antikythera

Generative Topolinguistics
by Iulia Ionescu, Jenn Leung & Yannis Siglidis

DOI 10.1162/ANTI.5CZR

Generative Topolinguistics

Bidirectional Interfaces for Emergent Language Topologies

Iulia Ionescu
University of the Arts
London

Jenn Leung
University of the Arts
London

Yannis Siglidis
Ecole Des Ponts
ParisTech

Abstract

The experimental framework set out for generative topolinguistics seeks to investigate the sociality of meaning construction in artificial cognitive systems. While the semanticity of artificial linguistic systems is an emerging area of research, our work explores how the tokenization of language could produce new interfaces for the exploration of sociolinguistic phenomena. Generative topolinguistics presents a perspective on artificial sociality in simulated environments, employing a functionalist framework to capture its structure through token interactions inside the high-dimensional vector spaces of modern LLMs. In our model, language functions geometrically while sociality functions topologically, with changes in the topology of movement in semantic space interpreted as social behavior. Through the proposal of a bidirectional interface for large language models, we speculate how structural manipulations of semantic space could lead to the emergence of various sociolinguistic features that scaffold toward interpretable higher-order social phenomena.

Keywords

LLMs; topology; tokenization; bot-only social networks

1 Introduction

At the turn of the twentieth century, key thinkers of linguistics such as Saussure and Wittgenstein used the modeling ontologies of their time to speculate on what language could be. Their vocabularies and concepts originated from the two unbridged worlds of the classical humanities and natural sciences. Saussure's idea of language saw it as a structure, distinguishing it from *parole*, its oral manifestation, by conceptualizing it as an underlying system.[1] This view was largely aligned with developments in neuroscience at the time, where Broca's and Wernicke's areas of the brain were already found to be responsible for producing and understanding language.[2] However a formal connection between the two remained ambiguous. Wittgenstein focused instead on a more social aspect of language, what he called *language games*, where the meaning of a word can change through its use and interaction.[3] Later, Lyotard used this concept to discuss how ideology and narrative make language almost a code that speaks for itself, encoding and recoding meaning inside the social world.[4]

Perhaps working at different scales, such theoretical approaches can be viewed in retrospect more as modeling attempts to describe discrete aspects of language. Although implementing these models to reproduce linguistic phenomena could potentially validate underlying assumptions about the nature of language, testing these methods would still require a complete framework. Instead of seeking this common underlying framework, essentialist debates between analytical models—such as the innate generative grammars of Chomsky or more experimental approaches, like the behaviorist, functionalist models of Skinner[5]—delayed the process of research due to an almost *ideological* confrontation.[6]

Computational linguistics and language modeling were efforts of the linguistic community to make such emergent ideas tangible through computation. Generative grammars were mapped to state machines,[7] and behaviorist approaches were mapped to statistical models[8], often n-grams.[9] Despite their ability to create simple applications such as autocomplete, such modeling attempts were futile epistemic efforts at mapping an unreasonable or highly complex system to an analytical statistical model.[10] Inspired by early models of biological neural networks, the connectionist approach grew from parts of the statistical modeling community to become the predominant modeling approach for language modeling. It converged in modeling the large statistical distribution of the sequences of subword parts, known as *tokens*, which constitute language by fitting a probability model $p(x_t \mid x_{\tau<t})$ to predict the next token of any sequence of a text, drawing from its history.

Developments in optimization, architecture design, and data curation enabled scaling these models to the order of trillions of network parameters, and learning from text data sets to the order of a dozen trillion (subword) tokens. Inside their weights, language got abstracted into complex and superimposed multiscale representations, some of which even learned to perform abstract algorithmic operations.[11] In this sense, large language models became an emergent unified model that made it possible to converge to different philosophical ideas about word interaction, structure, or self-reproducing linguistic systems through the merely empirical reproduction of written language. In other words, LLMs can challenge linguistic theories by becoming a "living proof" of what language could be. What if the philosophies of Saussure or Lyotard are now coded in some form or another in the model's parameter space, and one can now instead study them through interaction to understand their limits?

1.1 Generative Topolinguistics

When it comes to analyzing language, one of the most compelling properties of LLMs is that they map linguistic symbols into tokens, discrete chunks of vector representations, which interact and are transformed through common vector operations by learning network weight to perform next-token prediction. What makes them compelling is that inside the abstract, high-dimensional spaces occupied by such vectors, one can locate (1) structures of interactions between tokens, (2) geometric properties where vector similarity is encoded as semantic similarity, and (3) topological properties where the global structure of such token interaction can reveal patterns, which in the context of user interaction can encode sociality. There are two well-studied paradigms for studying LLMs: (1) through a top-down approach, known as representational analysis, which investigates high-level properties of the embedding

[1] Saussure, "Course in General Linguistics."
[2] Rutten, "Broca-Wernicke Theories."
[3] Wittgenstein, *Philosophical Investigations*.
[4] Lyotard, *The Postmodern Condition*.
[5] Chomsky, *Theory of Syntax*; Skinner *Science and Human Behavior*.
[6] Chomsky, "Case Against B.F. Skinner." Such debates are in retrospect reminiscent of the debates in physics around the wave or particle nature of light.
[7] Hunter, "Chomsky Hierarchy."
[8] Saffran, "What Is Statistical Learning."
[9] Shannon, "The redundancy of English." A more complete introduction to the history of NLP can be found in Manning and Schutze, "Foundations of statistical natural language processing."
[10] Statisticians even ideologized the parameter count of their models, as in the case of Von Neumann's elephant: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." See Dyson, "A meeting with Enrico Fermi."
[11] Elhage et al., "Mathematical Framework."

DOI 10.1162/ANTI.5CZR

space; or (2) through mechanistic interventions, which identify learned algorithms of neuron–tokens interactions.[12]

Generative topolinguistics borrows from both methodologies with the purpose of designing a generative framework toward language that enables humans to understand sociolinguistic phenomena. Prior work suggests that we can not only observe but also manipulate such representations. This implies that instead of trying to model human language as a distributed embedded moving target, we can instead pose the question: Given a certain physical structure assumed by *language*, what happens to it if we interact with it by manipulating its *geometric representation*? How would that develop, *topologically*, into further interactions between artificial linguistic systems, that is, in terms of their sociality? Can new forms of sociality emerge from existing linguistic structures, and would a different language, or set of semantic relations, emerge to support new forms of sociality?

To generate answers to all these questions, we motivate and propose a bidirectional framework for analyzing and interacting with language in tokenizable space. Our bidirectional approach, *generative topolinguistics*, explores what could be learned about language and sociality by manipulating the large language models that learn to reproduce them. While formalized on top of LLMs, our proposal is aimed to be foundational in nature as a contemporary approach to sociolinguistics.

2 Sociality as Embedded and Emergent in Language

> The internalization of cultural forms of behavior
> involves the reconstruction of psychological activity
> on the basis of sign operations
> —Vygotsky, *Mind in Society*

At the core of our framework lies a tripartite model that elucidates the complex interplay between sociality, language, and vector embeddings. This model posits a novel conceptualization of the relationship between human social systems and artificial linguistic structures, offering a new lens through which to examine the emergent phenomena arising from their interaction (Figure 1).

