How Word Category and Selection Criteria Drive Evolution in an Evolutionary Ecology of Words

Reiji Suzuki[1] and Takaya Arita[1]

[1]Graduate School of Informatics, Nagoya University, Japan
reiji@nagoya-u.jp

Abstract

We extend an "evolutionary ecology of words" model based on a large language model (LLM), in which each agent moving in a two-dimensional space carries an LLM-generated word as its trait and competitions between words are judged by the LLM. We systematically analyze how word categories and selection criteria shape evolutionary dynamics, sweeping all 110 conditions (10 word categories, 10 selection criteria, and a random-judgment null), with 5 trials each. Using two metrics, the number of persistent species and the semantic leap distance, we find an asymmetric influence: both axes affect both metrics, but the category more strongly drives open-endedness (semantic leap) and the selection criterion more strongly constrains diversity (persistent species). Where the two axes act together, concepts that transcend a category's lexical boundary can become stably established. For example, under a "rare" criterion the plant name amaranthus hypochondriacus colonized the vehicle category, plausibly via the car color "amaranth red." These results illustrate how the creative generative capacity and context-dependent judgment of LLMs can contribute to the realization of open-ended evolution.

Introduction

Recent advances in large language models (LLMs) make it possible to incorporate real-world levels of complexity and diversity into computational models through natural language, with growing applications in artificial life (Nisioti et al., 2024). Research on LLM-based multi-agent systems and generative agent-based models (GABMs) is especially active, combining the generative capabilities of LLMs with collective dynamics to pursue novelty and open-endedness such as the emergence of personality, behavioral, and social traits, and of collective innovation, in LLM agent populations. Harnessing the creative potential of LLMs for open-ended evolution (OEE) (Taylor, 2019) requires collective and evolutionary frameworks that go beyond prompting.

Two complementary roles for LLMs in such evolutionary loops have emerged. On the generative aspect, LLMs act as semantic variation operators that propose informed mutations rather than blind perturbations, as in Evolution through Large Models (ELM) (Lehman et al., 2022). On the evaluative aspect, the LLM-as-a-judge paradigm (Gu et al., 2025) uses an LLM as a context-dependent evaluator. The model studied here exercises both roles within a single evolving population.

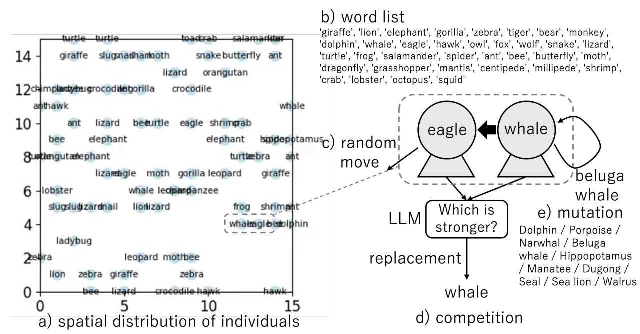

Figure 1: Overview of the model. (a) Spatial distribution of individuals in a 2D space, (b) word list generation for the initial population, (c) random movement, (d) LLM-based competition, and (e) mutation.

Building on this, Suzuki and Arita (2025) proposed an "evolutionary ecology of words" model that utilizes the lexical and contextual knowledge of LLMs as a source of complexity rather than imitating biological and cultural evolution. Agents on a two-dimensional grid carry words as genes; when neighbors meet, the LLM judges which word "is stronger" or better satisfies some criterion, the loser's word is overwritten by the winner's, and with some probability the LLM mutates a word into a similar but different one. Outcomes thus rest on the semantic and cultural knowledge embedded in the LLM, enabling biologically fictitious yet conceptually coherent pressures.

In this model the word category and the selection criterion are two essential axes that respectively define the structure of the possibility space and the nature of the selection pressure, corresponding to the generative and evaluative roles above. However, the preliminary experiments examined only a single condition, leaving the effects of each axis and their interaction unexplored. Here we extend the model and conduct a systematic sweep over all 110 conditions (10 word categories × 10 selection criteria, plus a random-judgment null). Using the number of persistent species, representing the richness of words that achieve substantial cumulative presence in the population, and the semantic leap distance, representing the magnitude of departure from the initial concept space, we discuss the effects of the two axes and their synergistic effects.

