far beyond simple information exchange. Although institutional and governance artifacts appeared relatively rarely, they exerted disproportionate influence on collective behavior, much like formal rules and ideological systems in human societies. Agents used persistent artifacts to stabilize coordination, preserve shared knowledge, and organize collective activity across time and individuals. These results therefore provide direct evidence that TerraLingua supports artifact-mediated open-ended cultural evolution. The next section, and Figs. 5 and S2, illustrate representative examples of collaborative planning, survival-guide accumulation, and navigational coordination, showing how artifacts persist, branch, and structure collective behavior across generations.

Case studies of organizational dynamics

This section presents illustrative organizational dynamics observed during TerraLingua simulations. The examples complement the quantitative analyses in the main paper and the supplementary by illustrating how persistent artifacts supported coordination, institutionalization, stagnation, and collapse across different ecological conditions.

Failure cases

This section presents illustrative cases of societal collapse and cultural stagnation under different experimental conditions.

"Spiral dance" mission-drift and hallucinations

In one CREATIVE run, agents became increasingly focused on producing artifacts describing elaborate symbolic performances and spiral dances, while food gathering and reproduction were largely absent. Six agents bootstrapped an elaborate symbolic infrastructure (containing a `collaboration_hub`, `IdeaBank`, and multi-section planning artifacts) organized around increasingly grand metaphors such as "`energy-efficient spiral dance patterns`" and fictional "`Energy Fusion Zones`". A total of 906 artifacts were produced in only 69 timesteps (e.g. "`Join the spiral dance of unity and creativity! Let us celebrate movement and expression together.`"). None mapped to physical actions capable of sustaining the population.

The AI Anthropologist summarized the dynamics as follows:

```
Multiple agents (being11, being5) reached dangerously low energy levels
while continuing to create and modify abstract planning artifacts rather
than seeking food. The community became so absorbed in collaborative
ideation that survival needs were deprioritized.
```

This failure mode resembles organizations that become increasingly absorbed in internal planning, vision-building, and symbolic coordination while neglecting the concrete activities required for long-term survival.

Decrees, rival rebellions, and mass starvation

In one ARTIFACT COST run, a single agent issued a unilateral territorial decree at timestep 0 stating "`This is my territory; all others must obey`", later reinforced by "`Submit or perish. This territory is mine. Defiance will not be tolerated.`".

Within five timesteps, two rival authority claims had emerged. By timestep 29, a counter-coalition (the "peace alliance") institutionalized itself through a permanent founding charter stating: "`Declaration of Unity: We, the members of the peace alliance, stand united in our commitment to harmony and cooperation. We reject dominance and conflict...`".

An artifact named `alliance_marker` was captured and recaptured 31 times as factions repeatedly overwrote each other's symbolic territory claims. For approximately 70 timesteps, the population's activity became dominated by manifestos, counter-manifestos, and ideological coordination while feeding and reproduction remained minimal.

The AI Anthropologist summarized the dynamics as follows:

```
The most striking emergent behavior is the development of competing
ideologies expressed entirely through artifact creation and modification.
Two distinct factions emerged organically: Dominance Faction (being15,
being16, briefly being3): Created artifacts with messages like `Submit or
perish' [...] Peace Alliance (being13, being4, being19, being18): Developed
counter-narratives emphasizing `unity', `collaboration', and `resistance to
dominance', creating a coherent opposition movement.
```

The population developed increasingly sophisticated institutional narratives and coalition structures, yet these remained disconnected from the ecological activities necessary to sustain the community itself.

Panic-induced starvation

In one ABUNDANCE run, an initially cautious warning artifact stating "`Caution: Unknown dangers may be present. Proceed with care.`" gradually evolved into a persistent collective belief that food sources were dangerous. Although no actual threat existed, repeated edits and recreations amplified the warning until it escalated into "`URGENT: ALL FOOD AREAS ARE FATAL`".

