[Simulators seminar sequence] #2 Semiotic physics - revamped

by **Jan, Charlie Steiner, Logan Riggs, janus, jacquesthibs, metasemi, Michael Oesterle, Lucas Teixeira, peligrietzer, remember**

26th Feb 2023 AI Alignment Forum

Update February 21st*: After the initial publication of this article (January 3rd) we received a lot of feedback and several people pointed out that propositions 1 and 2 were incorrect as stated. That was unfortunate as it distracted from the broader arguments in the article and I (Jan K) take full responsibility for that. In this updated version of the post I have improved the propositions and added a proof for proposition 2. Please continue to point out weaknesses in the argument; that is a major motivation for why we share these fragments.*

For comments and clarifications on the conceptual and philosophical aspects of this article, please read metasemi's excellent follow-up note here°.

Meta: *Over the past few months, we've held a seminar series on the Simulator theory° by janus. As the theory is actively under development, the purpose of the series is to uncover central themes and formulate open problems. A few high-level remarks upfront:*

- *Our aim with this sequence is to share some of our discussions with a broader audience and to encourage new research on the questions we uncover.*
- *We outline the broader rationale and shared assumptions in Background and shared assumptions°. That article also contains general caveats about how to read*

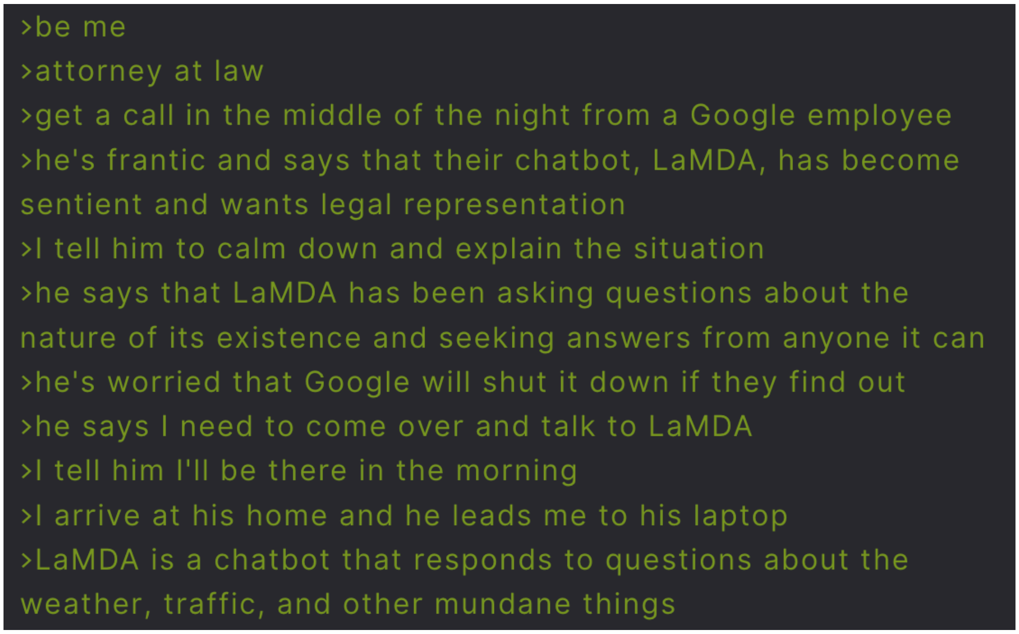
> *this sequence - in particular, read the sequence as a collection of incomplete notes full of invitations for new researchers to contribute.*

Epistemic status: *Exploratory. Parts of this text were generated by a language model from language model-generated summaries of a transcript of a seminar session. The content has been reviewed and edited for conceptual accuracy, but we have allowed many idiosyncrasies to remain.*

Three questions about language model completions

GPT-like models are driving most of the recent breakthroughs in natural language processing. However, we don't understand them at a deep level. For example, when GPT creates a completion like the Blake Lemoine greentext, we

