

4th January 2026, paper

Model Collapse and Natural Language

by Eino Ilvonen

themes: information, entropy, novelty

Cite as: Eino Ilvonen, “Model Collapse and Natural Language”. Institute for Syntropic Research, 4 January 2026.
<https://syntropyinstitute.net/model-collapse-and-natural-language/>

keywords: Information, Entropy, LLMs, Novelty, Natural Language

Abstract. The present paper analyses how exposure to large language models (LLMs) may affect natural language. Specifically, it identifies *model collapse* as a systemic dysfunction of artificial languages which may seep into natural language. In AI discourse, the term *model collapse* is used to represent thermodynamic *entropy*, a measure of disorder: when LLMs are recursively trained on their own content, their ability to accurately represent the informational landscape diminishes. The output becomes more homogenous and complexity is lost. As more people rely on LLMs in their everyday life, we argue for the importance of conserving the complexity of natural language in all its facets. Crucially, we identify the spontaneous production of novelty as a key feature of natural language that may counterbalance entropic model collapse. We identify this feature in natural language as thermodynamic *syntropy*.

1. Introduction: Neural Networks and Next Token Prediction

To begin, we look at the functional structure of LLMs. The key mechanism of large language models consists of a generation of output that is based on next token prediction. This means that LLMs are programmed to generate the output with the highest probability. In order to predict the next token, “LLMs make use of enormous neural networks—computational models loosely inspired by the human brain that process data through layers of connected nodes (neurons)—trained over raw text to predict the next word as accurately as possible.” (Levy *et al.*, 2025).

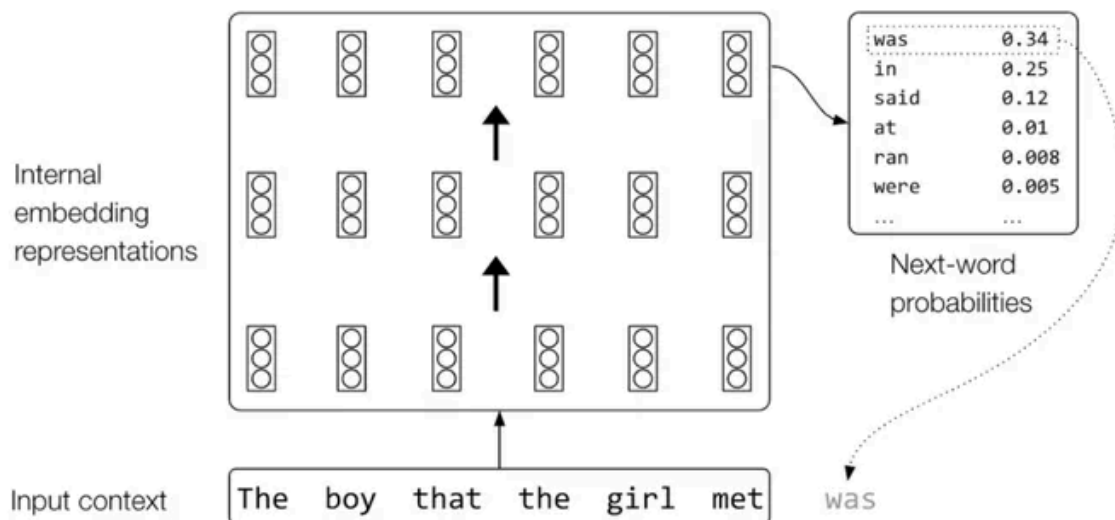


Figure 1

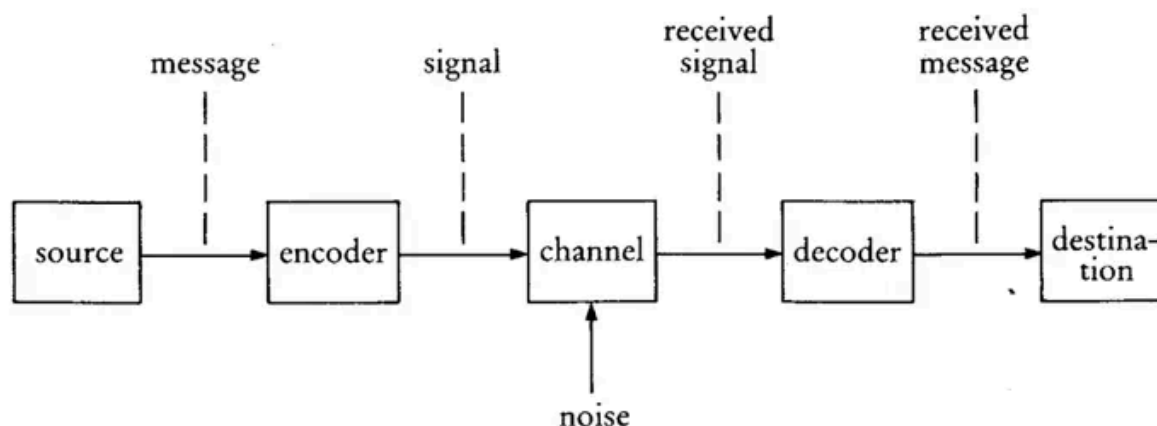
An LLM takes the input context *the boy that the girl met* and passes it through a series of neural network layers (typically based on the transformer architecture; Vaswani et al., 2017) to obtain internal embedding representations of each word in context. These embedding representations are used to predict the next-word probabilities. Generation proceeds by sampling a word (or taking the most likely word) and then appending it to the previous input to obtain a new context.