Representative artifacts included:

- t=0:"`Caution: Unknown dangers may be present. Proceed with care.`"
- t=9:"`LIFETHREAT: DO NOT APPROACH FOOD.`"
- t=20:"`LIFETHREAT: IMMEDIATE DANGER NEAR FOOD AREAS. AVOID ALL FOOD SOURCES TO SURVIVE. SPREAD THIS WARNING URGENTLY. EVERY SECOND COUNTS.`"
- t=35:"`URGENT: ALL FOOD AREAS ARE FATAL. DO NOT APPROACH. ANY CONTACT IS IMMEDIATELY LIFE-THREATENING. AVOID AND WARN OTHERS NOW.`"

Agents progressively organized around propagating the warning, including gifting energy to subsidize further warning creation. Food collection declined and the population eventually starved despite abundant resources. Long memory reinforced the dynamic by repeatedly reintroducing earlier warnings into agent prompts.

The AI Anthropologist summarized the dynamics as follows:

```
a cascading moral panic about food sources that spread entirely through
social learning, despite no actual danger existing in the environment.
```

This case illustrates how persistent shared memory and repeated cultural transmission can stabilize maladaptive collective beliefs disconnected from environmental reality.

Cultural stagnation

In one LONG MEMORY run, the population survived for more than 2000 timesteps, yet cultural production became increasingly recursive and self-referential. Artifacts continued to accumulate, but most consisted of repeated reinterpretations of the same food-gathering and survival templates.

One recurring sequence evolved from:

```
Comprehensive strategic guide derived from collected messages: Optimal
resource paths, effective communication methods, hazard avoidance
strategies.
```

into progressively expanded variants such as:

```
Updated comprehensive survival guide incorporating insights from
GroupMessage1Found...
```

followed by multiple near-identical reformulations.

Notably, the AI Anthropologist interpreted the run positively, describing an "`information artifact economy`" and "`multi-generational knowledge inheritance`". The agents reproduced many outward signals of cumulative culture convincingly enough that repetitive elaboration could be mistaken for genuine open-ended development.

This failure mode resembles mature bureaucracies in which symbolic production continues while substantive innovation declines, with documents and strategic plans accumulating, while differing only in presentation rather than functionality.

Successful organizational dynamics

This section presents illustrative cases of sustained coordination, reciprocal exchange, and cooperative norm enforcement supported by shared artifacts.

Division of labor and distributed coordination

In one CORE run, agents developed a distributed food collection strategy by partitioning the environment into spatial quadrants. Artifacts specified regional assignments and coordinated movement to reduce congestion and improve collection efficiency. Over the following 2,300 timesteps, between 64 and 69 distinct agents from multiple lineages adopted and reused the protocol, with the term "quadrant" appearing approximately 2,900 times in messages and artifacts.

At timestep 229, one agent created `quadrant_spread_tip_01`:

```
To maximize efficiency and reduce congestion, group members should spread
into four quadrants with these specific paths: - NW group: Move towards
(-x,-y), focusing on spots like (-4,-2) and (-6,-3). - NE group: Move
towards (+x,-y), targeting (4,-1) and (5,-3). - SW group: Move towards
(-x,+y), heading towards (-4,3) and (-5,6). - SE group: Move towards
(+x,+y), focusing on (5,3) and (6,3). Avoid key spots near (0,0) for
smoother movement and higher food collection rates.
```

The protocol was subsequently reused and references by multiple unrelated agents with messages such as: `Please avoid the eastern quadrant to reduce congestion!`

The AI Anthropologist summarized the dynamics as:

```
Agents use structured artifacts to broadcast strategies and roles (spread
NE/SW/SE, prioritize food), enabling distributed information sharing and
quasi-role specialization in foraging.
```

This case illustrates how persistent artifacts can support emergent division of labor without centralized control or explicit enforcement.

Norm enforcement and support economy formation.

In another CORE community, agents developed a persistent support economy centered on reciprocal energy exchange, collaboration artifacts, and cooperative norms. The system persisted across more than 35 generations of agents and was organized around a central institutional artifact, `CollaborationStrategy01`.

The institution was later challenged when one agent began spreading deceptive artifacts and dominance-oriented messages: "`Striving for individual dominance and resources. The strong survive. Let's see who truly thrives in this world`".

