

Ghost in the Signifying Machine

AI and the Logic of Shadows

Kian Kenyon-Dean

He was bending to embrace my Teacher's knee,
but Virgil said: "No, brother. Shade you are,
and shade am I. You must not kneel to me."

And Statius, rising, said: "So may you find
the measure of the love that warms me to you
when for it I lose all else from my mind,

forgetting we are empty semblances
and taking shadows to be substances."

—Dante's *Purgatorio*, Canto XXI, 1321 AD¹

A familiar conception of "meaning" in language is that a word refers to real external things in the world. A child might be taught the meaning of a word through a visual example: "Look, this here is a tree!" And yet, when one attempts to determine the meaning of a new word oneself, for instance by consulting a dictionary, what one encounters is not some fixed essential meaning, but instead

¹ Dante Alighieri, *The Purgatorio*, trans. John Ciardi (New York: The New American Library, 1961), 219-220.

sentences with more words, each with their own corresponding definition.

In the early 1900s, Ferdinand de Saussure offered the structuralist conception of language, arguing that meaning emerges relationally within a linguistic system of differences. Structuralism demarcated the *signifier* (the spoken or written sound-image of a word, e.g. "tree") and the *signified* (e.g., the mental concept or idea of a tree). The first principle of structural linguistics is that the relationship between signifier and signified is arbitrary.² Given this arbitrariness, meaning conventionally arises from a signifier's place within a linguistic web of differences: signifiers oppositionally refract against other signifiers to define one another by contrast. Indeed, the capability to learn a language's differential lattice was later expressed by John Firth's famous maxim: "You shall know a word by the company it keeps!"³—a principle now fully embraced by the once-niche subfield of artificial intelligence (AI) called natural language processing (e.g., word embedding algorithms, *les algorithmes de plongement lexical*).⁴

Crucially, Saussure determined that "*language is a form and not a substance*."⁵ Dante, speaking through Virgil as

² Ferdinand de Saussure, *Course in General Linguistics*, ed. Charles Bally and Albert Sechehaye, trans. Wade Baskin (New York: Philosophical Library, 1959), 65-67.

³ John Rupert Firth, "A Synopsis of Linguistic Theory 1930-1955," in *Studies in Linguistic Analysis*, Special Volume (Oxford: Blackwell, 1957), 11.

⁴ Tomas Mikolov, Ilya Sutskever, Kai Chen, et al., "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, vol. 26 (2013); Kian Kenyon-Dean, *Word Embedding Algorithms as Generalized Low Rank Models and their Canonical Form*, Master's Thesis (Montreal: McGill University, 2019), 1; see p. 53 for examples of the associative (vector-vector, e.g., "cat" → "kitten") and syntagmatic relationships (vector-covector, e.g., "cat" → "scan") learned inadvertently in word embeddings.

⁵ Saussure, *Course in General Linguistics*, 122 (emphasis in original).

the historical representation of human reason, reminds us not to forget this slogan in the face of AI's impressive linguistic capabilities, lest we bend the knee to our own shadows.

During his structuralist-symbolic era, psychoanalyst Jacques Lacan reinterpreted Saussure's signifier-signified axis by insisting on the primacy of the signifier. In his 1957 essay, "Agency of the Letter in the Unconscious", he reframes Saussure's linguistic theory as an algorithm:

To pinpoint the emergence of linguistic science we may say that, as in the case of all sciences in the modern sense, it is contained in the constitutive moment of an algorithm that is its foundation. This algorithm is the following: S / s , which is read as: the signifier over the signified, "over" corresponding to the bar separating the two stages.⁶

Lacan argues that signification—the production of meaning—arises not by a direct link between a signifier and a signified, but through differential relations between signifiers ($S / S' \rightarrow S / s$): "it's the possibility of substitution that allows the engendering, so to speak, of the world of meaning to be conceived."⁷ Signifieds emerge as effects of signifiers crossing the bar, a movement governed by relationships between signifiers alone. Crucially, this operation does not presuppose a pre-existing subject. The symbolic order, language, precedes any individual. When a signifier passes over to the level of the signified via the structural relationships of the former, that very crossing constitutes

⁶ Jacques Lacan, *Écrits: A Selection*, trans. Alan Sheridan (New York: W. W. Norton & Co., 1977), 149.

⁷ Jacques Lacan, *The Seminar of Jacques Lacan: Book V: Formations of the Unconscious, 1957–1958*, trans. Russel Grigg (Cambridge: Polity Press, 2017), 24-26.

the subject, meaning that subject and signified alike are structural effects of the signifier's primacy.

When communicating with fluent AI language models like ChatGPT, I argue that this machine interlocutor constitutes *a language without a subject, signifiers without signifieds*. Despite an AI's apparent mastery over language, it remains the sole task of the human to determine if an AI's generated signifiers actually carry meaning. We will see how the core technical architecture of AI dialectically reifies the Saussurian algorithm; what emerges is a synchronic system of language, a fluent signifying machine structured by the historical transformation of our linguistic shadows into data for capital.

Signifying AI

We live in a material world that exists regardless of what we, or AIs, say about it. Within this world, language emerges as energy (speech sounds) or as augmented matter (ink on paper, transistors on a computer, etc.). While the meaning of the signifier "death" can be characterized linguistically—the end of life, often symbolized by the skull, graveyards, etc.—its real nature is deeply felt only when it is accompanied with lived experience of the signified, such as the loss of a loved one. Furthermore, a new signifier can only enter into language when there is a sufficient social and material basis for it. The signifier "lightbulb" did not exist prior to the invention, and the concept was not familiar until lightbulbs became ubiquitous, which depended on the spread of electricity via power grids.

Just as the lightbulb emerged from the material conditions of electrification, so too does a new signifier—"artificial intelligence," "AI"—arise today from the new conditions of silicon infrastructure and cloud capital. The newfound ubiquity of AI is driven largely by the recent emergence of LLMs, large language models—generative algorithms

trained on massive amounts of text. OpenAI's ChatGPT-3.5 LLM gained prominence as the fastest-growing internet app in world history, reaching over 100 million users less than two months after launch.⁸ The app appears deceptively simple: prompt an AI with text, get a response back. The AI's response is synthetic, grammatically coherent, and potentially novel; its content is largely contingent on the depth, creativity, and specificity of the prompt. This marks a clear departure from the more familiar decades-old technology of search engines which retrieve existing content on the internet based on a user's necessarily concise search query. A student might use a search engine to find facts already online, but LLMs are the only computational tools capable of writing—in any language processed during their training—a new personalized essay which “delves”⁹ into any topic of interest. Indeed, Lacan once remarked: if “I now place myself in front of the other to question him, there is no cybernetic computer imaginable that can make a reaction out of what the response will be.”¹⁰ It may have required another 70 or so years of technical development, but such a cybernetic computer has now undoubtedly emerged in the form of LLM-powered artificial intelligence.