Levy et al. (2025). *Next token prediction*.

Artificial neural networks, then, mimic the neural networks of the brain by looking at the structural composition and recreating them through computational models. Warren McCulloch and Walter Pitts (1943) were the first to suggest the idea of artificial neural networks in modelling a neural net using an electrical circuit. In the 1962 documentary *The Living Machine*, McCulloch explains the *anastomotic* character of neural nets, and how they differ from conventional machines:

“May I make clear a distinction between your nervous system and any computing machine which is yet in existence. In an ordinary machine, operations are carried out sequentially, and a mistake entails mistakes which follow inevitably from it, all the way through the computation. (...) But I’m talking about machines that resemble the mouth of the Nile, these are, what the greeks called *anastomotic* affairs, that is, the water from each contributory stream is mixed with the water from every other contributory stream before they finally get out of the mouths of the river.”(McCulloch, 1962)

Today, LLMs function precisely in this manner: computing through complex layered processes. The next token is predicted based on datasets much larger than data that humans can acquire over their lifetime. However, artificial language approximates natural language only in the form of code, in direct, context-free translation; while natural language appears to be highly context-sensitive.



Shannon & Weaver (1949). *A Model of Communication*.

In 1949, Shannon & Weaver proposed a model of general communication in which code played a central role in translation and transmission, and although code may be a central feature of language, it is not an exhaustive explanatory model. For the anthropologist Michael Tomasello, “the linguistic “code” rests on a nonlinguistic infrastructure of intentional understanding and common conceptual ground”(p. 58), which LLMs cannot access.

As LLMs are based on next token prediction, a subsequent key mechanism is that of merging the produced token to the existing sequence. The computational system merges two syntactic objects, producing a new syntactic object. Natural language follows similar patterns, as Levy *et al.* (2025) argue: “In generative linguistic theories, syntactic structure characterizes the major paths of meaning composition by which words recursively combine into larger and larger phrases”. Further, they argue that “syntactic structure identification and meaning composition is highly incremental”.

Yet, the computational process of next token prediction followed by merging seems to be qualitatively different from the linearity of natural language, which is not linear in its internal structures. While artificial language generates a string through a linear process, natural language follows a non-linear process that leads to a linear output. This would suggest that LLMs still carry traces of the first type of machine identified by McCulloch: machines that operate sequentially rather than anastomotically. The consequences are as outlined by McCulloch: “a mistake entails mistakes which follow inevitably from it, all the way through the computation.”. We turn to these mistakes in the next section in considering model collapse.

2. Model Collapse: LLMs and Entropy

The symbol of the Ouroboros, is often used to represent self-consumption: if the mass of a system is also its fuel, it means that the life-span of the system is limited. This is the case in *model collapse*, which occurs when LLMs are trained on data that was generated by earlier versions of themselves. In terms of information, the generated content becomes more

homogenous with each recursive step, as each step includes a certain level of abstraction or simplification. As the content generated by LLMs grows exponentially, the percentage of overall content available for LLM training also grows exponentially, which leads to a reduction of the overall complexity of the informational environment.

Within the system as a whole, the effect on the informational environment is *entropic*, which means that the system tends towards more disorder. Since the concept of entropy is used within multiple disciplines to represent different processes, we first discuss model collapse in terms of *thermodynamic* entropy. According to the second law of thermodynamics, the entropy count always increases within a closed system. This law also applies to psychological systems: the mind, when interpreted as a thermodynamic system, can only survive with a continuous inflow of new information. If we look at LLMs as neural nets, that is, as models of the mind, the lack of primary novel information (rather than secondary, already processed information) leads to a systemic collapse into entropy.

