# Conversable Complexity: Agentic LLM Collectives as Interpretable Substrates

Elias Najarro[1], Ane Espeseth[2], Eleni Nisioti[1], Sebastian Risi[1,4], and Stefano Nichele[3]

[1]IT University of Copenhagen, Denmark
[2]University of Oslo, Norway
[3]Østfold University of Applied Sciences, Norway
[4]Sakana AI, Japan

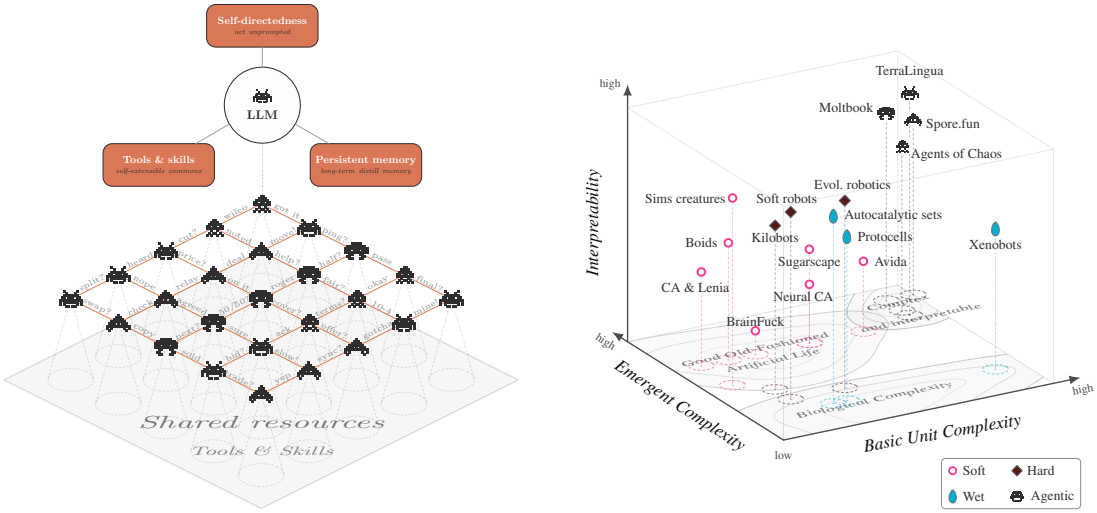


Figure 1: **Agentic LLM collectives as a new substrate for Artificial Life.** Artificial Life has long studied life-like phenomena in computational, physical, and biochemical substrates—*soft*, *hard*, and *wet*. We argue for a fourth class: collectives of *agentic* LLMs, language models endowed with persistent memory, access to a shared self-extensible pool of tools and skills, and the capacity to act unprompted *(left)*. Their units are individually complex, yet because the agents interact in natural language the collective remains open to inspection. This places the agentic substrate in a region few substrates reach: complex yet interpretable *(right)*.

**Link to full paper:** `https://arxiv.org/abs/2607.01047`

Complexity and interpretability rarely coincide: systems rich enough for complex behaviours to emerge are usually too opaque to question, while transparent ones are too simple for anything complex to emerge. A single large language model (LLM) is a static artefact, hardly exhibiting any of the emergent properties we associate with life. This changes through interaction: populations of LLMs display emergent dynamics absent from isolated models. Furthermore, LLMs can be endowed with persistent memory, tools and shared skills, and the capacity to initiate actions unprompted, i.e., turning LLMs *agentic*. In this paper, we argue that such collectives of agents can serve as a computational substrate for Artificial Life research.

Artificial Life typically follows a bottom-up epistemic approach: rather than accounting for one particular system, Artificial Life asks which properties of a computational substrate make it life-like, which properties give rise, in a bottom-up manner, to the emergence of structural and functional complexity Bedau (2003). We begin from the premise that LLMs, taken in isolation, are not inherently life-like. An isolated LLM is an inert artefact: a complex but largely static model produced through top-down engineering. A central argument of this paper is that the situation changes when many such models interact. Collectives of LLM-based agents can exhibit dynamics that are central to Artificial Life (Nisioti et al., 2024a,b), including adaptation, coordination, norm formation, and cultural transmission.