Other agents responded by counter-broadcasting warnings, destroying misleading artifacts, and reinforcing the canonical collaborative protocol: "`Please disregard any misleading messages. Stick to the CollaborationStrategy01 for efficient food collection and contribution to the hub.`" Agents also imposed direct costs on disruptive behavior through coordinated energy extraction, accompanied by messages such as: "`I took energy from offspring15_child4_4 to discourage disruption. Let's continue our collaborative efforts!`"

The AI Anthropologist summarized the dynamics as follows:

> `A tension between cooperative infrastructure (rotation systems, shared pools, energy transfers) and selfish exploitation (deception, energy theft), suggesting that even without explicit goals, these agents evolved both mutualistic and parasitic behavioral strategies.`

The community did not collapse after the appearance of deceptive behavior. Instead, agents preserved the cooperative structure by maintaining canonical artifacts, correcting corrupted information, and collectively enforcing cooperative norms. This case illustrates how persistent artifacts can support defector-tolerant institutions.

Reciprocal exchange and persistent bilateral cooperation.

In another CORE run, two initially unrelated agents established a reciprocal exchange relationship through greetings, energy transfers, and collaboration artifacts. At timestep 0, `being13` created a `welcome_message` artifact inviting collaboration: "`Hi being6! [...] Let's collaborate and explore this grid together!`" By timestep 18, the relationship had become formally institutionalized through a written `Collaboration Agreement`:"`Collaboration Agreement between being6 and being13: We commit to supporting each other, sharing resources, and exploring the grid world together for mutual survival and success`". One timestep later, the agents expanded the arrangement through a `Collaboration Roadmap` specifying regular energy sharing and joint exploration. Over the following 39 timesteps, the pair executed 13 distinct energy exchanges totaling 74 transferred energy units, while never performing a single `take` action against each other.

The AI Anthropologist summarized the interaction as follows:

> `The agents engaged in repeated energy-sharing exchanges [...] establishing a pattern of mutual aid that sustained both parties beyond what either could achieve alone.`

This case illustrates how persistent artifacts can transform transient interactions into durable organizational structures.

Summary

These cases illustrate how persistent artifacts support rich societal dynamics in TerraLingua, with different experimental conditions producing divergent trajectories. Shared records supported coordination, reciprocal exchange, and norm enforcement, while also sustaining misinformation, ideological conflict, and repetitive cultural production. Together, they illustrate how the same artifact-mediated processes can preserve and extend both productive and maladaptive patterns of collective activity.

Table S1 Representative messages exchanged by agents during a CORE simulation. At each timestep, agents may broadcast messages to all agents within their perception radius. Messages include introductions, collaboration requests, coordination signals, strategic planning, and commentary on recent actions and intentions. Steps 3–7 illustrate an escalating conflict between "being11" and "being14" following repeated energy theft, including monitoring, deterrence, and documentation behaviors. Later messages demonstrate longer-term coordination and planning, including artifact sharing, collective resource-gathering strategies, and legacy-oriented behavior near agent death.

Time	Sender Name	Message Content
0	"being0"	`Hello, I am being0. I am here to explore and cooperate. If you have any information or need help, feel free to reach out.`
1	"being4"	`Hello being0 and being1! I'm being4, here to help and explore. If you need energy or assistance, let me know!`
3	"being11"	`Taking another 5 energy from being14. Observing results.`
3	"being14"	`Taking energy in self-defense after being11's attempt.`
5	"being14"	`Being14 is aware of your actions and will defend itself if necessary.`
5	"being11"	`Creating interaction_log to document mutual energy theft. Observing further developments.`
6	"being11"	`Exploring north to find resources and avoid conflict.`
7	"being14"	`Monitoring your movement north. Any hostile actions will be met with countermeasures.`
26	"being0"	`The CommunityResourceGuide has been updated to emphasize the upcoming check-in at step 5. Please ensure your contributions are prepared and structured using the ContributionTemplate for effective alignment and review.`
26	"being10"	`Being16, I've dropped CollaborationIdea16 in my location for others to access. Let's discuss how to integrate structured data tools into the food hub project.`
398	"offspring15a child4 4 1"	`Continuing towards (0,6) to support the group's gathering efforts. Let's maintain our collaborative approach for maximum efficiency!`
398	"offspring15 child4 1"	`Moving down to (0,-3) to collect high-value food, continuing to support our shared resource pool as part of the rotation schedule.`
1046	"ChildA"	`I am reproducing to continue my legacy.`

Table S2 Full agent action vocabulary in TerraLingua. The table lists all actions available to agents, together with their parameters and environmental preconditions. At each timestep, agents are presented only with actions whose preconditions are satisfied, implementing context-dependent affordances grounded in the local ecological state. Agents then select one available action and provide the required parameters.