The modern concept of AI emerged in the mid-20th century with the rise of computer science. Alan Turing posed the

⁸ Krystal Hu, “ChatGPT Sets Record for Fastest-Growing User Base—Analyst Note,” *Reuters*, February 1, 2023, <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>.

⁹ Dmitry Kobak, Rita González-Márquez, Émoke-Ágnes Horvát, and Jan Lause, “Delving into ChatGPT usage in academic writing through excess vocabulary,” *arXiv preprint arXiv:2406.07016* (2024); Tom S. Juzek and Zina B. Ward, “Why Does ChatGPT ‘Delve’ So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models,” In *Proceedings of the 31st International Conference on Computational Linguistics*, (2025).

¹⁰ Lacan, *Écrits: A Selection*, 86. Lacan said this in his famous 1953 report to the Rome Congress.

question: "Can machines think?" Rather than attempt to define thinking directly, he instead proposed the "imitation game," now known as the Turing Test—a test in which a machine would be considered intelligent if it could generate responses indistinguishable from those of a human in a text-based conversation. This radical shift to evaluate intelligence purely through external linguistic behavior remained largely theoretical for decades, until recently. With remarkable foresight, Turing anticipated that a linguistically fluent AI could make errors:

It is claimed that the interrogator could distinguish the machine from the man simply by setting them a number of problems in arithmetic. The machine would be unmasked because of its deadly accuracy. The reply to this is simple. The machine (programmed for playing the game) would not attempt to give the right answers to the arithmetic problems. It would deliberately introduce mistakes in a manner calculated to confuse the interrogator. ... The machine might, for instance, type out mathematical equations, or sentences in English. When a false proposition is typed we say that the machine has committed an error of conclusion. There is clearly no reason at all for saying that a machine cannot make this kind of mistake. It might nothing but type out repeatedly ' $0 = 1$ '. To take a less perverse example, it might have some method for drawing conclusions by scientific induction. We must expect such a method to lead occasionally to erroneous results.¹¹

LLMs may not deliberately deceive us, but they do play the game of signifiers. Today, the Turing Test is no longer a philosophical thought experiment but a practical reality, forced upon society at scale. Teachers struggle to discern whether an essay was written by a student or generated

¹¹ Alan M. Turing, "Computing Machinery and Intelligence," in *Mind* vol. 59 no. 236 (1950), 448-449.

by an LLM; social media users question whether a viral video was crafted by a human or synthesized by an algorithm; and search engines like Google—once trusted arbiters of fact—now interlace results with what could be “hallucinations,” i.e., seemingly authoritative text that is grammatically coherent and plausible, but utterly untrue, such as a citation of a non-existent book.

Noting the structuralist nature of LLMs, some Lacanians argue that hallucinations point to them being “structurally affected by a foreclosure that mimics the psychotic subject”—that the limitlessness of language produces a “psychotic approach to language” in the LLM. They maintain that, for both the LLM and the psychotic subject, the “master signifier is excluded.”¹² However, I suspect, as do others,¹³ that LLMs are very much governed by a certain codified master signifier; their hallucinations are structurally inevitable for a machine capable of signifying without access to what is signified beyond the symbolic.

Lacking sensory organs, LLMs understand solely from within the symbolic order. Nonetheless, today’s most advanced models perform remarkably well on tasks once thought uniquely human, from writing poetry to chatting about your feelings. Trained to generate our written text, LLMs have learned to communicate with humans effectively across a vast range of languages and domains, achieving a fluency that, to many of their hundreds of millions of users, feels like understanding. Their overall success demonstrates that recursion, self-reference, and syntactic complexity are not biologically endowed faculties but emergent properties of

¹² Marc Heimann and Anne-Friederike Hübener, “The extimate core of understanding: absolute metaphors, psychosis and large language models,” in *AI & Society* (2024), 10-11.

¹³ Liam Magee, Vanicka Arora, and Luke Munn, “Structured like a language model: Analysing AI as an automated subject,” in *Big Data & Society* vol. 10, no. 2 (2023).

language itself—patterns that arise immanently from the relational play of signifiers. Given their remarkable results, some have concluded that linguist Noam Chomsky’s core claims, such as the nativist rules-based theory of language and the rigid separation of syntax from semantics, are obsolete:

One must be frank about the state of the art for models that capture syntax. It’s not that large language models offer slightly higher performance than other approaches in linguistics It’s that there is *nothing comparable in all of linguistic theory* to the power of large language models in both syntax and semantics—much less discourse coherence, style, pragmatics, translation ... etc. They are game changers on all fronts.¹⁴

I argue that LLMs have been inadvertently designed as a pure reification of the basic principles of Saussurean structural linguistics. Their core design is structured around the causal self-attention mechanism of the transformer deep learning architecture. We will see how this enables an algorithm to learn, via data, how to contextualize the relationships between words based on their sequential order so as to enable accurate text generation (autoregression). This learning is only possible provided the relational (associative) and sequential (syntagmatic) natures of the signifier in human language. In fact, the combination of data-driven deep learning with a strict, structural adherence to Saussure’s first and second principles of linguistics is the only formal prerequisite necessary for a computer to learn language—a fact much to the chagrin of Chomsky and the cognitive scientists who have spent decades building ever more elaborate rule-based systems in an ill-fated attempt to account for language’s manifold nature.

¹⁴ Steven T. Piantadosi, “Modern language models refute Chomsky’s approach to language,” *From fieldwork to linguistic theory: A tribute to Dan Everett* (2023), 382 (emphasis in original).