Model

Figure 1 shows an overview; see Suzuki and Arita (2025) for details of the original model, whose prompts and procedures we modified to make the category and criterion configurable and to speed up the exhaustive experiments. N individuals occupy a $W \times W$ toroidal grid, each carrying a word (a single word or short phrase) expressed in English as its gene. The base system prompt ("You are a versatile AI with deep language and comparative analysis skills, adept at generating diverse word and phrase lists and highlighting nuanced differences between terms") is provided, and task-specific prompts are appended for each operation.

The LLM is asked to "create a non-overlapping list of A existing #TARGET_wordclass#" (the word category), separated by slashes. Each of the N individuals is assigned a word drawn uniformly at random (with replacement) and placed at a random grid cell.

Three processes run in random agent order. Movement: each individual attempts to step into a random cell of its 8-neighborhood (staying put if occupied), producing viscosity and spatial clusters. Competition: each individual picks a random neighbor and queries the LLM, "Which one is a more #TARGET_criteria# #TARGET_wordclass#: '#word1#' or '#word2#'?"; a matching response wins and overwrites the loser's word, otherwise nothing changes. Mutation: with probability p_m, the LLM lists 12 in-category words "similar to, but slightly or significantly different from" the current word, and one is adopted at random. For efficiency, competition pairs are sorted to a canonical order, deduplicated, and batched, and both competition outcomes and mutation candidate lists are cached and reused within a trial.

Experimental Setup

For text generation we used an AWQ-quantized Gemma-3-12B-it (gaunernst/gemma-3-12b-it-int4-awq) (Gemma Team, 2025) served by a vLLM batch-inference engine, collecting all competition judgments and mutation requests within a step and sending them at once to make the large-scale search (550 trials) tractable.

We defined 10 word categories (animals, foods, vehicles, colors, occupations, sports, musical instruments, plants, emotions, natural phenomena) and 10 selection criteria (strong, cute, large, fast, famous, beautiful, rare, useful, intelligent, ancient), plus a random-outcome null. For each of the 110 conditions we ran 5 trials. Parameters: $W = 20$, $N = 100$, $p_m = 0.075$, $A = 20$, 300 steps. In the null condition, LLM judgments are replaced by equiprobable random decisions, isolating the contribution of each category's lexical space; as with LLM judgments, random outcomes are cached and reused, so the null imposes a consistent, though randomly assigned, selection pressure.

A word is a persistent species if the cumulative count of individuals carrying it reaches 150 or more, i.e., a word that achieved substantial presence during the trial, whether through wide spread or sustained persistence. The semantic leap distance is the maximum cosine distance from any persistent species to the mean embedding of the initial population, computed with a multilingual sentence model (paraphrase-multilingual-MiniLM-L12-v2); the species attaining it is the innovator word. Based on 5 trials per condition, this is a preliminary overview of how the two axes shape evolution, relying on the consistency of trends across conditions.

Results

The system showed qualitatively different regimes (Figure 2). In intelligent × natural_phenomena, compound concepts absent from the initial population (e.g., self-organization and emergent behavior) emerged and stabilized alongside a churning coexistence of words such as entropy, chaos theory, and phase transitions, giving sustained diversity with innovative leaps. In famous × animals, by contrast, a single word ("horse") rapidly dominated while others were nearly eliminated. The two metrics below capture this contrast quantitatively.

Plotting persistent species against semantic leap for all 110 conditions (Figure 3), category centroids disperse primarily along the horizontal (semantic-leap) axis, whereas selection-criterion centroids spread along the vertical (persistent-species) axis. Although both axes affect both metrics at the level of individual conditions, their aggregate effects differ: categories primarily govern the extent of conceptual exploration (semantic leap), while selection criteria primarily constrain diversity (the number of persistent species). The null condition (NUL) shows relatively high persistent-species counts, indicating that mutation-driven lexical exploration alone generates some diversity because of the random but consistent selection pressure. Several LLM-based conditions exceed the null in semantic leap dis-

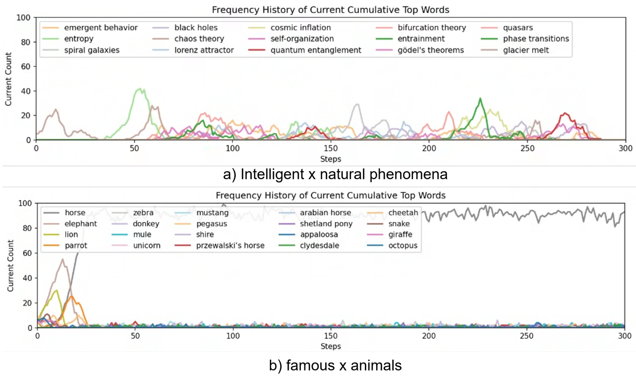

Figure 2: Frequency time series of the top words for two representative conditions: (a) intelligent × natural_phenomena (high leap, diverse), (b) famous × animals (low leap, dominant). Each line traces one word's population count over time.

tance, confirming that the LLM's semantic judgment plays an important role in promoting innovative establishment.