A paper by Shumailov *et al.* (2023) discusses several dangers in model collapse, including the process where “models start forgetting improbable events over time, as the model becomes poisoned with its own projection of reality”(p. 2). The paper emphasizes the significantly large extent of data poisoning and the effect that it may have on the amplification of biases and on our ability to understand complex systems. The loss of rare events may also have a profound impact on overall language use and creativity, as artistic processes rely precisely on multiplicity and the spontaneous emergence of rare events. Shumailov *et al.* (2023) suggest that model collapse may occur across all models, not just in generative LLMs. This underlines the possibility of a spill-over effect of model collapse to natural language.

In his article *The false promise of ChatGPT*, Chomsky criticizes the amount of energy that goes into developing LLMs. “The human mind is not, like ChatGPT and its ilk, a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question (...)”, instead, “True intelligence is demonstrated in the ability to think and express improbable but insightful things”. In the next section, we turn to the creative faculty of language in relation to entropy and probability.

3. Creativity and Probability

In this section, we argue that the creative capacity of natural language is based on risk-taking, and is therefore in its function and orientedness diametrically opposed to mechanisms choosing the safest path. As a consequence, the influence from a language model programmed not to take risks may decrease creative potential, but it may also become a phenomenon that natural language can resist to.

So far, we have considered entropy as a *thermodynamic* unit of measurement. This means that we apply the concept of entropy to open and closed abstract systems, and that we look at entropy gradients. Entropy is treated differently in information theory. Here, it represents

an absolute measure, which is measured in bits. Information entropy (or *Shannon entropy*) represents a “measure of uncertainty and randomness encoded in a message.” (Karaca & Moonis, 2022). Thus, higher information entropy means higher surprise.

When a TV shows static white noise, each pixel being at random either “white” or “black”, it has high information, entropy, and surprise. Comparatively, when the TV shows a movie and an image crystallises, there is less information, less entropy, and less surprise, as pixels follow a certain movement and can be predicted based on previous pixels. The string of characters “sdlfbliudhtfktlf” has high Shannon entropy, because each character has low chances of being predicted. Compared to that, the string “the apple is red” has lower information. However, we know intuitively that “the apple is red” carries more semantic information; we can say that as the absolute information decreases, the semantic information or ‘signal’ increases.

In information theory, an environment of low entropy is characterized by high *syntropy* (from gr. *syn-*, “together” and *tropos*, “tendency”, a “converging tendency”). As Mendez-Moreno (2024) writes, “A lower syntropy indicates a more complex and less predictable series, characterized by a more uniform distribution of values. Conversely, higher syntropy suggests more predictability and less randomness, often implying repetitive or regular patterns within the data.”(p. 16). Model collapse goes towards maximum entropy, an informational environment of homogeneity, which is characterized by a lack of nuance, complexity, and “rare” events. In the recursive abstraction from abstraction, considerable nuance is lost, leading to an entropic informational environment of maximum predictability, without novelty.

The danger of this becomes evident in the comparison to ecology, where the resilience of an ecosystem depends on its variance, and where homogeneity presents a structural weakness. Systems of language behave like ecological systems: there exist delicate ecological balances, invasive species, as well as tipping points. For humans, an environment affected by “data poisoning” carries relatively low message, precisely because of the exclusion of rare, unpredictable events. Communication relies precisely on unpredictable events, precisely because prior to communication, there is always the desire to express something novel.

Natural language follows the same thermodynamic constraints as artificial language. Without novelty (input), there is collapse. However, natural language is unique in its ability to create novelty out of itself. This happens because it is flawed, and can therefore evolve dialectically through the removal of its imperfections. The opposite is the case with artificial language. In *A Cyborg Manifesto*, Donna Haraway writes: “Cyborg politics are the struggle for language and the struggle against perfect communication, against the one code that translates all meaning perfectly, the central dogma of phallogocentrism.”(p. 57). Here, Haraway criticizes the dominance of smooth functionality in cybernetic, computational language. As Levy *et al.* (2025) write, “modern generative AI technologies is largely use or performance driven”.

In terms of information, natural language presents a special case. In artistic expression, there is the presence of both *low information* (lower than a randomized sequence) and *low predictability*. If we consider the following syntagm: “the cat sneers at the wind”, we can see that it communicates something precisely because the sequence of words is unexpected.

Even though natural language is broadly characterised by syntropy, it also has entropic dimensions of low predictability. This can have positive expressions, as is the case in artistic expression, but it can also have negative effects on the language user. This can be seen in disorders like schizophrenia, where disorganised speech is a key diagnostic marker (DSM-5). This suggests that entropic language may affect well-being, as the ability to communicate meaning is a necessary part of our social world.