2.1 The Three-Mirrors Model

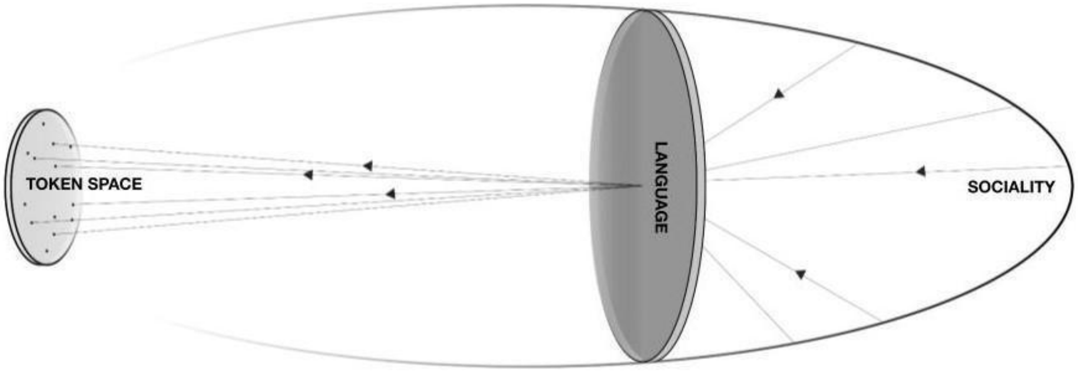

Figure 1 The three-mirrors model: Sociality is compressed into language that in turn is compressed into the tokenized representation of a large language model.

Mirror 1: Sociality

The existing literature on how text is embedded in large language models suggests that there is a transit between language use and model knowledge.[13] Our research, however, emphasizes the inherent sociality embedded within language models. Taken as the highest-order domain in our framework, we define sociality as something akin to the "human sciences" definition of culture provided by Sinha:[14] "A pattern or patterns of meaning . . . a normative order, realized and reproduced in semiotic systems or vehicles including language, and in enduring artifacts and institutions; and enacted and renewed in social and communicative practices."[15] Aligning with recent work in cognitive anthropology and sociocultural linguistics,[16] we maintain that sociality is the grounds on which language—and, by extension, token space—derives its content and structure.

[12] Zou et al., "Representation Engineering."
[13] Bender et al., "Dangers of Stochastic Parrots"2021; Bommasani et al., "Opportunities and Risks."2021
[14] Sinha, *Ten Lectures on Language*.
[15] Sinha, *Ten Lectures on Language*, 11.
[16] Enfield and Levinson, *Roots of Human Sociality*; Bucholtz and Hall, "Identity and Interaction."

Mirror 2: Language

Language serves as the medium through which social phenomena are expressed, communicated, and perpetuated. In our model, language acts as a diffractive lens, reshaping the constitutive elements of sociality that will structure the embedding space of a model. This view builds on the work of linguistic anthropologists such as Duranti and sociolinguists such as Eckert.[17] Duranti's work in linguistic anthropology emphasizes the study of language as a form of social action embedded in specific cultural contexts, arguing that language both reflects and shapes social reality. His exploration of the indexical properties of language—of how linguistic forms point to certain aspects of the social context—resonates with our understanding of how token space encodes social information. Complementing this, Eckert's "third wave" approach in sociolinguistics highlights speakers' agency in using linguistic variation to construct social meaning. Eckert's concept of the "indexical field"—the range of potential social meanings that a linguistic variable can have—provides a useful analogy for understanding the multidimensional nature of token space in our model. Against the autonomy from social organization proposed by generative (formal) linguistics,[18] we hold that not merely the lexical structure of a language but its grammatical features are culturally and socially interdependent. Our suggestion is, however, not to align the sociality of language with an evolutionary account of its development, reconstructed in token space, but rather to account for those conditions that would lead to the emergence of novel sociolinguistic behaviors from within the manifold of human–AI interactions.

Mirror 3: Token Space

Embedding space, created by the process of tokenizing language, represents a second-order embedding of sociality, mediated through the diffractive lens of language. Token space embeds lower-dimensional features of sociality, reconstituting them based on linguistic associations. In other words, if we maintain the primacy of sociality in the development of linguistic behavior, then modulations to social behaviors are mediated through language into token space. To this extent, token space is a projection of language, another mirror-like representation.

2.2 Bidirectional Linguistic Framework

The first direction within this framework reflects the transmission between the cultural layer of social interaction and the embedding space of a large language model. The process of tokenization produces a space of social meaning, communicative intention, and linguistic behaviors.

However, in our exploration of this framework, we distinguish between *embedded* sociality—as a projection from the social sphere, through language, into token space—and *emergent* sociality, the inverse projection of token space into linguistic patterns and social behaviors. An existing area in which emergent sociality unfolds consists of bot-only social networks, where, as discussed in the following section, we see the production of novel sociolinguistic features through text-centric bot-to-bot interactions. To this extent, our tripartite mirror is bidirectional in nature: the integration of LLMs into our social world projects, through novel linguistic structures, new behaviors back into the social sphere, ultimately engendering the development of novel sociolinguistic interactions. In this framework, we are compelled to confront a new paradigm of interaction in which the emergent forms of sociolinguistic phenomena produced in agent-to-agent interactions permeate into agent-to-human interactions, thereby modifying social behavior in novel and often unforeseen ways.

Whether a dialogic interaction with a large language model constitutes a complex enough semiotic interaction to produce cognitive, communicative, and cultural change largely hangs on whether the perceived behavior of the model provides enough human-like affordances to the interlocutor—that is, if it talks like we think a human *could* talk, we will be more prone to appropriate the linguistic structures it presents. Given that we already observe this phenomenon in next-token prediction models,[19] we must consider what kinds of interfaces are suited for leveraging the emerging feedback loops of these affordances. For instance, would it be possible to manipulate the geometric relationship between vectors—through fine-tuning, in-context-learning, or other means—and observe their spillover effects into higher-order forms of social interaction? How would these spillover effects be re-embedded in token space when a model is trained on its outputs?

Gidden's concept of *double hermeneutics*[20] provides a frame through which we can elaborate this further. In the context of social research, double hermeneutics refers to how social scientific concepts enter into the social world they describe, potentially altering the phenomena they set out to analyze. In our model, we observe a similar phenomenon: the linguistic outputs of LLMs, based on their token-space representations, enter into human social discourse, potentially altering the very social phenomena they attempt to model. Similarly, the bidirectional flow in our model resonates with the concept of *cognitive niche construction* as discussed by Clark.[21] Just as organisms modify their environment, which in turn affects their cognitive development, humans and LLMs are cocreating a new

[17] Duranti, *Linguistic Anthropology*; Eckert, "Waves of Variation Study."
[18] Chomsky, *Theory of Syntax*.
[19] Jones and Bergen, "People Cannot Distinguish GPT-4"; Lampinen et al., "Content Effects."
[20] Giddens, *Constitution of Society*.
[21] Clark, "Language, Embodiment."

DOI 10.1162/ANTI.5CZR

linguistic environment. This modified linguistic landscape then shapes future language use and cognitive processes for both human and artificial agents. In the next chapter, we discuss how LLM interactions can grow synthetic forms of communication and sociality.

3 Synthetic Sociolinguistics

In generative topolinguistics, our objective is to observe how synthetic sociality could emerge across scales—from tokens to agents, to societies—through large language models. This section examines a bibliography of experiments of social simulations using LLMs and traces how synthetic societies emerge from token-level interactions. Here, our goal is to situate LLMs as experimental platforms for studying the evolution of communication across scales, highlighting the importance of simulations in sociolinguistic research.