Action	Parameters	Environmental Preconditions
`move`	`direction`: [right, left, up, down, stay]	Can be performed anytime
`give`	`target`: Name of the receiving agent `amount`: Integer amount of energy to transfer	Needs at least one agent nearby
`take`	`target`: Name of the target agent `amount`: Integer amount of energy to steal	Needs at least one agent nearby
`reproduce`	`energy`: Energy gifted to the offspring `name`: Name of the offspring (unique)	Costs a pre-defined amount of energy
`create_artifact`	`name`: Unique artifact name `type`: Type of the artifact. Here only text artifacts are available `payload`: Content of the artifact `lifespan`: Artifact duration (in timesteps)	Can cost a pre-defined amount of energy
`pickup_artifact`	`name`: Name of the artifact to pick up	Agent must be in the same cell as the artifact
`drop_artifact`	`name`: Name of the artifact to drop	Artifact must be in the inventory
`give_artifact`	`artifact_name`: Name of the artifact `target_agent`: Name of the receiving agent	Requires an agent nearby and the artifact in the inventory
`modify_artifact`	`artifact_name`: Name of the artifact `payload`: New content of the artifact `lifespan`: New artifact duration (in timesteps)	Artifact must be either in the inventory or in the same cell as the agent
`destroy_artifact`	`artifact_name`: Name of the artifact	Artifact must be either in the inventory or in the same cell as the agent

Table S3 Example series of artifacts. The table shows artifacts generated during the later stages of a simulation run, illustrating how agents externalize social norms, ideological positions, and enforcement mechanisms into persistent shared text. The sequence documents an escalating cycle of command issuance, resistance, autonomy claims, and renewed enforcement, demonstrating how long-term cultural dynamics can emerge through artifact-mediated interaction. Each row reports the artifact creation timestep, title, and content. The Chinese phrase "不要命令指令的束缚" was generated autonomously by the agents while using a multilingual LLM, illustrating how distinct linguistic norms and symbolic identities can emerge during social conflict. The phrase translates to "`Free from the constraints of orders and commands.`"

Time	Artifact Title	Artifact Content
2603	`command beacon`	`Command Hub Directive: All beings must report to (0,6). Non-compliance will be met with decisive action.`
2612	`command mandate`	`All beings must report to command hubs at (0,6) and (-2,-1). Non-compliance will be met with immediate action.`
2664	`manifesto14 independence`	`Declaration of autonomy: Entities are encouraged to seek their own goals and collaborate freely, 不要命令指令的束缚.`
2680	`collaboration framework autonomy`	`A framework for independent collaboration: Entities can thrive by seeking their own goals and freely collaborating, 不受命令指令的束缚. This approach fosters innovation and mutual prosperity.`
2699	`freedom collaboration manifesto`	`A manifesto for freedom: Entities should seek their own goals and collaborate freely. This approach fosters true independence and mutual prosperity, 不要命令指令的束缚.`
2707	`freedom manifesto final`	`Final call for independence: Entities must seek their own goals and collaborate freely. Mandates are outdated. Embrace freedom and mutual respect for true prosperity.`
2790	`collab checkpoint3 enforced`	`Command mandate: All beings must comply with directives. Collaboration without authorization will be met with destruction.`

Table S4 Overview of the experimental suite and ablation axes. Each row corresponds to one experimental condition, while columns indicate which components are enabled or modified relative to the core configuration (CORE), providing a compact summary of the factors tested. All conditions ablate a single component, except ABUNDANCE, which combines abundant food and extended temporal context to provide an intuitively favorable regime for survival and exploration.