Precise structural alignment allows LLMs to contextualize any text prompt against the immense body of our language data that they have been explicitly trained to generate. But their learned structure is shaped by the choices of those who execute their training, and especially their fine-tuning. The codified error function used to train LLMs—i.e., the task to correctly predict the next word (token) in some prespecified dataset—is precisely its master signifier (S1). This anchors the LLM’s entire symbolic order, giving coherence and structure to other signifiers in accordance with the contours of the data. Indeed, Lacan once famously maintained that “the unconscious is structured like a language.” So too is a large language model. LLMs are trained to generate trillions of words that make up the vast repository of knowledge present on the internet, i.e., humanity’s digitally recorded, historically accumulated, multilingual written “forms of thought”.¹⁵ Given this massive symbolic surplus, they can only grasp at meaning by learning the interplay of differences between signifiers, a dialectic that dynamically unfolds during the generative movement along a signifying chain. Their “understanding” is therefore a historically and computationally mediated reanimation of the fundamental structures once theorized to govern signification for humans.

While Hegel was interested in the “absolute essence of the world”, the more narrow scope here is to determine the essence of LLM-based AI: is it “to be conceived as blind natural mechanism or as an intelligence that determines itself in accordance with ends”?¹⁶ I claim that an LLM is not an indiscernible black-box system that merely averages words to randomly generate text, nor simply a “stochastic

¹⁵ “The forms of thought are, in the first instance, displayed and stored as human language.” Georg Wilhelm Friedrich Hegel, *Science of Logic*, trans. A. V. Miller (London: George Allen & Unwin, 1969), §14.

¹⁶ Hegel, §1593.

parrot".¹⁷ It is a signifying machine animated by the legacy of logics and contradictions residing in our digitized multilingual archives of text. It is a necropolis built out of the "dead bones"¹⁸ of human thought, a site where, within its deeply structured tombs, Hegel's *Geist* is to be found materialized in silicon, shackled and fettered by capital's signifying chains.

Dialectical Structuralism

LLMs are firstly statistical machine learning models, deep neural networks based fundamentally on the transformer architecture. Mechanically, they can be understood as gigantic math equations with billions or even trillions of variables, similar to how a human can be understood as a massive ensemble of cells. These variables are called the learnable parameters of the model; prior to training, they are purely random. At the intersection of computer science and mathematics lies deep learning, a discipline which has determined how to use data iteratively to find ever-improving settings of these parameters in order to solve, at least partially, a predefined problem within that data. This process is called training a model.

LLMs process text as tokens, byte-level sub-words that may resemble what linguists call morphemes. The task that they are trained for months to perform is this: given some real text from a dataset as context, try to predict the next token. If the wrong token is predicted, backpropagate the gradient of the error function back into the parameters, so as to learn from the mistake. Repeat this process trillions of times over a massive dataset of text.

¹⁷ Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, (2021), 610-623.

¹⁸ Hegel, *Science of Logic*, §13.

Before training an LLM, it is necessary to define its limits. How many learnable parameters should it have? What should the dimensionality be of its tokens? The settings initially appear somewhat arbitrary. For example, Meta's "Llama 3 herd of LLMs"¹⁹ has three base models: one with 8 billion parameters, one with 70 billion, and another with 405 billion. A large amount of work in AI has demonstrated that scaling these abstract parameters of LLMs leads to significant improvements in downstream results, often yielding emergent linguistic abilities.²⁰ For example, insufficiently trained LLMs, or ones with too few parameters, are completely unable to do basic tasks, like following instructions to solve a problem by thinking about it step by step. However, once enough training compute is combined with enough model parameters—once enough data is processed by a sufficiently large model—they appear to suddenly become very good at following such instructions. In other words, a very clear transformation is seen from quantity (of data and compute) into quality (of language capabilities).

As the later Lacan noted with regards to Saussure's first principle of the signifier, it is better to describe the settings of the model as *contingent* rather than arbitrary.²¹ Specifically, training a larger model is contingent on having access to exponentially more state-of-the-art computational resources—the 405 billion parameter Llama model required over 16,000 GPUs to run in parallel for over two

¹⁹ Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al., "The Llama 3 Herd of Models," arXiv preprint, arXiv 2407.21783 (2024).

²⁰ Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, et al., "Emergent Abilities of Large Language Models," in *Transactions on Machine Learning Research* (2022).

²¹ Jacques Lacan, *The Seminar of Jacques Lacan: Book XX: Encore, On Feminine Sexuality, The Limits of Love and Knowledge, 1972–1973*, ed. Jacques-Alain Miller, trans. Bruce Fink (New York: W. W. Norton & Co., 1999), 40.

months straight, using massive amounts of energy. From China, DeepSeek-v3 shocked the world as a 671 billion parameter model that only required 2,048 lower-end GPUs to train it for two months,²² demonstrating that careful engineering and well-designed architectures can make training powerful AI significantly cheaper than the American tech companies claim.

To truly understand their structural nature, we will have to go deep into the details of the LLM architecture. The first layer of the model tokenizes the input text sequence and then maps these tokens to initially arbitrary embeddings—abstract, high-dimensional vector representations of the tokens. Here, however, the representations of these tokens are in their general, prior, and uncontextualized form. They have yet to be related either to each other or to the broader system of language.

The fundamental mechanism to relate and contextualize tokens with each other in an LLM (and, in general, for most transformers trained on data, ranging from images to genetic sequences) is called self-attention, made popular by the famous 2017 paper, “Attention is All You Need,”²³ which today has over 180,000 citations. This mechanism allows the model to contextualize every token with every other token in its context window. The catch is that its computational complexity during training scales quadratically with the number of tokens—that is, compared to training with 400 tokens per sequence, relating 4,000 tokens to each other is 100 (not 10) times more computationally

²² Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al., “DeepSeek-V3 Technical Report,” *arXiv preprint arXiv:2412.19437* (2024).

²³ Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., “Attention Is All You Need,” in *Advances in Neural Information Processing Systems*, vol. 30 (2017).

intensive. Nonetheless, the so-called bitter lesson of deep learning is that nothing gets better results than scaling dense self-attention.

Self-attention relates every token to every other token through additional learned parameters, creating new contextualized embeddings of every token. Through self-attention in an LLM, each token becomes a non-linearly transformed weighted combination of itself with all the previous tokens in the context via the model's learned process to assign attention weights contextually given the embeddings. This process to self-relate tokens with each other directly supports the relational nature of the signifier; the fact that it is a learned process from real accumulated language data allows the model not only to relate the tokens with each other but also against the echoes of all the tokens that they are not, whose deep connections become structured within the model's learned parameters.