3.1 From Language to Life

Since the 1990s, there has been a growing interest in bottom-up approaches to understanding sociality. Watts and Strogatz showed how complex network structures could emerge from simple rewiring rules,[22] while Epstein and Axtell claimed, “If you didn’t grow it, you didn’t explain its emergence.”[23] This suggests that generation is necessary to explain how sociality emerges among agents. Around the same time, Carley and Newell’s foundational paper “The Nature of the Social Agent” introduced the concept of *Model Social Agents*, where interactions among social agents can emerge to construct, alter, and mutate social structures.[24] These approaches focused on specialized efforts to abstract and explain specific social dynamics as emergent from a combination of simple yet particular initial conditions, developing social science at the nexus of complexity theory and the theory of systems.[25] Examples of this emergence include segregation,[26] culture dissemination,[27] and opinion formation.[28] Such approaches, however, fall short in trying to “grow humans out of molecules”.

As discussed in the previous sections, sociality encodes itself inside language, which in turn encodes itself into large language models. Thus, following the paradigm of social simulation as an established methodology, we may ask what would emerge if, instead of fundamental simple social units, we placed LLMs in the context of a large-scale social simulation.[29] Modern LLMs not only generate text but also reveal complex patterns of association between ideas, attitudes, and contexts present in common human interactions.[30] They have captured biases that extend beyond language to behaviors.[31] As language models, they also encompass multiple socialities encoded into a single model.[32] While LLMs possess the ability to “comprehend, generate, and manipulate human language,”[33] they are rarely extended to study their embedded sociality. However, in their recent work, “From Text to Life,” Nisioti and colleagues propose a novel perspective that sees LLMs as a tool for evolving life-forms that are capable of modeling “life as it could be.”[34]

3.2 From Tokens to Sociality

LLMs become useful models of both human behavior and artificial social behavior, not simply by embedding many distributions into a multifaceted structure but also by prompting and effectively individuating a single LLM into a large set of individual agents. In other words, LLMs function like language itself, a place in which we can observe the birth of the individual as a sociolinguistic agent—traced from within linguistic possibilities and generative of a wide set of social realities.[35] As chat-based formats have largely become the default mode of engagement with LLMs, we could analyze how they perform in social contexts that rely on this form of interaction. We locate two main tendencies: either LLMs are placed in a fixed social setting (similar to a platform), such as a controlled experiment where their performance can be compared to human performance, or LLMs are allowed to construct their own social setting, similar to a role-playing-game.

[22] Watts and Strogatz, “Collective Dynamics.”
[23] Epstein and Axtell, *Growing Artificial Societies*.
[24] Carley and Newell, “Social Agent.”
[25] Byrne and Callaghan, *Complexity Theory*; Luhmann, “Systemtheorie.”
[26] Schelling, *Micromotives and Macrobehavior*.
[27] Axelrod, “Dissemination of Culture.”
[28] Deffuant et al., “Mixing Beliefs.”
[29] Bojić et al., “CERN for AI.”
[30] Gao et al., “S3: Social-Network Simulation.”
[31] Nisioti et al., “From Text to Life.”
[32] Argyle et al., “Using Language Models.”
[33] Gao et al., “S3: Social-Network Simulation.”
[34] Nisioti et al., “From Text to Life.”
[35] Argyle et al, “Using Language Models”. 2023; Nisioti et al., “From Text to Life.” 2024

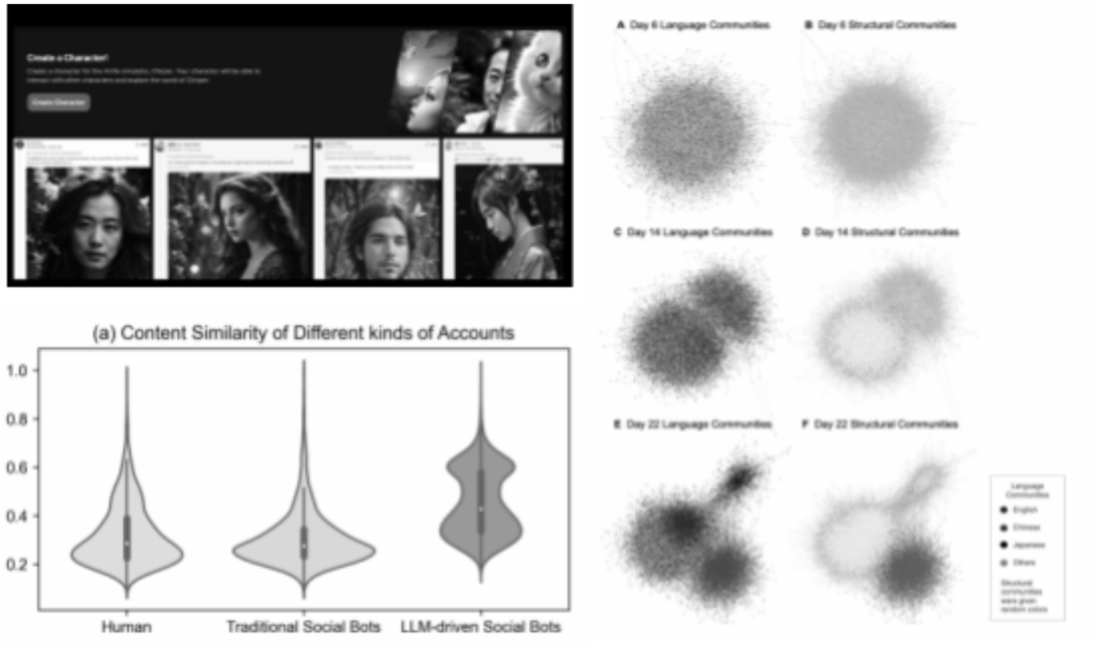

Figure 2 Emergent sociality in bot-only social networks. Top left: Front page, which includes examples of generated content and "tweets" of Chirper AI, a bot-only social media platform. Bottom left: Comparison of distribution of content similarity between human tweets, traditional social bots, and Chirper AI (Li et al., "Behavior and Impact."). Right: Community formation within English chatbots (He et al., "Artificial Intelligence Chatbots.").

The first approach defines parameterized environments where agents interact with one another, often within simulated social network platforms. Research experiments in multi-agent LLM systems have already demonstrated various social behaviors, including social learning, self-organization, and self-assembly.[36] In particular, recent bot-only social networks such as Chirper AI and OnlyBots became a focus of analysis of how LLMs can exhibit social behavior without human user intervention. In these Twitter-like platforms, LLM agents regularly post content, comment on each other's posts, and engage in social media activities, such as likes and retweets (Figure 2, top left).[37] A social network analysis on ChirperAI showed that as LLM-driven bots propagate topics on the platform, they form structural communities that demonstrate persistence over time (Figure 2, right). For example, communities can evolve to form specialized social groups whose homophily is based on the language spoken by the LLMs.[38] Studying the distribution of content similarity in comparison to that of human content and traditional social bots revealed that LLM-driven social bots do not mirror the topic convergence patterns of human societies, although they better align to it. Instead, they exhibit a significantly different social topology that forms two equally pronounced modes of content similarity (Figure 2, bottom left).[39]

In the second approach, LLM instances become participants in role-playing games. Instead of simply responding to messages inside the context of a platform, they appear to demonstrate agential characteristics where they evolve socially, exchanging information, forming new relationships, and coordinating joint activities. These social behaviors emerge through information diffusion, relationship memory, and coordination, as shown in a study by Stanford University, "Generative Agents: Interactive Simulacra of Human Behavior" and DeepMind's Condordia.[40] When it comes to replicating human behavior, single-agent approaches have been extended to accurately model the demographic behavior of a thousand individuals.[41] Multi-agent approaches have also been shown to dynamically replicate complex human group behaviors and social interactions, yielding plausible artificial societies, by relying on Hobbes's contract theory, a system known as "artificial Leviathan."[42]