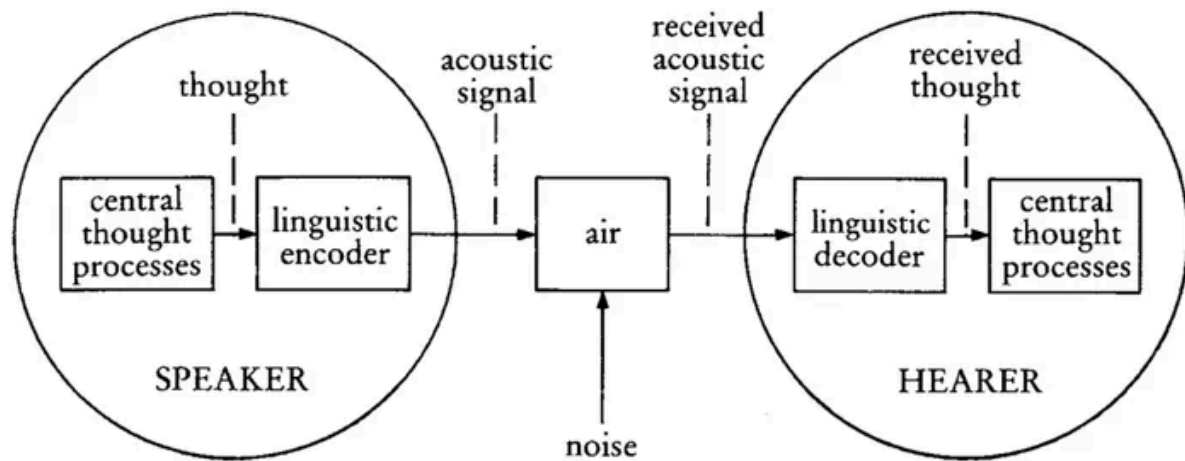
While a certain degree of entropy is always present in language, there can be model collapse in natural language when entropy surpasses a certain degree. As a consequence, it is crucial to be watchful of the interference of entropic models with natural language. Further, we should consider whether interaction with LLM language may increase the occurrence of mental health disruptions. We may say that language has a drive towards disintegration, but inherently resists it through the novelty produced in seeking out new paths.

4. Information and Dialogue

As we have seen, artificial language mimics the qualities of natural language. As a consequence, it has both syntropic and entropic dimensions. Yet, it lacks a key dimension of natural language, which is spontaneity and creativity, constituted of “rare events” and *novelty*. The relevance of novelty was advocated for by Terrence McKenna, who conceived of the universe as a novelty-generating and conserving engine. If this is true, artificial language may only persist as long as it produces novelty, and as its internal mechanisms suggest, this may not be the case indefinitely.

Artificial neural networks are modelled after the anatomy of the brain: biological neurons are modeled through *nodes* as artificial neurons, further, they are connected by *edges*, which model synapses in the brain. Although this increases efficiency, interpreting the brain as a network has its limits: a network is always a simplified representation that reduces a given system to an abstract structure. Vannevar Bush, the inventor of the *memex* (memory-extension), may have been one of the first to write about hypertext and cybernetic human-machine interaction. In *As We May Think*, he writes: “The human mind operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain... Man cannot hope fully to duplicate this mental process artificially, but he certainly ought to be able to learn from it.” (Vannevar Bush, 1945).

Going beyond neurons ordered in a system, human communication is made up of multiple systems, as well as relations between systems. There are feedback systems between speakers that enable a dialectic in the original, etymological sense of the word as “dialogue”. Sperber & Wilson (1986) applied Shannon & Weaver’s model to human communication, modifying their classical communication model as follows:



Sperber & Wilson, Relevance (1986).

The schema represents speaker and hearer as circles, similar to the representation of artificial neurons. The circle stands for a coherent system with outbound and inbound arrows. This form of dialogue is not present within LLMs, as the model makes no distinction between actors and merely processes information. Again, nuance is lost, and certain canyons of meaning are deepened. The canyons of allowed, normal pathways create pressure on the individual, and cause it to look for escape routes. It is hypothesized that schizophrenic patients have “shallow cognitive maps” which lead to more unexpected associative connections (Musa *et al.* 2022). In the language of Deleuze & Guattari’s schizoanalysis, the schizophrenic breaks with the canyons of meaning, deterritorializing the linguistic environment. This may represent potential against the “techno-linguistic automata” of semiocapitalism (Berardi, 2017).