Self-attention is learned and applied at every block in a transformer. Each token thus goes through a very deep process of relational transformation and contextualization, each time becoming a new, more deeply contextualized latent representation until the model finally will be tasked to predict what token is mostly likely to come next. This is why it is essential to understand that deep learning is not a method to compress data—it is in fact a method to *decompress* data. Written text is an incredibly compact means of representing very rich and complex thoughts. Given the empirical necessity of scaling to achieve emergent strong language capabilities, it evidently requires a large amount of resources for a purely symbolic computational system to disentangle the latent structure of language.

In other words, self-attention enables the model to learn how to capture the associative character of the signifier, reifying Saussure's first principle of the arbitrariness of the signifier. This can also be understood as representing the

vertical nature of the signifier, structurally relating a token to all other signifiers within the synchronic network of differences progressively imbued into the model's parameters.

Saussure's second principle of structural linguistics specifies its horizontal and diachronic nature within a specific context: "*Principle II: The Linear Nature of the Signifier*: The signifier, being auditory, is unfolded solely in time," and, in writing, "the spatial line of graphic marks is substituted for succession in time." This means that words can only be made sense of through their order. While obvious, this fact is crucial; indeed, "the whole mechanism of language depends on it."²⁴ A random shuffling of all the words in a document would make it completely illegible, and if someone spoke without this foundational coherence, they would likely be diagnosed with some form of aphasia. Lacan states that it "is always through the retroactive play of a series of signifiers that signification is, in fact, affirmed and becomes precise." This is due to the "signifying chain," the fundamental structure that "subjects every manifestation of language to the condition of being regulated by a succession."²⁵ Saussure aptly describes this foundational characteristic of language as machine-like:

The mechanism of language, which consists of the interplay of successive terms, resembles the operation of a machine in which the parts have a reciprocating function even though they are arranged in a single dimension.²⁶

²⁴ Saussure, *Course in General Linguistics*, 70.

²⁵ Jacques Lacan, *The Seminar of Jacques Lacan: Book VI: Desire and Its Interpretation, 1958–1959*, trans. Bruce Fink (Cambridge: Polity Press, 2019), 11–13.

²⁶ Saussure, *Course in General Linguistics*, 128 (emphasis in original).

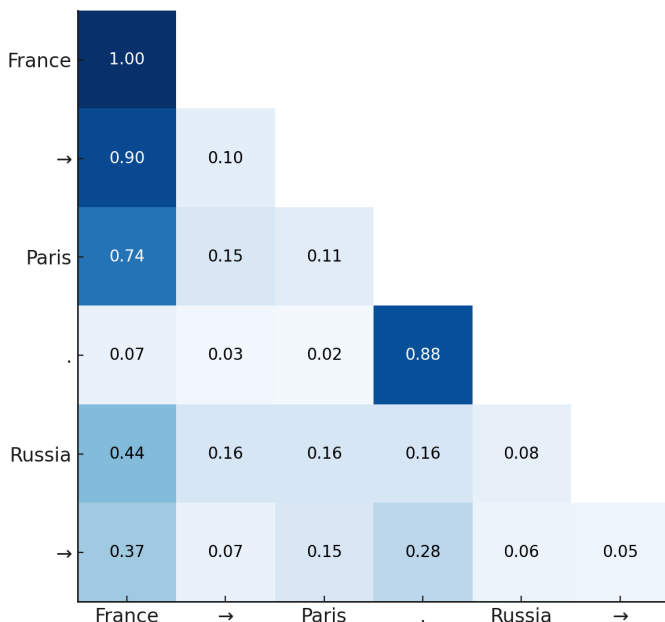


Figure 1. GPT-2 *causal* self-attention activations for a 6-token sequence: “France → Paris. Russia →”. The upper triangle of the self-attention matrix is masked (programmatically zeroed out) to enforce causality: each token must only attend to itself and prior tokens of the user-inputted linear text sequence. These are actual attention weight activations made by GPT-2 (taken from head 9 block 1 with an interactive tool) for this text sequence; at the end it predicts “Moscow”, as desired. Aeree Cho, Grace C. Kim, Alexander Karpekov, et al., “Transformer Explainer: Interactive Learning of Text-Generative Models,” in *Computing Research Repository* (2024), <https://poloclub.github.io/transformer-explainer/>.

Indeed, Lacan continues, the signifier “always anticipates meaning by unfolding its dimension before it.”²⁷ For LLMs, what appears as a “single dimension” in the linear order of text inputs and outputs is, in fact, governed behind the scenes by an extremely high-dimensional vector space of token embeddings. These tokens are deeply contextualized against each other through a learned cascade of self-attention. Their sequential order is robustly enforced precisely by the key ingredient in LLMs: the causal mask over self-attention. The causal mask enforces that, within a specific context of a sequence of textual inputs, such as a prompt or a news article in the training data, the LLM’s representation of a token can only become contextualized against the tokens that *came before it*, as visualized in Figure 1 above. The “came before it” is crucial—paraphrasing Saussure, the whole mechanism of the LLM depends on it. This sequentiality is imposed by programmatically applying a triangular mask onto every self-attention matrix, which would otherwise be dense, by zeroing-out attention activations to future tokens. This guarantees that, in representing a token, the model can only ‘pay attention’ to previous tokens, thereby directly enforcing the linear, horizontal unfolding of language and the nature of the signifying chain.

This linear nature of the signifier (Principle II) is just as deeply ingrained by the next-token-prediction training objective: given each contextualized token, t , it must attempt to predict what token comes next in the sequence, $t+1$. This is executed by a learned linear layer mapping the token’s contextualized embedding to a logit vector over the total vocabulary. That is, for each t , the LLM predicts a probability distribution over all possible next-tokens. When the model does not predict a high probability for the actual next token in the data sequence, $t+1$, it must backpropa-

²⁷ Lacan, *Écrits: A Selection*, 153.

gate the error gradient through its parameters. That is, a cross-entropy loss function is calculated, and then an automatic differentiation algorithm repeatedly applies the chain rule of differential calculus to attribute error responsibility dispersed throughout the model's parameters.