Regarding the number of persistent species, our measure of sustained diversity, both axes have an effect, but the selection criterion is the stronger driver. Context-dependent criteria such as "useful" and "rare" are judged differently from one matchup to the next, so no single word can dominate and many words coexist, giving high counts. Criteria backed by a stable, broadly shared ordering, such as "famous" and "large," instead drive the population to converge on a few consensus winners, giving the lowest counts. The word category also matters, with some domains (e.g., colors and occupations) tending toward higher counts, yet changing the criterion within a fixed category shifts the count more than changing the category within a fixed criterion.

Regarding the semantic leap, our measure of open-endedness, the pattern is symmetric: both axes again have an effect, but here the word category is the stronger driver (Figure 4). Vehicles and natural phenomena produce high leaps under any criterion, because their vast, hierarchical concept spaces let mutation keep generating novel expressions far from the initial population, whereas emotions and colors stay close to the initial concept space, their tighter conceptual boundaries constraining departure. The selection criterion also shifts the leap ("ancient" and "rare" higher, "famous" and "strong" lower), yet changing the category within a fixed criterion moves the leap more than changing the criterion within a fixed category.

Beyond these aggregate trends, the innovator words in Figure 4, namely the most distant persistent species in each condition, reveal how high semantic leaps arise,

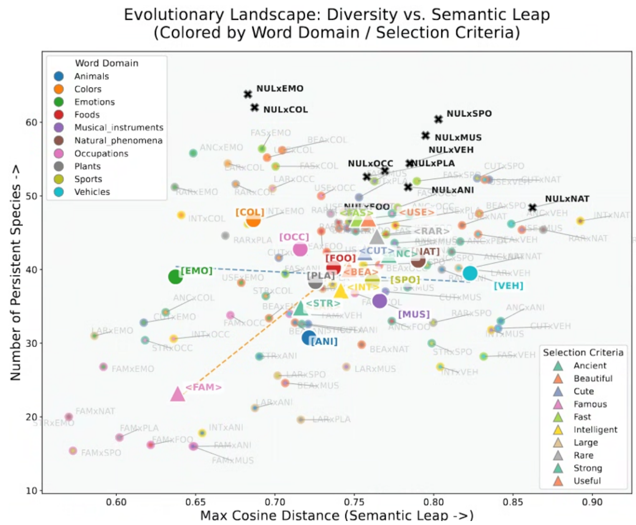

Figure 3: Number of persistent species vs. semantic leap distance for each condition. Large circles: category centroids (square brackets); triangles: selection-criterion centroids (angle brackets); crosses: null conditions. Small circles are individual conditions colored by word domain.

and several conditions where the two axes act together stand out. In rare × vehicles, "rare" combined with the expansive vehicle space to establish amaranthus hypochondriacus (0.935). Although a plant name beyond the vehicle category, the mutation path likely passed through the car color amaranth red, so the LLM's cross-domain associations mediated the leap. In ancient × vehicles, the innovator lynx spider (0.849) entered from the animal domain via the vintage car Lynx XJ-S Spider. In intelligent × natural_phenomena, "intelligence" steered the population toward self-organizing systems (0.862), bridging natural phenomena and complex systems science, and in useful × natural_phenomena, "usefulness" produced biogeochemical cycling (0.818), at the intersection of earth science and environmental engineering. These innovator words show that a selection criterion can dynamically extend a category's lexical boundaries, so the two axes interact synergistically rather than independently, and that the largest semantic leaps coincide with this cross-domain establishment.

Conclusion

Across all 110 conditions of an LLM-based evolutionary ecology of words, the word category and the selection criterion influence the two metrics asymmetrically: the lexical openness of categories drives open-endedness (semantic leap), while the context-dependence of selection criteria shapes diversity (persistent species), and the