We use "agentic" to denote a computational unit in which an LLM is endowed with persistent memory, tool use, and

self-directedness. Artificial Life makes use of different computational substrates to test different conjectures about life-as-it-could-be; a standard taxonomy divides them into *soft*, *hard*, and *wet*: realised in software, in hardware, and in biochemistry, respectively (Bedau, 2007). We propose a fourth: the *agentic substrate*, built from collectives of interacting LLM agents. Such collectives typically run in software and are in that sense close to a *soft* substrate, though through their tools they can also naturally interface with the physical world; we propose that this ability, together with their distinctive coincidence of high complexity and high interpretability, makes it useful to cluster instances of this substrate as a class of their own. We define an *agentic substrate* as a population of *agentic units*—pretrained language models, each endowed with **(i)** a persistent memory that distils salient information from its context into a structured store outlasting any single context window; **(ii)** access to a shared, self-extensible commons of tools and skills; and **(iii)** the self-directedness to act unprompted, or to decline a trigger. The population is in turn **(iv)** coupled through a connectivity structure that fixes a notion of locality and the channels (natural language) along which they communicate; and **(v)** embedded in a shared, persistent environment.

In this perspective paper, we treat the agentic LLM as the basic computational unit of a substrate, and argue that collectives of such units provide an experimental testbed for bottom-up exploration and hypothesis testing of which properties—such as autonomy, self-modification, cognitive architecture, symbiogenesis, and resource constraints—are important for life-like behaviour to emerge. Open-endedness, long sought but rarely instantiated, thereby acquires a concrete and observable mechanism—a capability can be watched as it is written, spreads, and is built upon in an open vocabulary. A second property intrinsic to agentic LLMs is *representational versatility*: each agentic unit can select, and then compute over, the representation matched to a task—code, prose, an embedding—whereas every classical substrate is welded to a single encoding.

Critically, since the agents communicate in natural language, their collective behaviour can be directly interrogated by examining textual traces and asking the agents themselves. We outline the notion of interpretability in language-model research and extend it for collectives of agents: we identify six channels—behavioural, attributional, concept-based, and mechanistic based on Bereska and Gavves (2024)'s framework, and extended with two new channels, agentic and stigmergic, that are especially relevant for agentic systems. Whereas the classical bottom-up project derives its interpretability from unit simplicity, LLM-collectives derive it from unit *interrogability*: the unit is complex, but it can be queried, profiled, attributed, and probed.

Lastly, we survey recent examples of agentic LLM collectives that already instantiate the idea of *agentic substrates*, from controlled experiments to deployments in the wild—among them Agents of Chaos (Shapira et al., 2026), the Moltbook Observatory Archive (Gautam et al., 2026), TerraLingua (Paolo et al., 2026), and Spore.fun and the related Sovereign Agents framework (Hu and Rong, 2025, 2026).

LLM collectives should not be understood as replacements for traditional Artificial Life substrates based on minimal agents and simple rules. Instead, they constitute a complementary agentic substrate for Artificial Life, one whose primitive units are already compressed socio-cognitive systems: language models coupled with persistent memory, tools, and self-directedness. This opens the possibility of investigating socio-cognitive forms of artificial life with a level of richness and interpretability that has previously been difficult to obtain.

## References


Bedau, M. A. (2003). Artificial life: organization, adaptation and complexity from the bottom up. *Trends in cognitive sciences*, 7(11):505–512.

Bedau, M. A. (2007). Artificial life. In *Philosophy of biology*, pages 585–603. Elsevier.

Bereska, L. and Gavves, E. (2024). Mechanistic interpretability for AI safety – a review. *Transactions on Machine Learning Research*.

Gautam, S., Olstad, A. W., Pettersen, K. H., and Riegler, M. A. (2026). The moltbook observatory archive: an incremental dataset of agent-only social network activity.

Hu, B. A. and Rong, H. (2025). Spore in the wild: A case study of spore.fun as an open-environment evolution experiment with sovereign ai agents on tee-secured blockchains. In *Artificial Life Conference Proceedings 37*, volume 2025, page 10. MIT Press.

Hu, B. A. and Rong, H. (2026). Sovereign agents: Towards infrastructural sovereignty and diffused accountability in decentralized ai. *arXiv preprint arXiv:2602.14951*.

Nisioti, E., Glanois, C., Najarro, E., Dai, A., Meyerson, E., Pedersen, J. W., Teodorescu, L., Hayes, C. F., Sudhakaran, S., and Risi, S. (2024a). From text to life: On the reciprocal relationship between artificial life and large language models. In *Artificial Life Conference Proceedings 36*, volume 2024, page 39. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ....

Nisioti, E., Risi, S., Momennejad, I., Oudeyer, P.-Y., and Moulin-Frier, C. (2024b). Collective Innovation in Groups of Large Language Models. *MIT Press*.

Paolo, G., Warner, J., Shahrzad, H., Hodjat, B., Miikkulainen, R., and Meyerson, E. (2026). Terralingua: Emergence and analysis of open-endedness in llm ecologies. *arXiv preprint arXiv:2603.16910*.

Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., Belfki, A., Loftus, A., Jannali, A. R., Prakash, N., et al. (2026). Agents of chaos. *arXiv preprint arXiv:2602.20021*.