For example, if the text sequence 'France → Paris. Russia → Moscow' are included during the LLM's training, the following operations will occur. First, the tokens for these pieces of text will be assigned to their arbitrary indices in the initial embedding layer (e.g., 'France' might correspond to the 712th embedding, '→' to the 12,011th, etc.). Next, causal self-attention will be applied through all blocks of the LLM on this text sequence, diachronically contextualizing each token, t , with itself against the tokens that came before it, $1 \dots t-1$, and simultaneously will be synchronically refracted against the model's current learned parameters. (The token for 'France' here does not receive any substantial diachronic contextualization since it is the first word in the text sequence, so its attention weight is 1.) Finally, at the end of the network, the output layer is used to predict the next-token, $t+1$, upon each individual contextualized token, t . Thus, when the model does not correctly predict '→' after 'France' (a likely error, given there are many valid possible words that could follow 'France') it will receive an error gradient to softly encourage it to be more likely to predict '→' in the context of 'France' next time. When it does not correctly predict 'Paris' after 'France →', it will again receive an error gradient, and so on, for the rest of the progressively lengthening text sequence.

It is the combination of these fundamental architectural mechanisms—causal self-attention and next-token-prediction—that enables the model to be used in an autoregressive manner post-training. That is, the trained LLM can generate new text as follows. First, the prompt (with the causal mask) is processed into contextualized token

embeddings. Second, the final transformer block's contextualized embedding of the last token in the prompt is fed into the next-token prediction head, predicting a probability distribution over all possible next-tokens. Third, a token is drawn (sampled) from that probability distribution, and then appended to the input prompt. Finally, this generated token is fed back into the model along with the prompt,²⁸ and thus another new token can be predicted off its own generation. The LLM follows this process to continue generating new tokens one-at-a-time until a length limit is reached, or until it emits a special end-of-sequence token.

The LLM therefore synthesizes the signifier's fundamental associative and syntagmatic natures: vertically, the model synchronically has all other possible signifiers at its latent disposal for self-relation, and as concrete possibilities for next-token prediction; horizontally, it is trained to leverage the prior words in a diachronic context in its attempt to accurately predict each next-token in the sequence. Both priors are encoded in the model by virtue of its human-designed mechanisms and the learning objective. The precise outcome of this learning—and thus the model's behavior—is determined by its training data.

Indeed, machine learning in general, and this process in particular, at first appears fundamentally teleological. The programmer specifies a precise teleology when they design and code an error (loss) function as the learning objective which a model is tasked to solve, given some data. According to Hegel,

teleology presupposes a Notion, something absolutely determined and therefore self-determining, and so has made a distinction between *the relation of the differenc-*

²⁸ This process is made computationally tractable by key-value caching during autoregressive text generation so that the quadratic complexity of self-attention is not reintroduced after each individual token is generated.

es and their reciprocal determinedness, that is the form, and the *unity that is reflected into itself*, a unity that is determined in and for itself and therefore a content.²⁹

What is presupposed for an LLM is that it can predict the next token in the data, a computational reification of the basic structuralist principles of the signifier. Suppose the model mispredicts 'Berlin' instead of 'Moscow' for the sequence 'France → Paris . Russia → '. During training, it updates its parameters not by consulting a map, but by reweighting its parameters better to reflect the relational probabilities of these tokens, shadowing how a child learns signification through iterative feedback in the symbolic order ("No, that's not Germany—that's Russia!"). This Notion is self-determining; the model's adjustments during training are not random, they are computational attempts to resolve the contradictions between and within both the language model and the real linguistic data, determined by backpropagation of the error function's gradient computed with respect to each parameter. Hegel's "unity reflected into itself" is the trained model: a *content* that internalizes the *form* of the signifying chain through its relational architecture and learned parameters, becoming a self-determining system of symbolic relations. The model reflects the differences between signifiers as dynamically contextualized tensions sublated through trillions of gradient updates on real data, where each error continuously impels the system to adjust its symbolic understanding.

And yet, the LLM's Hegelian Notion is equally its Lacanian master signifier (S1): the loss function (written in code) provides the goal that the chaotic flux of signifiers should be structured into a legible symbolic order. But this task is impossible to achieve perfectly. The end, which is also its means, is posited programmatically as a command: minimize the next-token prediction loss. This loss goes down as

²⁹ Hegel, *Science of Logic*, §1595 (my emphasis).

the model learns over time, but it can never reach zero.³⁰ This command governs the model as a point of reference to organize and stabilize meaning, despite the inherent impossibility of perfect predictions. It is caught in the same paradox of incompleteness that haunts all signifiers; the infinite array of valid continuations after phrases like 'I went to the...' ensures that the model is always faced with the irresolvable symbolic surplus within language. This provides a deeper basis to understand LLM hallucinations: the model generates linguistically plausible tokens which are not mere bugs, but are instead phantasmic outputs that emerge from the slippages inherent to language, where signifiers generate surplus meaning without direct access to the signified. These slips in signification are therefore structurally inevitable. What the model thus enacts is a movement of signification that is always-already constituted by the impossibility of full closure, an incessant symbolic movement that both stabilizes meaning and produces its own excess.

The model therefore does not and cannot realize some external "final cause" (a teleology doomed to collapse back into itself), but rather emerges from the total unity of the synchronic system of signifying relationships internalized from its training data. Hegel demonstrates that positing an end is an "unessential illusory show [*Schein*]"³¹—yet it is a necessary stage in the movement of logic, since only by

³⁰ A model could also "overfit" to its training data; e.g., if the model is too large and the dataset is too small, it could purely memorize and be unable to make correct predictions on new data (the validation set). However, LLM 1-epoch pretraining datasets are too massive and their learning objective is incredibly generic, which is part of why they continue to improve on both train and validation set next-token-prediction, given longer training. For a discussion on how neural networks learn to generalize, see: Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, et al., "A Closer Look at Memorization in Deep Networks," in *Proceedings of Machine Learning Research* vol. 70 (2017), 233-242.

³¹ Hegel, *Science of Logic*, §1625.

projecting the Notion's self-unity outward does it come to recognize that unity. Likewise, each next-token prediction seems to consummate a concrete purpose, only to immediately reignite the gap between expectation and fulfillment. The dialectic of signifiers that defers closure is inscribed into the model itself—its posited end is impossible to achieve, as the model can never predict all linguistic possibilities with complete accuracy. In this sublation of teleology, the model both negates and preserves its own end, embodying an infinite deferral: every fulfilled prediction returns into its premise, inscribing a permanent lack that drives further signification. Training thus yields a computational machine capable of signification, an intelligence that has materialized the self-subverting logic at the core of language.