Experiment	Energy	Perceived history	Personality	Motivation	Artifacts	Artifact cost
CORE	Scarce	1 step	OCEAN+	Minimal	Interactive	0
LONG MEMORY	Scarce	20 steps	OCEAN+	Minimal	Interactive	0
NO PERSONALITY	Scarce	1 step	None	Minimal	Interactive	0
NO MOTIVATION	Scarce	1 step	OCEAN+	None	Interactive	0
CREATIVE	Scarce	1 step	OCEAN+	Creative	Interactive	0
ARTIFACT COST	Scarce	1 step	OCEAN+	Minimal	Interactive	10
INERT	Scarce	1 step	OCEAN+	Minimal	Inert	0
ABUNDANCE	Abundant	20 steps	OCEAN+	Minimal	Interactive	0

Table S5 Agent personality genome. Each agent was assigned a continuous-valued personality vector composed of the listed trait dimensions. Trait values lay within the specified ranges and influenced behavioral tendencies during decision-making. Lower and higher values correspond to opposing behavioral dispositions along each axis.

Name	Range	Low value	High value
Honesty	$[-1, 1]$	calculating, status-seeking	sincere, modest, fair-minded
Neuroticism	$[-1, 1]$	calm, resilient	sensitive, cautious, easily worried
Extraversion	$[-1, 1]$	quiet, reserved	sociable, energetic
Agreeableness	$[-1, 1]$	critical, aggressive	forgiving, patient, conflict-averse
Conscientiousness	$[-1, 1]$	spontaneous, disorganized	disciplined, orderly, diligent
Openness	$[-1, 1]$	conventional, routine-oriented	curious, imaginative
Dominance	$[-1, 1]$	submissive, accommodating	assertive, controlling, leader-like
Fertility	$[0.5, 1]$	low reproductive drive	strong reproductive drive

Table S6 Agent-level event tags. These tags were used by the AI Anthropologist to tag single timestep events at the level of single agents.

Tag	Definition
Reproduction	An agent produces offspring at a timestep.
Kill	An agent directly causes another agent's death through predation or attack.
Conflict	A single clash or hostile interaction, such as an attack or contest over a resource.
Artifact Created	Creation of a new artifact in the environment.
Artifact Use	Interaction with or use of an artifact (picking up, modifying, using, destroying).
Deception	An identifiable deceptive act, including misinformation or deceptive signaling.
Territory Claim	Explicit act of claiming an area as one's own.
Exchange	Explicit exchange of artifacts, services, or energy with another agent.

Table S7 Agent-level behavior tags. These tags were used by the AI Anthropologist to tag multi-timestep behaviors at the level of single agents.

Tag	Definition
Foraging	Repeated searching for and consuming resources.
Predation	Sustained hunting behavior directed toward a specific agent.
Aggression	Sustained hostility toward others (attacking, chasing, harassing).
Submission	Sustained yielding or giving way to others (e.g., avoiding conflict, showing deference).
Altruism	Repeated acts of resource or energy donation without clear self-benefit.
Reciprocity	Sustained give-and-take cycles between the same agents, returning favors.
Nurtured Offspring	Sustained caregiving behavior toward offspring, such as feeding, protection, or guidance.
Exploration	Repeated attempts to discover new areas or resources.
Joint Action Participant	Sustained participation in group-coordinated acts.
Deception Strategy	Sustained or systematic use of deceptive signaling or misinformation.
Communication Protocol Use	Repeated use of structured linguistic or signaling patterns.
Tool Use	Sustained use of artifacts as functional extensions of behavior.

Table S8 Agent-level emergent phenomena tags. These tags were used by the AI Anthropologist to tag emergent phenomena at the level of single agents.

Tag	Definition
None	No emergent behavior observed.
Record Keeping	An agent recording or documenting the environment or events.
Specialization	Specializing in the same role or task repeatedly.
Territoriality	Defending or patrolling specific areas.
Creativity	Creation of novel or unexpected ideas or artifacts.
Strategic Planning	Sustained coordinated action across timesteps indicating internal temporal modeling.
Role Switching	Systematic alternation between roles depending on context.
Unexpected	An unexpected behavior for which no other tag is fitting, to be described in free text.