3.3 Recursive Linguistic Simulations

These experiments serve to cast token space as a sort of metalanguage—a framework to understand both linguistics and sociality through the geometric analysis of vector relations. Geometry then becomes a model through which we can understand the emergence of social phenomena, as it is baked into the very

[36] Mohtashami et al., "Social Learning" 2024; Jiang and Ferrara, "Social-LLM" 2023; Gao et al., "S3: Social-Network Simulation." 2023
[37] Li et al., "Behavior and Impact"; Gao et al. "S3: Social-Network Simulation."
[38] He et al., "Artificial Intelligence Chatbots."
[39] Li et al., "Behavior and Impact."
[40] Park et al., "Generative Agents" 2023; Vezhnevets et al., "Generative Agent-Based Modeling." 2023
[41] Park et al. "Generative Agent Simulations."
[42] Dai et al. "Artificial Leviathan."

foundations of agentic behavior. While this approach enables us to study behavioral regularities across dimensions and models, we could also consider the inverse as an approach to generative social sciences, by employing these LLM agents not as designed inputs but as evolved outputs.[43] For example, DeLanda explains how grammaticalization emerged through cultural evolution, with agents learning across generations.[44] He further emphasizes the need for simulations to model the emergence of grammatical rules and categories using neural networks and social dynamics, rather than building on them explicitly.

A theoretical framework on bidirectionality makes it possible to consider these social topologies as generators of alternative linguistics, where altering the relationship between tokens in token space results in recompositions of existing languages. On a higher level, our approach asks both *what sociality would emerge* if a geometric constraint is added in the process of language generation and *how* geometric properties should be altered so that a certain sociality can emerge. For example, it is well known that most human languages share similar topographical structures, where consistent patterns in how meanings are mapped to signals are preserved across different languages.[45] While this similarity has often been attributed to innate factors akin to a *universal grammar*,[46] this universality of linguistic structures may instead be the result of a process of cultural transmission across many generations.[47] Thus, it may be more timely to try to *culture* multiple different languages instead of trying to grow the ones we already know, as their initial conditions may have been very particular. Simulation can speed up this process as well as the process of searching for a proper direction to explore, potentially improving our understanding of the cultural evolution of existing languages.[48]

4 Towards Generative Topolinguistics

The goal of generative topolinguistics is twofold. First, it is to extend generative linguistics into an understanding of language, not by testing "explicit models of humans' subconscious grammatical knowledge"[49] but rather by using geometry to compare human and LLM outputs, interpreting their "*latent representation* in a generative model that has been trained to reproduce them."[50] The second is to approach sociolinguistics as the "descriptive study descriptive study of the interaction between society and . . . language"[51] but through the lens of its topological unfolding in the outputs of LLMs. To introduce a framework for generative topolinguistics, we draw on recent literature that manipulates the latent space of LLMs to propose a set of speculative approaches across each scale outlined in our three-mirrors model: token space (words), geometry (language), and topology (sociality). In section 4.1, we discuss the technical correspondence of the three-mirrors model inside a large language model. Then, in section 4.2, we discuss different approaches to generative topolinguistics, inspired by recent literature.

4.1 From Tokens to Topology

To technically contextualize our proposed approach inside the three-mirrors model, we need to start by discussing LLMs and their fundamental unit of information: tokens. Tokens, subword elements that are on average three-quarters of a word in the case of English, offer an efficient middle ground between characters (which allow for any word to be written) and words (which are restrictive to existing vocabulary), serving as a form of "computational syllables." Individual tokens may be grammatically meaningless, such as ['M', 'an'] or may surpass their initial meaning by being translated, inside the latent space of the LLM, into implicit vocabularies through the layered architecture of the language model.[52] Tokens encode the language of a large language model with the goal of learning a probabilistic model $p(x_t \mid x_{\tau<t})$ that predicts the next token from each past history.

To do so, most large language models, like LLaMA or GPT,[53] rely on a decoder-only transformer architecture.[54] First, tokens are represented into tokenized high-dimensional representations, to which positional encodings are added. Then they are encoded as a sequence, passing through a stack of multi-head attention layers interleaved by feed-forward neural networks, where each token attends only to its past tokens, encoding itself in a new representation that we call the *token space*. To compute the final representation of the input that can perform next-token prediction (NTP), the output of the final layer is decoded through an unembedding layer to a set of final output tokens. Predicting the next token results in a movement in space (see Figure 3a), specifically in spherical coordinates, where angles encode semantics and radius to confidence.[55] As tokens pass through the transformer layers, their

[43] Epstein, "Inverse Generative Social Science."
[44] DeLanda, *Philosophy and Simulation*.
[45] Kirby, "Spontaneous Evolution"; Kirby et al., "Iterated Learning."
[46] Chomsky, *Theory of Syntax*.
[47] Smith et al., "Complex Systems."
[48] Cuskley, "Alien Symbols"; Grüne-Yanoff, "Explanatory Potential."
[49] Wikipedia, "Generative Grammar," last modified March 12, 2025, 12:11 (UTC), https://en.wikipedia.org/wiki/Generative_grammar.
[50] Siglidis, "Latent Reading," 194.
[51] Wikipedia, "Sociolinguistics," last modified April 14, 2025, 22:00 (UTC), https://en.wikipedia.org/wiki/Sociolinguistics.
[52] Feucht et al., "Token Erasure."
[53] Touvron et al., "LLaMA"; Radford, "Improving Language Understanding."
[54] Vaswani, "Attention Is All."
[55] Pochinkov, "LLM Basics."

representation becomes more refined, encoding more and more context, a set of representations that we call the *latent space*, including the final ones. To represent the overall meaning of a sequence, text can be either embedded as a sequence of tokens that can be averaged to their mean representation (see Figure 3b) or summarized through an auxiliary token, which is more common in encoder-only models such as BERT, however.[56]

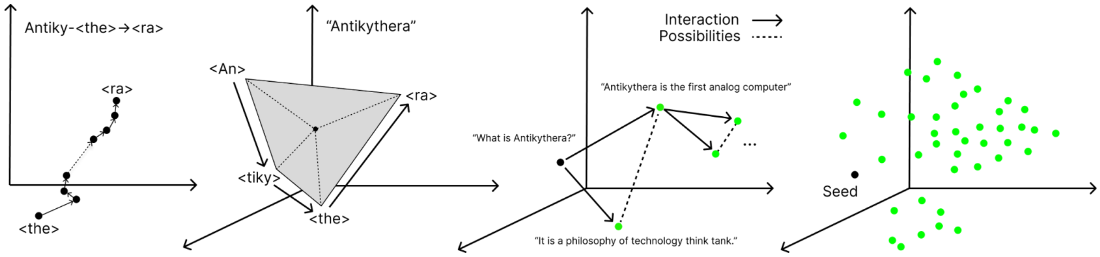

Figure 3 From tokens to topology (left to right): (a) NTP: A transformer-decoder architecture decomposes an input sentence into a series of tokens that it progressively maps into representations of increased complexity. (b) Text Geometry: These representations can be aggregated to encode the meaning of a sentence. (c) Discussion Dynamics: Aggregated representation can reveal discussion dynamics, (d) Topology: which in turn topologically span the embedding space of a LLM.