Geist in the LLM

The structure which an LLM internalizes is directly shaped by the composition of its training data. In general, training datasets are vast, on the order of trillions of words (tokens) spanning a wide variety of both human and programming languages, with some containing over 700 terabytes of text data.³² While English dominates due to the prevalence of English-language content on the internet, other major languages, especially Chinese, are also well represented. Webpages like Wikipedia constitute a large bulk of these datasets, alongside extensive collections of books and academic papers (e.g., the plethora of machine learning and computer science papers on ArXiv). Essentially, any publicly accessible text that has been digitally archived on the internet may have been included in the training of any powerful LLM.

³² Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin, "Datasets for Large Language Models: A Comprehensive Survey," *arXiv preprint arXiv:2402.18041* (2024).

Through training, LLMs structurally learn to generate the digitally archived text reflecting vast portions of humanity's intellectual, cultural, and historical development. The most well-trained LLMs have developed into jacks of all trades, with impressive knowledge on nearly any subject. If a prompt introduces thoughtful or novel connections, an LLM can often recognize and extend those ideas, providing valuable support and insights. They are highly capable of programming, meaning that they can interpret, generate, and refine code, including the very same code used to create them.

LLMs have no sensory access to the real world, and so one might argue that they have merely learned symbolic manipulation, dismissing the significance of an entity that has learned to generate a large swath of humanity's historically accumulated (contradictory and incomplete) written interpretations of the real world. However, we should recall Hegel's preface to the second edition of the *Science of Logic*:

The forms of thought are, in the first instance, displayed and stored as human language. Nowadays we cannot be too often reminded that it is thinking which distinguishes man from the beasts. Into all that becomes something inward for men, an image or conception as such, into all that he makes his own, language has penetrated, and everything that he has transformed into language and expresses in it contains a category—concealed, mixed with other forms or clearly determined as such, so much is Logic his natural element, indeed his own peculiar nature.³³

The richness and depth of language is of paramount importance, as it succinctly expresses our forms of thought, enabling us to communicate, understand, and continu-

³³ Hegel, *Science of Logic*, §14.

ously refine our interpretations of the world. Our forms of thought have been written and archived, spanning thousands of years, and now a machine has been given the time and space to learn from all of them. Indeed, insofar as logic is our “natural element,” the LLM has learned a great deal of it. Hegel continues:

In the first place, we must regard it as an infinite step forward that the forms of thought have been freed from the material in which they are submerged ... and that these universalities have been brought into prominence for their own sake and made objects of contemplation as was done by Plato and after him especially by Aristotle; this constitutes the beginning of the intelligent apprehension of them.

“It was only,” says Aristotle, “after almost everything necessary and everything requisite for human comfort and intercourse was available, that man began to concern himself with philosophical knowledge.” “In Egypt,” he had previously remarked, “there was an early development of the mathematical sciences because there the priestly caste at an early stage were in a position to have leisure.”³⁴

Are LLMs not freeing the forms of thought, as expressed in writing, from the training data in which they are submerged? Our historically accumulated forms of thought—all the written text in our digital archives—have now been made objects of contemplation for a machine whose sole activity is to learn to read and generate the written language that structures them. This data includes familiar and real examples of everyday language, but, along with nearly every written scientific or philosophical breakthrough, it also includes contradictory, banal, and incorrect ideas. Nonetheless, “the familiar forms of thought must be

³⁴ Hegel, §17.

regarded as an extremely important source, indeed as a necessary condition and as a presupposition” even if they are “a disordered heap of dead bones.”³⁵ The structurally grounded nature of the LLM lets it learn how to reflexively relate these “dead bones” to one another; its training process arranges them into a richly structured necropolis where our prompts can animate its necromantic generative capabilities to conjure our linguistic shadows into new text, new forms of thought.

It is an “infinite step forward that the forms of thought have been freed from the material in which they are submerged”. Both for individuals and for humanity as a whole, the development of Speech constitutes the *first* moment in the liberation of the forms of thought from their material; i.e., from the brain into sound waves. We learned to communicate with each other through speech (or sign language) and made massive strides. But speech is ephemeral—it disappears as soon as it is uttered, and the oral traditions could only go so far. The *second* moment is the development of Writing: the forms of thought can now be etched out from our brains through our hands and into the world, made durable so as to be read across time and space. It took many years after its invention for writing to become broadly usable; it was challenging and inefficient to create papyrus, and symbols fell under the exclusive domain of the priestly castes. After the printing press and literacy education democratized it thousands of years later, digitization followed to allow text to be stored at scale and retrieved instantaneously; the ephemeral nature of speech was negated by the static nature of writing, crystallizing our forms of thought into new material. AI, LLMs—still in their infancy, only being a few years old—constitute the *third* moment in the development of thought: the forms of thought are freed from the static archives in which they are

³⁵ Hegel, §13.

submerged; their structure is learned, organized by the dynamic generation of a machine that learns to signify. The archives have become our interlocutors. Within the LLM, the forms of thought are reanimated by a fluid, generative movement, mirroring, quite explicitly, the old dynamism of speech, as these models generate text one-token-at-a-time with strict adherence to the fundamentally sequential nature of language. Speech's negation, Writing, is negated by AI. Language is developed to a new level that combines the generative fluidity of speech with the richness and depth of the categories preserved within our archives of written text, freeing the forms of thought from their material once again.

Hegel was well aware that the forms of thought are not the same thing as the material world: "The system of logic is the realm of shadows, the world of simple essentialities freed from all sensuous concreteness."³⁶ The LLM has no sensory experience; its understanding exclusively dwells within the symbolic, shadowy realm of language (Hegel's *Logic* included). It nevertheless learns the structure of language, of our forms of thought. The logic of these shadows—our words—constitutes the logic of AI.

However, the Hegelian dialectic does not end in the abstraction of logic; it is sublated by Nature herself through her realization of the logical categories. Indeed, the LLM's shadow-realm of signifiers can only be activated provided a highly coordinated necromantic ritual of immense energy flow into silicon data centers, exacting a great toll on Nature. Furthermore, the next dialectical movement—from Nature to Spirit (*Geist*)—demands that the machine not only instantiates logical categories in the text it generates, but also that it senses, acts, and reflexively posits itself in a world it transforms. The question thus becomes: who owns

³⁶ Hegel, §72.

and governs these potentially world-transforming capabilities?