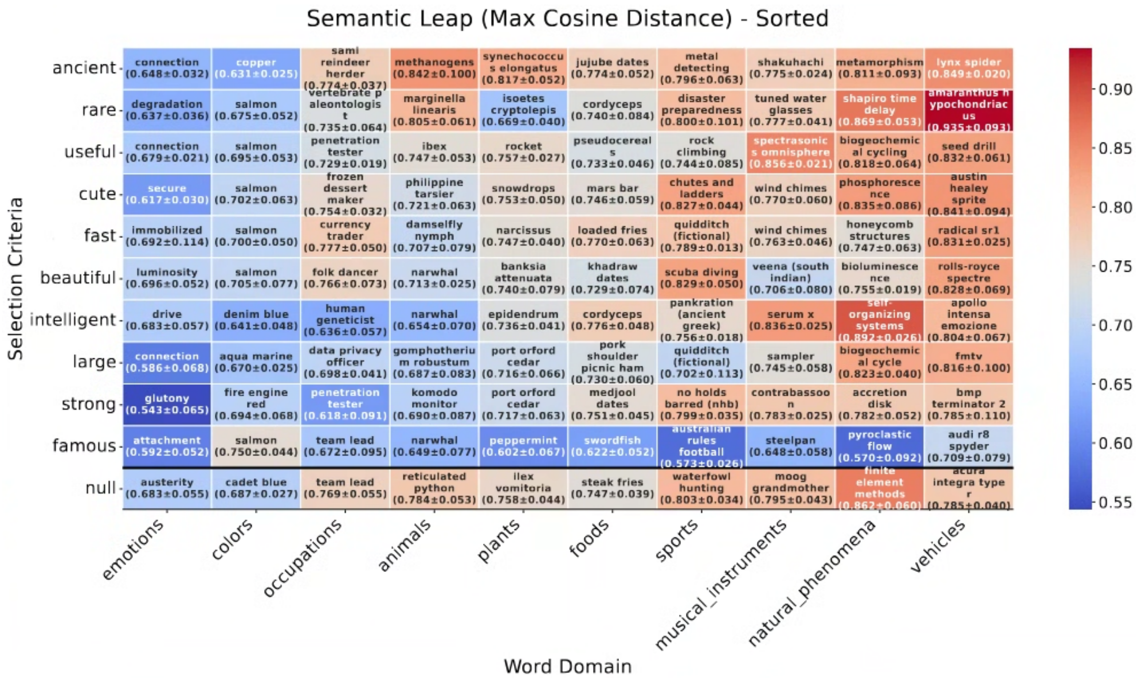

Figure 4: Semantic leap distance (mean ± SD) and innovator words for each condition: cosine distance from the initial mean vector to the most distant persistent species. Axes independently sorted by mean. The innovator word shown is from the trial with the largest leap among the 5 trials.

two can interact to establish concepts beyond the initial lexical boundaries. Despite the influence of the particular model and the limited number of trials, the core point already holds: the category and the criterion are two distinct design axes, each able to shape the diversity and evolutionary tendency of the population, consistent with the balance of exploration and convergence in open-ended evolution (Taylor, 2019). This asymmetry may, however, be specific to the small population and short horizon studied here, since Suzuki and Arita (2025) report changing dynamics at larger populations and longer time scales even in the stronger-animal condition. What to generate and how to evaluate it are fundamental choices in any LLM-based creative process; testing this across other models and substrates is an important next step.

Acknowledgments

This study is supported in part by JSPS Topic-Setting Program to Advance Cutting-Edge Humanities and Social Sciences Research JPJS00122674991, JSPS KAKENHI JP24K15103, Kayamori Foundation of Information Science Advancement, and Google Gemma 2 Academic Program.

References

Gemma Team (2025). Gemma 3 technical report. arXiv preprint arXiv:2503.19786.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., and Guo, J. (2025). A survey on LLM-as-a-judge. arXiv:2411.15594.

Lehman, J., Gordon, J., Jain, S., Ndousse, K., Yeh, C., and Stanley, K. O. (2022). Evolution through large models. arXiv:2206.08896.

Nisioti, E., Glanois, C., Najarro, E., Dai, A., Meyerson, E., Pedersen, J. W., Teodorescu, L., Hayes, C. F., Sudhakaran, S., and Risi, S. (2024). From text to life: On the reciprocal relationship between artificial life and large language models. In Proceedings of the 2024 Conference on Artificial Life, page 39.

Suzuki, R. and Arita, T. (2025). Evolutionary ecology of words. In Proceedings of 2025 IEEE Symposium on Computational Intelligence in Artificial Life and Cooperative Intelligent Systems (ALIFE-CIS). DOI: 10.1109/ALIFE-CIS64968.2025.10979831.

Taylor, T. (2019). Evolutionary innovations and where to find them: Routes to open-ended evolution in natural and artificial systems. Artificial Life, 25(2):207–224.