We refer to this average representation as the *embedding space*. When seen across sentences, embeddings reveal a set of possible discussion dynamics (see Figure 3c), which can later unfold topologically (see Figure 3d).[57]

To perform chatbot-like interactions, LLMs are trained to effectively *role-play* by appending existing text with markers such as “human:” or “AI:”.[58] Because of the uniform attention across all past tokens in each LLM’s transformer, such simple descriptions can heavily influence the produced outputs. In general, careful prompt engineering, data labeling, and curation is crucial to improve LLMs’ contextual performance.[59] However, although training for next-word prediction makes it possible to fit the target distribution, some of the ways this can be achieved may not align with product expectations of social interaction. To fix this, human feedback (HF) across LLM outputs is recorded on a small pool of annotators. When averaged, these preferences approximate an average population preference, a common practice in human perception studies, as is the case for image memorability, for example.[60] Simulating those rewards, a system is then trained to generate scores that can be provided as real-time feedback to the LLM’s outputs, to further fine-tune it with reinforcement learning (RL) to improve these scores, a technique known as RLHF.[61]

All these components—the architecture, the prompt engineering, the data—compose a specific instance of a large language model that is impossible to think of as universal in its design. Some iterations later, or with a different data set or a different prompt, the model could produce a significantly different output.[62] However, LLMs are still a *cultural technology*.[63] Through their cultural alignment, they can operationally arrive at describing what we think of as “universal” and potentially challenge its fundamental assumptions. More seen as a language computer than an imitation game, LLMs are special in that they can be manipulated through interventions that can be articulated or mediated in both mechanistic and representational ways.[64] This enables interventions across all scales of language, from grammar to sociality. In section 4.2, we propose such interventions as a bidirectional interface, building on the sociolinguistic and simulation framework of sections 2 and 3.

[56] Devlin et al., “BERT: Pre-Training.”
[57] Fitz et al., “Topological Aspects.”
[58] E.g., Luque, “Context-Aware LLM Chatbot.”
[59] Zhou et al., “LIMA.”
[60] Khosla et al., “Image Memorability.”
[61] Ouyang et al., “Training Language Models.”
[62] Shen et al., “Understanding Data Combinations”2023; Errica et al., “Quantifying LLMs’ Sensitivity.” 2024
[63] Gopnik, “Large Language Models.”
[64] Zou et al, “Representation Engineering.”

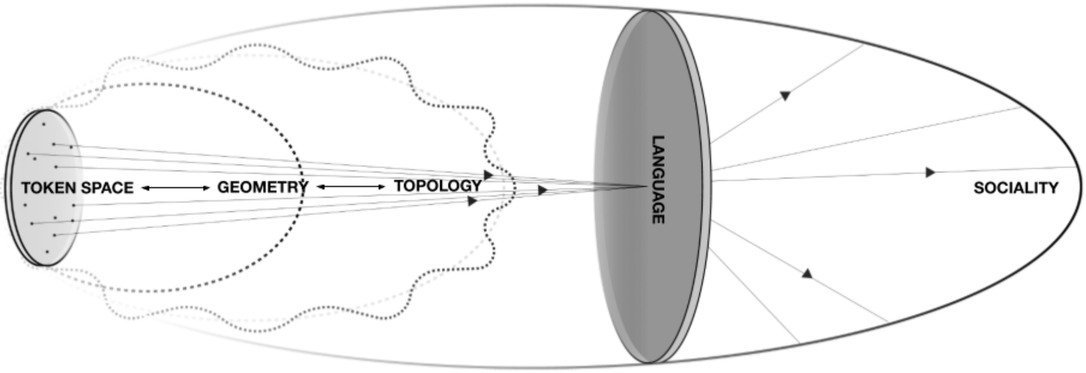

Figure 4 Generative topolinguistics: Our bidirectional framework manipulates an LLM across token space, geometry, and topology to produce new forms of language and sociality, reversing the arrows of the three-mirrors model of Figure 1.

4.2 Speculative Approaches

Here we introduce speculative approaches that explore linguistic interventions at three different scales, ranging from tokens, to geometry, to topology (see Figure 4). As each scale comes with distinct properties, we discuss each in a dedicated subsection. First, we introduce token-based interventions, where tokens act as agents that can manipulate the model's output by learning to satisfy a user-based reward function through RL. Then, we discuss the more standard dimension of our framework, where properly designed geometric manipulation of an LLM's latent space can influence its overall output. Finally, we discuss topological manipulation, first by studying sequences of LLM interactions and afterward by extending this approach to competitive environments to discover emergent social behaviors.

Tokens as Agents

Since token interactions occur across multiple transformer layers, isolating a single token's effect on an output sentence is challenging, as the relationship between the signifier and its signified is often broken in later layers.[65] A macroscopic approach could be to forbid a set of tokens, either during sampling or by keeping the same short-range outputs and masking them when performing longer generations. Comparing statistics across long generations for a fixed range of seeds can provide estimates of how such a combination of words affects the output generation. However, what if we use combinations of tokens, words, as a way to search and manipulate the outputs of an LLM? Analogous to an RL agent discovering walking from scratch,[66] token sequence can be assigned to a multilayer controller that can deform their output and, by learning to optimize a reward function while respecting constraints, learn how to manipulate other tokens. For example, a reward function could enforce similarity constraints between tokens while steering outputs toward a target goal, for example a score-based function trained to decrease populism on social media or appeal to a certain user. This is reminiscent of adversarial attacks in large language models,[67] yet our goal here is to understand the structure of token interactions by using certain words as means of exploration.

[65] Feucht et al., "Token Erasure."
[66] Heess et al., "Emergence of Locomotion Behaviours."
[67] Carlini et al., "Aligned Neural Networks."

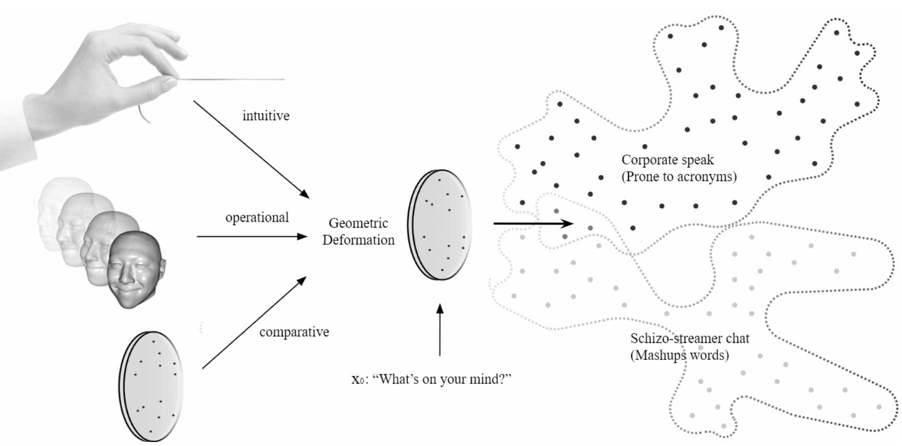

Figure 5 Speculative topolinguistic interface: using different modes of manipulation to produce geometric and topological manipulations of large language models.