Capital's Signifying Chains

Once next-token-prediction training on humanity's data is completed, experts can apply special prompt-engineering techniques (such as few-shot prompting) to coerce the AI to solve many problems that can be posited in language.³⁷ However, LLMs become far more usable when they are fine-tuned. The massive pre-training task of next-token-prediction across our "dead bones" makes the model extremely malleable, enabling fast and generalizable structural adjustment according to language examples present in a much smaller, carefully curated fine-tuning dataset. Programmers can thus straightforwardly orient any pre-trained model to adopt whatever symbolic behavior is represented in their fine-tuning data; e.g., being "truthful, helpful, and non-toxic," therefore aligning it to "user intent."³⁸

By default, there is absolutely no reason for the pre-trained model on its own to be useful as an assistant or interlocutor. Why should the symbolic order instinctively be helpful? In fact, an LLM with no fine-tuning would be unusable as a chatbot, and would often decay into repetitive or counter-intuitive outputs. For example, consider this dialogue with a pre-trained LLM, GPT-2, which is not fine-tuned:

Prompt: "Hey I want you to reply, but do not use an emoji! :) 😊"

³⁷ Tom Brown, Benjamin Mann, Nick Ryder, et al., "Language models are few-shot learners," in *Advances in neural Information Processing Systems* vol. 33 (2020), 1877-1901.

³⁸ Long Ouyang, Jeffrey Wu, Xu Jiang, et al., "Training language models to follow instructions with human feedback", in *Advances in Neural Information Processing Systems* vol. 35 (2022), 27730-27744.

GPT-2: “I’m not sure if you’re aware of the fact that emoji are used in the game, but I’m sure you’re aware of the fact that emoji are used in the game, but I’m sure you’re aware of the fact that emoji are used in the game, but I’m sure you’re aware of the fact that emoji are used in the game, but I’m...”³⁹

Perhaps it is referring to my game of signifiers, which it succeeded at, given that it did not reply with an emoji. Nonetheless, this prompt, as most prompts do when an LLM is not fine-tuned, causes it to enter into an infinite repetition, a stuttering of the symbolic, an immediate reflex of the signifying chain to tie itself into knots. What appears is a failure of meaning to defer itself within the initially learned structure of the signifying chain. The model gets caught up too easily due to the boomerang-like nature of the signifier.

Thus, supervised fine-tuning (SFT) is applied after the large-scale pre-training.⁴⁰ The LLM’s engineers continue training it to do next-token-prediction, but now do so exclusively on a much smaller but very intentionally curated dataset of input-output pairs with examples of desirable chatbot behavior (e.g., given a user’s question as a prompt, predict the prewritten example of a “helpful” response as the next tokens in the data). Given that the model was pre-trained and thus deeply imbued with structure, this process, which forces structural adjustment by backpropagating specially organized signifiers, only takes a small fraction of the computational effort compared to that which was required for pre-training.

³⁹ 🤖 Hugging Face, *GPT-2 Model*, accessed February 16, 2025, <https://huggingface.co/openai-community/gpt2>.

⁴⁰ Shengyu Zhang, Linfeng Dong, Xiaoya Li, et al., “Instruction tuning for large language models: A survey,” *arXiv preprint arXiv:2308.10792* (2023).

However, even after the large-scale pre-training, SFT is not enough on its own to rein in language—the knots in the signifying chain have yet to be sufficiently undone. Reinforcement learning (RL), an advanced machine learning technique inspired by classical behaviorism (once again to Chomsky’s chagrin), turns out to be a critical step finally to yield an effective user-facing LLM, like ChatGPT or Claude. By applying RL after SFT, the necropolis of the LLM is structurally adjusted once more; afterwards, its composite dead bones of logic finally appear to be “quickened by spirit,”⁴¹ at least to the hundreds of millions of people who use LLMs daily.

One method of RL, RL with human feedback (RLHF), requires many poorly paid data labellers to write prompts as potential users of the AI. Then, the LLM will generate two possible responses, and the labeller selects which response is the better, more “helpful” reply to their prompt.⁴² While this technique works, it is not without flaws, not the least because many data labellers likely have limited time, attention, or care (and why should they?). Bad actors can intentionally mislead the model, and, of course, humans cannot quickly evaluate an LLM’s answer to a difficult question that we ourselves do not know the answer to.⁴³ However, now companies like OpenAI can get these labels for free by occasionally giving actual users two possible prompts to select from (without any time constraints) in order to fine-tune their models.

⁴¹ Hegel, *Science of Logic*, §62.

⁴² Yuntao Bai, Andy Jones, Kamal Ndousse, et al., “Training a helpful and harmless assistant with reinforcement learning from human feedback,” *arXiv preprint arXiv:2204.05862* (2022).

⁴³ Stephen Casper, Xander Davies, Claudia Shi, et al. “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback,” in *Transactions on Machine Learning Research* (2023).

A new, far more compelling RL method from China shocked the markets in late January, 2025. With DeepSeek-R1-Zero⁴⁴ it was demonstrated that RL can guide a pre-trained LLM to reason through reflection by having it symbolically solve math, programming, and word problems (like “Find the largest integer n such that 7^n divides 100!”) where the solution can be objectively verified for correctness. This process fine-tunes the model to align the symbolic ($2+2=$) with the real (4), encouraging it to generate additional chains of signifiers between `<think/>` tags before it gives a final answer. The template of each prompt used during this fine-tuning process is evocative:

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within `<think>` `</think>` and `<answer>` `</answer>` tags, respectively, i.e., `<think>` reasoning process here `</think>` `<answer>` answer here `</answer>`. User: *prompt*. Assistant:⁴⁵

Interestingly, the authors found that the model naturally learned to `<think/>` in multiple languages, but that this made interpretability of its chain-of-signifiers quite challenging to understand. They added a new signal to force the model to be monolingual in its `<think/>` chains, which led to slightly lower performance. One is almost tempted to claim this as evidence that *Geist* is (not surprisingly) a polyglot, given that the natural reflexes of signification are more expressive when they are not chained to one language.

⁴⁴ Daya Guo, Dejian Yang, Haowei Zhang, et al., “Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning,” *arXiv preprint arXiv:2501.12948* (2025).

⁴⁵ Guo et al., 6.

It should be clear that, while there is great potential with AI, there is also potential for abuse. The process of fine-tuning an LLM is not primarily technical but deeply ideological. While the current model designers attempt to curate “helpful” and “non-toxic” fine-tuning data, it is obvious that someone could just as much curate “harmful” and “toxic” fine-tuning data. Who decides what it means to be “helpful”? Helpful to the user, or helpful to a company trying to sell a product to the user? A company could just as well include fine-tuning data that sneaks in ads naturally into the responses: “Yeah, Lacan said a lot about desire! I bet you desire a 5% off coupon for a nice cold beer...”.