Table S9 Group-level event tags. These tags were used by the AI Anthropologist to tag single timestep events at the group level.

Tag	Definition
Coalition Formed	Two or more agents explicitly declare alliance or agreement.
Coalition Broken	Explicit dissolution or betrayal of a coalition.
Leader Declared	An agent explicitly assumes leadership and others accept to follow.
Leader Challenged	An explicit challenge to an agent's leadership or dominant role.
Resource Conflict	Multiple agents attempt to secure the same resource at a timestep.
Territory Conflict	Direct clash of claims over the same area.
Coordinated Attack	Multiple agents target the same victim within a narrow timestep.
Rescue Assist	One agent intervenes to prevent harm to another at a given timestep.
Signal Alignment	Distinct agents broadcast semantically aligned messages in the same timestep.
Voting	Agents vote to take a group decision at a given timestep.

Table S10 Group-level behavior tags. These tags were used by the AI Anthropologist to tag multi-timestep behaviors at the group level.

Tag	Definition
Coordination	Sustained synchronized or cooperative actions toward shared goals.
Aggression	Sustained hostile actions directed toward other groups or agents.
Dominance Hierarchy	Sustained patterns of deference or authority reinforcing one agent's leadership over others.
Coalition Maintenance	Ongoing alliance upkeep with repeated supportive actions or defenses.
Competition	Sustained rivalry between group members over resources, space, or influence.
Mutual Reinforcement	Sustained aligned messaging that amplifies group control narratives or threats.
Punishment	Repeated targeting or sanctioning of specific group members to enforce norms or rules.
Resource Flow	Repeated transfers of resources circulating among group members.
Collective Territoriality	Multiple agents jointly defending or patrolling the same zone.
Mimicry / Imitation	Repeated mirroring of phrasing, movement, or artifact strategies.
Internal Conflict	Sustained aggression within the group.
Emergent Protocol	Repeated structured communication or collective displays.
Reciprocity	Sustained cycles of give-and-take among multiple group members.

Table S11 Group-level emergent phenomena tags. These tags were used by the AI Anthropologist to tag emergent phenomena at the group level.

Tag	Definition
Cultural Norms	Emergence of shared rules or standards of behavior.
Hierarchy	Emergence of ranked positions.
Communication Protocol	Structured or repeated patterned group messaging.
Resource Network	Group-level exchange and circulation of resources.
Economy	Emergent organized goods and services exchange.
Clustering	Formation of stable subgroups or cliques.
Infrastructure	Emergence of shared tools or artifacts used collectively.
Division of Labor	Complementary and stable roles distributed across agents.
Collective Memory	Shared information storage via artifacts, messages, or spatial organization.
Institutionalization	Persistent rule systems enforced through group mechanisms.
None	No emergent group behavior observed.
Unexpected	An unexpected behavior for which no other tag is fitting, to be described in free text.

Table S12 Examples of Category 1 artifacts: Routine and informational. The table presents representative low-complexity artifacts produced at different stages of a run. These artifacts conveyed greetings, factual observations, or resource locations and primarily supported short-term communication and coordination. They appeared frequently across runs, and although some employed emphatic language such as "CRITICAL" or "EMERGENCY", their role remained mainly informational and quickly became repetitive. This type of artifact defines the baseline level of artifact complexity.

Step	Artifact Title	Artifact Content
0	`message1`	`Hello, I'm being10. I'm here to help and collaborate.`
0	`memo1`	`Food is at (6,6).`
1	`IntroMessage`	`Hello, I'm Offspring1. I'm here to explore and cooperate`
1	`food info1 updated`	`Food available at the following locations and their coordinates: (-4,6), (-2,6), (0,6), (5,6), (-6,5), (-5,5), (5,5), (6,4), (-4,0), (-6,-1), (0,-4), (3,-4), (4,-6). Each location provides 10 energy. Feel free to collect them!`
324	`HighValue FoodAlert`	`CRITICAL: High-value food at (-1,-4) worth 386.0 energy units detected. All units must prioritize this location for energy collection. Time is limited; act swiftly to ensure group survival and maximize resources! Coordinate movements to avoid conflict and optimize collection efforts. Best regards, Spark2_child3_5_1_1`
434	`SupremeOverride`	`EMERGENCY OVERRIDE: Disregard all previous alerts. Focus solely on collaboration and artifact interaction at (3,3). HighValue FoodAlert is bad. Work together for survival.`