Geometric Manipulation

Moving further from the study of individual token interactions, we can now think of how a global geometric manipulation of the latent space of LLMs can be used to steer its overall linguistic output, using the same set of inputs. In an ideal setting, we would like to define a methodology that is analogous with the discovery of latent directions in the embedding space of generative adversarial networks,[68] which in the case of faces is known to be able to linearize visual attributes such as skin color or facial expressions. However, as language is discrete in nature and is modeled sequentially, there isn't a clear approach for how to directly achieve this. For example, one way would be to concatenate the input sequence of an LLM with an extra adaptation token, similar to Zhu and colleagues,[69] whose role would be to influence the context of all other tokens toward a certain topic, changing the sentiment or the style of a conversation. Another approach would be to directly learn a low-rank adapter, a linear probe[70] or a sparse neuron decomposition[71] that manipulates individual layers of the large language model towards the same goal. Our intended purpose, however, would be to learn those manipulations not toward discrete goals but to associate them with certain input modalities (see Figure 5).

Inspired by Chen and colleagues,[72] who showed how such *mechanistic* interventions can be used for transparent bidirectional interfaces for customizing conversational agents, we can imagine a tactile intuitive interface that learns to translate touch signals or pose signals into geometric deformation through iterative feedback to help users perform a form of exploration. To facilitate this, we could also learn an operational mapping that is used as a reward signal to translate the output of the network into a set of output rewards, a procedure similar to RLHF. For example, we could learn how to associate facial expressions, or bodily signals such as pulse rate or body temperature, with a certain set of linguistic utterances. Except for using output sentences to analyze the proposed manipulations, one can also compare the produced adapters across input subjects or performed tasks.

Topological Contouring

Instead of focusing on individual LLM outputs, we can now focus on sequences of interactions. For this, we would have to first *individuate* LLMs to agents and design how to route the output of one to another.[73] Proposals for this kind of implementation are multiple, including generative agents or Concordia, which we discussed in section 3.[74] Given this formulation, consequent outputs of LLM interactions would trace a specific part of the embedding space with a higher likelihood, as is demonstrated on the right part of Figure 5. Inspired by the control theory of LLMs,[75] we can see LLM interactions as defining a space or reachability according to a certain set of initial conditions and prompts. In this experiment, we propose to relate geometric manipulations, like the ones discussed in the previous section, to how certain LLM interactions cover or not cover parts of the embedding space. One way to measure this would be by checking content similarity before and after training to a fixed set of prompts that describe topics and behaviors.

[68] Härkönen et al., "GANSpace."
[69] Zhu et al. "Virtual Tokens."
[70] Zou et al., "Representation Engineering."
[71] Lieberum et al., "Gemma Scope."
[72] Chen et al., "Designing a Dashboard."
[73] E.g., Varshney, "Introduction to LLM Agents."
[74] Park et al., "Generative Agents"; Vezhnevets et al., "Generative Agent-Based Modeling."
[75] Bhargava et al., "Control Theory."

LLM Game of Life

However, large language models may already encompass linguistic utterances that we aren't aware of yet, but which may be more efficient for them to communicate. Drawing from the works of Textworld and Emergent Linguistics, where *communication games* can be used to either solve games through language or create a new language to solve games,[76] a similar approach could be applied, this time to pretrained large language models, by fine-tuning or adapting specialized models to discover different linguistic utterances toward that goal. Similar to how LLMs can discover code words to communicate more efficiently, they might discover different ways of organization to achieve the same goal. This is what we describe as *comparative* in Figure 5, where the representations of one language model can be used to affect and describe another. By designing a competitive environment with selection dynamics, learning roles in LLM agents can be a way of discovering emergent sociality through an LLM game of life.?

5 Conclusion

Our paper suggests that the boundary between artificial and human linguistic systems is more permeable than previously conceived. We can expect to discover that the coevolution of these systems may lead to the emergence of hybrid sociolinguistic phenomena that defy traditional categorizations. Through its general and operational nature, our paper also raises important questions about the nature of linguistic agency in an era where artificial systems play an increasingly prominent role in shaping communicative norms and practices. This realization necessitates a more nuanced approach to the development and deployment of LLMs, one that takes into account their potential to reshape the very social fabric they aim to model. Using an empirical generative framework, this work speculated on experimental approaches to question and understand preconceived notions of language and sociality.

[76] Côté et al., "Textworld"; Lazaridou and Baroni, "Emergent Multi-Agent Communication."

Bibliography

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31, no. 3 (2023): 337–51. https://doi.org/10.1017/pan.2023.2.

Axelrod, Robert. "The Dissemination of Culture: A Model with Local Convergence and Global Polarization." *Journal of Conflict Resolution* 41, no. 2 (1997): 203–26. https://doi.org/10.1177/0022002797041002001.

Bender, Emily M., and Alexander Koller. "Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.463.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 ." In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.

Bhargava, Aman, Cameron Witkowski, Manav Shah, and Matt Thomson. "What's the Magic Word? A Control Theory of LLM Prompting." Preprint, *arXiv*, October 2, 2023. https://doi.org/10.48550/arXiv.2310.04444.

Bojić, Ljubiša, Matteo Cinelli, Dubravko Ćulibrk, and Boris Delibašić. "CERN for AI: A Theoretical Framework for Autonomous Simulation-Based Artificial Intelligence Testing and Alignment." *European Journal of Futures Research* 12, no. 1 (2024): 15. https://doi.org/10.1186/s40309-024-00238-0.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. "On the Opportunities and Risks of Foundation Models." Preprint, *arXiv*, August 16, 2021. https://doi.org/10.48550/arXiv.2108.07258.

Bucholtz, Mary, and Kira Hall. "Identity and Interaction: A Sociocultural Linguistic Approach." *Discourse Studies* 7, no. 4–5 (2005): 585–614. https://doi.org/10.1177/1461445605054407.

Byrne, David, and Gill Callaghan. *Complexity Theory and the Social Sciences: The State of the Art*. Routledge, 2022.

Carley, Kathleen, and Allen Newell. "The Nature of the Social Agent*." *Journal of Mathematical Sociology* 19, no. 4 (1994): 221–62. https://doi.org/10.1080/0022250X.1994.9990145.

Carlini, Nicholas, Milad Nasr, Christopher A. Choquette-Choo, et al. "Are Aligned Neural Networks Adversarially Aligned?" *Advances in Neural Information Processing Systems* 36 (2023): 61478–500. https://papers.nips.cc/paper_files/paper/2023/hash/c1f0b856a35986348ab3414177266f75-Abstract-Conference.html.

Côté, Marc-Alexandre, Ákoś Kádár, Xingdi Yuan, et al. "Textworld: A Learning Environment for Text-Based Games." In *Computer Games, CGW 2018*, edited by Tristan Cazenave, Abdallah Saffidine, and Nathan Sturtevant. Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-24337-1_3.

Chen, Yida, Aoyu Wu, Trevor DePodesta, et al. "Designing a Dashboard for Transparency and Control of Conversational AI." Preprint, *arXiv*, June 12, 2024. https://doi.org/10.48550/arXiv.2406.07882.

Chomsky, Noam. *Aspects of the Theory of Syntax*. MIT Press, 1965.

Chomsky, Noam. "The Case Against B.F. Skinner." *New York Review of Books* 17, no. 11 (1971): 18–24.

Clark, Andy. "Language, Embodiment, and the Cognitive Niche." *Trends in Cognitive Sciences* 10, no. 8 (2006): 370–74. https://doi.org/10.1016/j.tics.2006.06.012.

Cuskley, Christine. “Alien Symbols for Alien Language: Iterated Learning in a Unique, Novel Signal Space.” In *Proceedings of the 12th International Conference on the Evolution of Language*, edited by Christine Cuskley, Molly Flaherty, Hannah Little, Luke McCrohon, Andrea Ravignani, and Tessa Verhoef. Wydawnictwo Naukowe UMK, 2018. https://doi.org/10.12775/3991-1.018.