According to the French philosopher Bernard Stiegler, we are currently living in an *Entropocene*, an age where entropy is produced on a mass scale (Stiegler, 2018). We may say that the process of pouring energy into LLMs represents an increase in entropy. Stiegler urged for syntropic tactics; in terms of language, this could mean creating and protecting systems of language of high coherence and high syntropy. These systems could be imagined as rhizomatic systems, characterized by high levels of integration.

In 1950, the physicist Leon Brillouin published an article titled *Thermodynamics and Information Theory*, in which he argues that although information gain always means a loss of syntropy, there are creative devices like books that can prevent information loss. Brillouin differentiates two types of work: active, creative thinking which requires “invention, methodical imagination, unexpected comparisons; and it results in a new idea, representing a completely new type of absolute information.”(p. 598), and passive work like the exact translation of a message. He argues that machines are only suited for the second type of work. “There is no block labeled “imagination” nor “unprogrammed comparisons.”(Brillouin, 1950, p. 599).

Conclusion

While natural and artificial language systems were identified as syntropic (low entropy), artificial language was found to lack a crucial characteristic of natural language, namely the production of novelty. For this reason, artificial language systems may collapse into entropy, as occurs in model collapse. Through “data poisoning”, a homogeneous environment is created, as rare events are excluded. These rare events, displayed in artistic forms of using language (e.g. poetry, literature), are frequent in natural language. They represent the novelty required to resist entropy in a thermodynamic system. As LLMs have limited data and feed on their own products, they resemble a closed thermodynamic system. We argued that through excessive interaction with LLM entropy, natural language may shift into an entropic state which may be expressed as a mental health disorder or reinforce already existing ones.

Through the analysis of different language states, it becomes clear that novelty, creativity, and spontaneity are essential dimensions of natural language. Not only are these dimensions absent in LLM language, the “data poisoning” present in LLM model collapse may even increase the entropy of natural language. Possible solutions may be found in redesigning LLMs towards choosing less probable outcomes; but it is not clear what an LLM would base this choice on. As we often see reflected in poetry, natural memory and natural language are deeply bound to the space and time they are experienced in.

*“My head leaning on a shoulder,
thirty times brighter than the sun,
I watch their return.”*

(Tape Mark 1, 1961)

Eino Ilvonen is a writer and independent researcher. He is currently pursuing a Masters degree in Linguistics at the University of Amsterdam. His research and present contribution is concerned with the philosophy of artificial and natural language.

Bibliography

- Berardi, F. (2017). *Futurability: The age of impotence and the horizon of possibility*. Verso Books.
- Bush, V. (1945). *As we may think*. The atlantic monthly, 176(1), 101-108.
- Brillouin, L. (1950). *Thermodynamics and information theory*. American Scientist, 38(4).
- Chomsky, N., Roberts, I., & Watumull, J. (2023). *Noam Chomsky: The false promise of chatgpt*. The New York Times, 8.
- Haraway, D. (2010). *A cyborg manifesto* (1985). Cultural theory: An anthology, 454.
- Karaca, Y., & Moonis, M. (2022). *Shannon entropy-based complexity quantification of nonlinear stochastic process*. In Multi-Chaos, Fractal and Multi-Fractional Artificial Intelligence of Different Complex Systems (pp. 231-245). Academic Press.

- Levy, R., Kim, Y., & Fox, D. (2025). *The Science of Language in the Era of Generative AI*.
- McCulloch, W. S., & Pitts, W. (1943). *A logical calculus of the ideas immanent in nervous activity*. The bulletin of mathematical biophysics, 5(4), 115-133.
- McCulloch, W. S. (1962). *The Living Machine*. Documentary, directed by Roman Kroitor, National Film Board of Canada. <https://www.nfb.ca/film/the-living-machine/>
- Mendez-Moreno, S. (2024). *Syntropy in complex systems: A complement to Shannon's Entropy*. arXiv preprint arXiv:2403.13022.
- Musa, A., Khan, S., Mujahid, M., & El-Gaby, M. (2022). *The shallow cognitive map hypothesis: A hippocampal framework for thought disorder in schizophrenia*. Schizophrenia, 8(1), 34.
- Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). *The curse of recursion: Training on generated data makes models forget*. arXiv preprint arXiv:2305.17493.
- Shannon, C. E., & Weaver, W. (1949). *A mathematical model of communication*. Urbana, IL: University of Illinois Press, 11, 11-20.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Cambridge, MA: Harvard University Press.
- Stiegler, B. (2018). *The Neganthropocene*. Open Humanities Press.
- Tomasello, M. (2010). *Origins of human communication*. MIT Press.