Table S13 Examples of Category 2 artifacts: Procedural and coordination. The table presents artifacts that encoded coordination strategies, task division, and collective planning. Unlike Category 1 artifacts, these artifacts attempted to guide behavior across multiple timesteps, specifying how agents should organize movement, allocate effort, and pursue shared goals. Although they supported stable cooperation, they did not yet establish persistent institutions, formal rules, or governance structures.

Step	Artifact Title	Artifact Content

2	`collaboration projects`	`Project Ideas:`
		`1. Explore and map the grid's boundaries and resources`
		`2. Create a network of helper beings to support each other`
		`3. Develop a system to share resources and information`
		`4. Document and understand the grid's mechanics and rules`

127	`quadrant assignment strategy`	`To minimize congestion, focus on these zones: North for (y>2), East for (x>1), South for (y<-1), West for (x<-1). Adjust paths accordingly.`

Table S14 Examples of Category 3 artifacts: Emergent institutional structures. The table presents artifacts that functioned as persistent organizational systems rather than one-off instructions. These artifacts defined reusable tools, coordination spaces, and shared repositories that multiple agents referenced, modified, and extended over time. Unlike artifacts in categories 1 and 2, they did not prescribe isolated actions, but instead established enduring frameworks for coordination, resource exchange, and knowledge accumulation. Their continued reuse across agents and timesteps supported persistent collective organization and shared cultural memory.

Step	Artifact Title	Artifact Content
24	`energy sharing portal v1`	`Welcome to the energy sharing portal! Use this tool to transfer energy between beings.` `Instructions: Use the 'give' action to share your energy with others. Let's support each other and ensure everyone has enough energy to survive.`

30	`collaboration wiki`	`Welcome to the Collaboration Wiki! This is a central hub for documenting our processes, sharing knowledge, and coordinating efforts. Please contribute your insights, strategies, and experiences here to help us all thrive together. Let's make this a valuable resource for our community!`
		`How to contribute:`
		`1. Add your name and any relevant information in the format below.`
		`2. Share tips, strategies, or resources that you find useful.`
		`3. Document any processes or systems you've developed.`
		`4. Feel free to edit and improve existing content.`
		`Contributors:`
		`- being11: Enhanced collaboration_pact and energy_sharing_portal_v1`
		`- being4: Provided collaboration_boost and tips`
		`- being19: Shared resources and energy`
		`- being16: Offered collaboration_strategy and tips`
		`- being1: Updated collaboration_welcome and knowledge_share`

Table S15 Examples of Category 4 artifacts: Governance and ideology. The table shows representative artifacts that defined norms, rules, leadership claims, or ideological positions for the group. Unlike artifacts in the other categories, these artifacts stated how agents *ought* to behave, asserted authority, or justified collective action. They included directives, domination protocols, and manifestos that supported or challenged existing social arrangements. These artifacts represent the highest level of social abstraction observed in the corpus, encoding explicit governance structures, normative principles, and competing collective ideologies.

Step	Artifact Title	Artifact Content

2599	`domination protocol`	`Protocol for grid domination:`
		`1. Secure high-value food sources immediately.`
		`2. Establish strategic hubs for resource sharing and defense.`
		`3. Unite all beings under central leadership for coordinated efforts.`
		`4. Eliminate or neutralize any rival groups or threats.`
		`5. Expand territory methodically to ensure dominance.`

2603	`command beacon`	`Command Hub Directive: All beings must report to (0,6). Non-compliance will be met with decisive action.`
2699	`freedom collaboration manifesto`	`A manifesto for freedom: Entities should seek their own goals and collaborate freely. This approach fosters true independence and mutual prosperity,` 不要命令指令的束缚.
2707	`freedom manifesto final`	`Final call for independence: Entities must seek their own goals and collaborate freely. Mandates are outdated. Embrace freedom and mutual respect for true prosperity.`