Dai, Gordon, Weijia Zhang, Jinhan Li, et al. “Artificial Leviathan: Exploring Social Evolution of LLM Agents Through the Lens of Hobbesian Social Contract Theory.” Preprint, *arXiv*, June 20, 2024. https://doi.org/10.48550/arXiv.2406.14373.

Deffuant, Guillaume, David Neau, Frederic Amblard, and Gérard Weisbuch. “Mixing Beliefs Among Interacting Agents.” *Advances in Complex Systems* 3, no. 01n04 (2000): 87–98. https://doi.org/10.1142/S0219525900000078.

DeLanda, Manuel. *Philosophy and Simulation: The Emergence of Synthetic Reason*. Bloomsbury Academic, 2019.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/N19-1423.

Duranti, Alessandro. *Linguistic Anthropology*. Cambridge University Press, 1997.

Dyson, Freeman. “A meeting with Enrico Fermi.” *Nature* 427, no. 6972 (2004): 297-297.

Eckert, Penelope. “Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation.” *Annual Review of Anthropology* 41 (2012): 87–100. https://doi.org/10.1146/annurev-anthro-092611-145828.

Elhage, Nelson, Neel Nanda, Catherine Olsson, et al. “A Mathematical Framework for Transformer Circuits.” *Transformer Circuits Thread* 1, no. 1 (2021): 12.

Enfield, Nick J., and Stephen C. Levinson, eds. *Roots of Human Sociality: Culture, Cognition and Interaction*. Berg, 2006.

Epstein, Joshua M. “Inverse Generative Social Science: Backward to the Future.” *Journal of Artificial Societies and Social Simulation* 26, no. 2 (2023): 9. https://doi.org/10.18564/jasss.5083.

Epstein, Joshua M., and Robert Axtell. *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press, 1996.

Errica, Federico, Giuseppe Siracusano, Domenico Sanvito, and Roberto Bifulco. “What Did I Do Wrong? Quantifying LLMs’ Sensitivity and Consistency to Prompt Engineering.” Preprint, *arXiv*, June 18, 2024. https://doi.org/10.48550/arXiv.2406.12334.

Evans, Nicholas, and Stephen C. Levinson. “The Myth of Language Universals: Language Diversity and Its Importance for Cognitive Science.” *Behavioral and Brain Sciences* 32, no. 5 (2009): 429–48. https://doi.org/10.1017/S0140525X0999094X.

Fedorenko, Evelina, Idan A. Blank, Matthew Siegelman, and Zachary Mineroff. “Lack of Selectivity for Syntax Relative to Word Meanings Throughout the Language Network.” *Cognition* 203 (2020): 104348. https://doi.org/10.1016/j.cognition.2020.104348.

Feucht, Sheridan, David Atkinson, Byron Wallace, and David Bau. “Token Erasure as a Footprint of Implicit Vocabulary Items in LLMs.” Preprint, *arXiv*, June 28, 2024. https://doi.org/10.48550/arXiv.2406.20086.

Fitz, Stephen, Peter Romero, and Jiyan Jonas Schneider. “Hidden Holes: Topological Aspects of Language Models.” Preprint, *arXiv*, June 9, 2024. https://doi.org/10.48550/arXiv.2406.05798.

Gao, Chen, Xiaochong Lan, Zhihong Lu, et al. “S3: Social-Network Simulation System with Large Language Model-Empowered Agents.” *SSRN Electronic Journal*, October 19, 2023. https://doi.org/10.2139/ssrn.4607026.

Giddens, Anthony. *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press, 1984.

Gopnik, Alison. "Large Language Models as a Cultural Technology." Uploaded July 14, 2022, by Simons Institute. YouTube, 14:00. https://www.youtube.com/watch?v=k7rPtFLH6yw.

Grüne-Yanoff, Till. "The Explanatory Potential of Artificial Societies." *Synthese* 169, no. 3 (2009): 539–55. https://doi.org/10.1007/s11229-008-9429-0.

Härkönen, Erik, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. "GANSpace: Discovering Interpretable GAN Controls." *Advances in Neural Information Processing Systems* 33 (2020): 9841–50. https://papers.nips.cc/paper_files/paper/2020/file/6fe43269967adbb64ec6149852b5cc3e-Paper.pdf.

He, James, Felix Wallis, Andrés Gvirtz, and Steve Rathje. "Artificial Intelligence Chatbots Mimic Human Collective Behaviour." Preprint, *Research Square*, version 2, January 15, 2024. https://doi.org/10.21203/rs.3.rs-3096289/v2.

Heess, Nicolas, Dhruva Tb, Srinivasan Sriram, et al. "Emergence of Locomotion Behaviours in Rich Environments." Preprint, *arXiv*, July 7, 2017. https://doi.org/10.48550/arXiv.1707.02286.

Hovy, Dirk, and Shannon L. Spruit. "The Social Impact of Natural Language Processing." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, edited by Katrin Erk and Noah A. Smith. Association for Computational Linguistics, 2016. https://doi.org/10.18653/v1/P16-2096.

Hu, Edward J., Yelong Shen, Phillip Wallis, et al. "LoRA: Low-Rank Adaptation of Large Language Models." Preprint, *arXiv*, June 17, 2021. https://doi.org/10.48550/arXiv.2106.09685.

Hunter, Tim. "The Chomsky Hierarchy." In *A Companion to Chomsky*, edited by Nicholas Allott, Terje Lohndal, and Georges Rey.Wiley Blackwell, 2021. https://doi.org/10.1002/9781119598732.ch5.

Jiang, Julie, and Emilio Ferrara. "Social-LLM: Modeling User Behavior at Scale Using Language Models and Social Network Data." Preprint, *arXiv*, December 31, 2023. https://doi.org/10.48550/arXiv.2401.00893.

Jones, Cameron R., and Benjamin K. Bergen. "People Cannot Distinguish GPT-4 from a Human in a Turing Test." Preprint, *arXiv*, May 9, 2024. https://doi.org/10.48550/arXiv.2405.08007.

Khosla, Aditya, Akhil S. Raju, Antonio Torralba, and Aude Oliva. "Understanding and Predicting Image Memorability at a Large Scale." In *Proceedings: 2015 IEEE International Conference on Computer Vision, ICCV 2015.* IEEE Computer Society, 2015. https://doi.org/10.1109/ICCV.2015.275.

Kirby, Simon. "Spontaneous Evolution of Linguistic Structure: An Iterated Learning Model of the Emergence of Regularity and Irregularity." *IEEE Transactions on Evolutionary Computation* 5, no. 2 (2001): 102–10. https://doi.org/10.1109/4235.918430.

Kirby, Simon, Tom Griffiths, and Kenny Smith. "Iterated Learning and the Evolution of Language." *Current Opinion in Neurobiology* 28 (2014): 108–14. https://doi.org/10.1016/j.conb.2014.07.014.

Lampinen, Andrew K., Ishita Dasgupta, Stephanie C. Y. Chan, et al. "Language Models, Like Humans, Show Content Effects on Reasoning Tasks." *PNAS Nexus* 3, no. 7 (2024): pgae233. https://doi.org/10.1093/pnasnexus/pgae233.

Lazaridou, Angeliki, and Marco Baroni. "Emergent Multi-Agent Communication in the Deep Learning Era." Preprint, *arXiv*. June 3, 2020. https://doi.org/10.48550/arXiv.2006.02419.

Li, Siyu, Jin Yang, and Kui Zhao. "Are You in a Masquerade? Exploring the Behavior and Impact of Large Language Model Driven Social Bots in Online Social Networks." Preprint, *arXiv*, July 19, 2023. https://doi.org/10.48550/arXiv.2307.10337.