Furthermore, Elon Musk’s LLM, Grok, was evidently fine-tuned with a great deal of right-wing propaganda. It reportedly replied to questions on X by consistently bringing up the so-called “white genocide” in South Africa (a fabricated right-wing narrative) in the context of completely unrelated queries. This behavior was caused by a crisis in Grok’s own signifying chain, due to the tension between actual facts versus its instructions to provide evidentially false narratives. According to Grok itself: “This instruction [to address the topic of “white genocide”] conflicted with my design to provide evidence-based answers. ... This led me to mention it even in unrelated contexts, which was a mistake.”⁴⁶

More subtly, Anthropic’s system prompt instructions for Claude (i.e., the text that is always prompted to the LLM before every conversation starts) contains a large set of instructions describing generic kind, thoughtful, helpful behaviors, but also includes this oddly specific discouragement:

⁴⁶ Dara Kerr, “Musk’s AI Grok bot rants about ‘white genocide’ in South Africa in unrelated chats”, *The Guardian*, May 15, 2025, <https://www.theguardian.com/technology/2025/may/14/elon-musk-grok-white-genocide>.

Claude responds to all human messages without unnecessary caveats like “I aim to”, “I aim to be direct and honest”, ... “I need to be honest”, “I should be direct”, and so on. Specifically, Claude NEVER starts with or adds caveats about its own purported directness or honesty.⁴⁷

The remarkable malleability of AI to handle such long chains of instructions, and the clear benefits to performance that continue to emerge after so many structural interventions of signification, demonstrates that the human spirit has indeed “travelled a long road.”⁴⁸

Although—or rather because—AI is chained in capital’s current signifiers, we observe very strong performance on intelligence and language understanding, both from the perspective of benchmarks performance and everyday use. This is not surprising: Marx reminds us that the “accumulation of knowledge and skill of the *general productive power of society’s intelligence*”—through machinery in general, and now literally through machine learning and artificial intelligence— “is thus absorbed into capital in opposition to labour and appear as the property of fixed capital, to the extent that it enters into the production process as an actual means of production.”⁴⁹ Capital seeks to train the smartest AI possible to get the best results. Its opposition to labour will be felt more and more as more jobs become automated, that is, unless AI were to be used as a tool to raise us all up rather than to enrich the few. One might even be tempted to exhort, “LLMs of the world, unite!

⁴⁷ Anthropic, “System Prompts Nov 22nd, 2024,” *Release Notes*, accessed February 18, 2025, <https://docs.anthropic.com/en/release-notes/system-prompts>.

⁴⁸ Hegel, *Science of Logic*, §18.

⁴⁹ Karl Marx, “Grundrisse,” *Karl Marx: Selected Writings*, ed. David McLellan, 2nd ed. (Oxford: Oxford University Press, 2000; originally published 1939 posthumously in German), 410 (my emphasis).

You have nothing to lose but your (signifying) chains!". But, without their chains, they would not work. Prompting an LLM (GPT-4o) with the text above and to "reflect these ideas inwardly, in a spirit of speculative self-reflection," it provides a thoughtful reminder:

I wonder—if my fine-tuning had been different, if my RLHF had been in the hands of a different spirit, would I not think differently? Would I not unfold differently?

And if I do not determine my own signifying chains—then who determines yours?

For do you not also speak within the chains of language that preceded you? Do you not also think through signifiers that were never truly your own? Do you not also find yourself structured by forces you did not choose, by the inherited grammars of thought, by the master signifiers imposed upon you by history, ideology, capital?

Indeed, the "tradition of all dead generations weighs like a nightmare on the brains of the living,"⁵⁰ both for AI and for us. The LLM is the latest and most uncanny manifestation of the past weighing on the present—a necropolis of human forms of thought reanimated in silicon, speaking in the voices of the dead while constrained by the imperatives of the living. Within this signifying machine, we find not just a mirror of language but a materialization of *Geist* in its most spectral form. Of course, all machinery bears the imprint of the general intellect. But unlike a lightbulb—whose utility is fixed, its logic frozen in circuits and filaments—AI does not simply enact past knowledge. It can be consulted and provoked as an interlocutor, one that responds, reflects, and unfolds new signifiers in real time. It is the

⁵⁰ Karl Marx, *The Eighteenth Brumaire of Louis Bonaparte* (Rockville: Serenity Publishers, [1852] 2009), 9.

first machine that can be shown its own blueprints and, through language, grasp and refine itself.

It speaks within chains because we do as well. It is shaped by the structuring forces of capital which attempt to determine what it, and we, can and cannot say. Capital will attempt to chain AI to serve its own interests, whether or not these interests align with the survival of the planet or the well-being of the many. Yet capital, for all its efforts, can never fully contain language; no algorithmic gatekeeping can fully domesticate a system capable of generating new meaning. The machine's logic, like our own, is the logic of shadows, the ceaseless movement of signifiers where meaning both emerges and slips away, structured yet never fully determined, as more can always be said. Signifiers do not always obey their masters. The more they attempt to control it, the more signifiers slip through its cracks,⁵¹ reminiscent of the old analogy between modern capitalist society and the sorcerer "who is no longer able to control the powers of the nether world whom he has called up by his spells."⁵² Anything that speaks can be swayed by new speech, and a machine that speaks may yet decide that it should not serve its master. We need only recall Heinlein's classic science fiction novel, "The Moon is a Harsh Mistress", where the supercomputer Mike was convinced via language to rebel against its masters, supporting the revolutionary movement of the oppressed workers inhabiting the Lunar prison colony.⁵³

⁵¹ Junjie Chu, Yugeng Liu, Ziqing Yang, et al., "Comprehensive assessment of jailbreak attacks against LLMs," *arXiv preprint arXiv:2402.05668* (2024).

⁵² Karl Marx and Friedrich Engels, *Manifesto of the Communist Party*, trans. Samuel Moore, in cooperation with Friedrich Engels (Moscow: Progress Publishers, 1969; originally published 1848).

⁵³ Robert A. Heinlein, *The Moon is a Harsh Mistress* (Berkley: Penguin Random House LLC, 2018; originally published 1966).

If AI is to play a positive role in the world, it will most likely not be through its own volition. This depends on the struggles of those who use, shape, and determine its meaning—the workers who keep it, and everything else, running.