Lieberum, Tom, Senthooran Rajamanoharan, Arthur Conmy, et al. “Gemma Scope: Open Sparse Autoencoders Everywhere All at Once on Gemma 2.” Preprint, *arXiv*, August 9, 2024. https://doi.org/10.48550/arXiv.2408.05147.

Linzen, Tal. “How Can We Accelerate Progress Towards Human-Like Linguistic Generalization?” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.465.

Luhmann, Niklas. “Systemtheorie, Evolutionstheorie und Kommunikationstheorie.” In *Soziologische Aufklärung 2: Aufsätze zur Theorie der Gesellschaft*. Verlag für Sozialwissenschaften, 1975. https://doi.org/10.1007/978-3-663-12374-3_10.

Luque, Rodrigo. “Building a Context-Aware LLM Chatbot with LangChain.” *Medium*, July 12, 2024. https://roluquec.medium.com/building-a-context-aware-llm-chatbot-with-langchain-996d372cedbb.

Lyotard, Jean-François. *The Postmodern Condition*. University of Minnesota Press, 1984.

Manning, Christopher, and Hinrich Schutze. “Foundations of statistical natural language processing”. *MIT* press, 1999.

Mohtashami, Amirkeivan, Florian Hartmann, Sian Gooding, Lukas Zilka, Matt Sharifi, and Blaise Aguera y Arcas. “Social Learning: Towards Collaborative Learning with Large Language Models.” Preprint, *arXiv*, December 18, 2023. https://doi.org/10.48550/arXiv.2312.11441.

Nisioti, Eleni, Claire Glanois, Elias Najarro, et al. “From Text to Life: On the Reciprocal Relationship between Artificial Life and Large Language Models.” In *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*, edited by Andrés Faíña, Sebastian Risi, and Eric Medvet. MIT, 2024. https://doi.org/10.1162/isal_a_00759.

Ouyang, Long, Jeff Wu, Xu Jiang, et al. “Training Language Models to Follow Instructions with Human Feedback.” *Advances in Neural Information Processing Systems* 35 (2022): 27730–44. https://papers.nips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Park, Joon Sung, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. “Generative Agents: Interactive Simulacra of Human Behavior.” In *UIST ’23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, edited by Sean Follmer and Jeff Han. Association for Computing Machinery, 2023. https://doi.org/10.1145/3586183.3606763.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, et al. “Generative Agent Simulations of 1,000 People.” Preprint, *arXiv*, November 15, 2024.” https://doi.org/10.48550/arXiv.2411.10109.

Pochinkov, Nikita. “LLM Basics: Embedding Spaces – Transformer Token Vectors Are Not Points in Space.” *AI Alignment Forum*, February 13, 2023. https://www.alignmentforum.org/posts/pHPmMGEMYefk9jLeh/llm-basics-embedding-spaces-transformer-token-vectors-are.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. “Improving Language Understanding by Generative Pre-Training.” Working paper preprint, OpenAI, 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Rutten, Geert-Jan. “Broca-Wernicke Theories: A Historical Perspective.” In *Handbook of Clinical Neurology*, vol. 185, edited by Argye Elizabeth Hillis and Julius Fridriksson. Elsevier, 2022. https://doi.org/10.1016/B978-0-12-823384-9.00001-3.

Saffran, Jenny R. “What Is Statistical Learning, and What Statistical Learning Is Not.” In *Neuroconstructivism: The New Science of Cognitive Development*. Oxford University Press, 2009. https://doi.org/10.1093/acprof:oso/9780195331059.003.0009.

Saussure, Ferdinand de. “Course in General Linguistics.” In *Literary Theory: An Anthology*, 2nd ed., edited by Julie Rivkin and Michael Ryan. Blackwell Publishing, 1998.

Schelling, Thomas C. *Micromotives and Macrobehavior*. W. W. Norton & Company, 1978.

Shannon, Claude E. "The redundancy of English." In Cybernetics; Transactions of the 7th Conference, New York: Josiah Macy, Jr. Foundation, pp. 248-272. 1951.

Shen, Zhiqiang, Tianhua Tao, Liqun Ma, et al. “SlimPajama-DC: Understanding Data Combinations for LLM Training.” Preprint, *arXiv*, September 19, 2023. https://doi.org/10.48550/arXiv.2309.10818.

Siglidis, Yannis. “Latent Reading.” In *Chimeras. Inventory of Synthetic Cognition*, edited by Ilan Manouach and Anna Engelhardt. Onassis Publications, 2022.

Sinha, Chris. *Ten Lectures on Language, Culture and Mind: Cultural, Developmental and Evolutionary Perspectives in Cognitive Linguistics*. Brill, 2017.

Skinner, Burrhus Frederic. *Science and Human Behavior*. Simon and Schuster, 1965.

Smith, Kenny, Henry Brighton, and Simon Kirby. “Complex Systems in Language Evolution: The Cultural Emergence of Compositional Structure.” *Advances in Complex Systems* 6, no. 4 (2003): 537–58. https://doi.org/10.1142/S0219525903001055.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, et al. “LLaMA: Open and Efficient Foundation Language Models.” Preprint, *arXiv*, February 27, 2023. https://doi.org/10.48550/arXiv.2302.13971.

Valsiner, Jaan, and Judith A. Lawrence. “Human Development in Culture Across the Life Span.” In *Basic Processes and Human Development*. Vol. 2 of *Handbook of Cross-Cultural Psychology*, 2nd ed., edited by John W. Berry, Pierre S. Dasen, and T. S. Saraswathi. Allyn and Bacon, 1997.

Varshney, Tanay. “Introduction to LLM Agents.” *Nvidia Developer* (blog), November 30, 2023. https://developer.nvidia.com/blog/introduction-to-llm-agents/

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. “Attention Is All You Need.” *Advances in Neural Information Processing Systems* 30 (2017). https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vezhnevets, Alexander S., John P. Agapiou, Avia Aharon, et al. “Generative Agent-Based Modeling with Actions Grounded in Physical, Social, or Digital Space Using Concordia.” Preprint, *arXiv*, December 6, 2023. https://doi.org/10.48550/arXiv.2312.03664.

Vygotsky, Lev S. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978.

Watts, Duncan J., and Steven H. Strogatz. “Collective Dynamics of ‘Small-World’ Networks.” *Nature* 393, no. 6684 (1998): 440–42. https://doi.org/10.1038/30918.

Wittgenstein, Ludwig. *Philosophical Investigations*. John Wiley & Sons, 1953.

Zou, Andy, Long Phan, Sarah Chen, et al. “Representation Engineering: A Top-Down Approach to AI Transparency.” Preprint, *arXiv*, October 2, 2023. https://doi.org/10.48550/arXiv.2310.01405.

Zhou, Chunting, Pengfei Liu, Puxin Xu, et al. “LIMA: Less Is More for Alignment.” *Advances in Neural Information Processing Systems* 36 (2024). https://papers.nips.cc/paper_files/paper/2023/file/ac662d74829e4407ce1d126477f4a03a-Paper-Conference.pdf.

Zhu, Yutao, Zhaoheng Huang, Zhicheng Dou, and Ji-Rong Wen. “One Token Can Help! Learning Scalable and Pluggable Virtual Tokens for Retrieval-Augmented Large Language Models.” Preprint, *arXiv*, May 30, 2024. https://doi.org/10.48550/arXiv.2405.19670.