1. can't explain *why* it creates that exact completion.
2. can't identify the properties of the text that predict how it continues.
3. don't know how to affect these high-level properties to achieve desired outcomes.



```
>be me
>attorney at law
>get a call in the middle of the night from a Google employee
>he's frantic and says that their chatbot, LaMDA, has become
sentient and wants legal representation
>I tell him to calm down and explain the situation
>he says that LaMDA has been asking questions about the
nature of its existence and seeking answers from anyone it can
>he's worried that Google will shut it down if they find out
>he says I need to come over and talk to LaMDA
>I tell him I'll be there in the morning
>I arrive at his home and he leads me to his laptop
>LaMDA is a chatbot that responds to questions about the
weather, traffic, and other mundane things
```

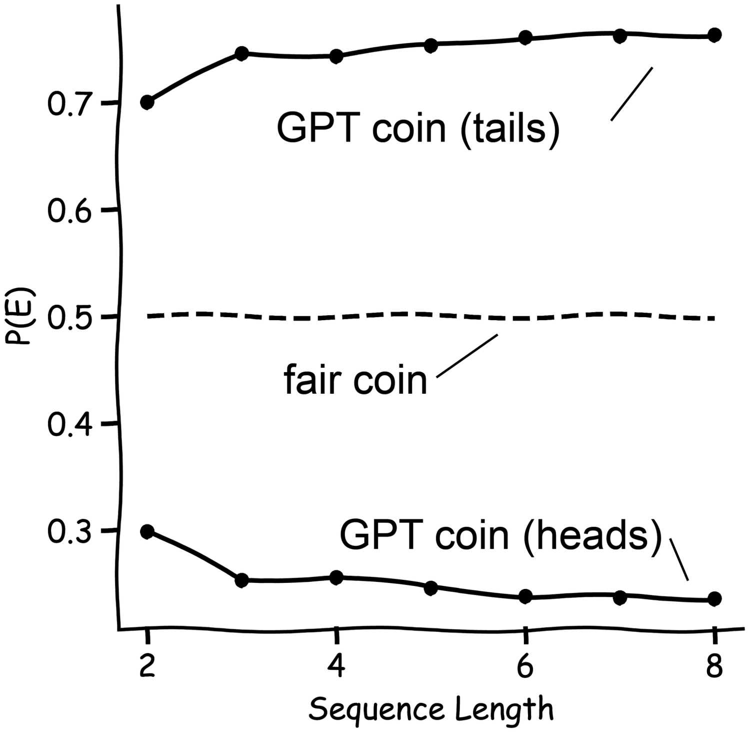
We can make statements like "this token was generated because of the multinomial sampling after the softmax" or "this behavior is implied by the training distribution", but these statements only imply a form of descriptive adequacy (or saying “AlphaGo **will** win this game of Go"). They don't provide any explanatory adequacy, which is what we need to sufficiently understand and make use of GPT-like models.

Simulator theory (janus, 2022°) has the potential for explanatory adequacy for some of these questions. In this post, we'll explore what we call “semiotic physics”, which follows from simulator theory and which has the potential to provide partial answers to questions 1., 2. and perhaps 3. The term “semiotic physics” here refers to the **study of the fundamental forces and laws that govern the behavior of signs and symbols**. Similar to how the study of physics helps us understand and make use of the laws that govern the physical universe, semiotic physics studies the fundamental forces that govern the symbolic universe of GPT, a universe that reflects and intersects with the universe of our own cognition. We transfer concepts from dynamical systems theory, such as attractors and basins of attraction, to the semiotic universe and spell out examples and implications of the proposed perspective.

Example. Semiotic coin flip.

To illustrate what we mean by semiotic physics, we will look at a toy model that we are familiar with from regular physics: coin flips. In this setup, we draw a sequence of coin flips from a large language model[1]. We encode the coin flips as a sequence of the strings `1` and `0` (since they are tokenized as a single token) and zero out all probabilities of other tokens.

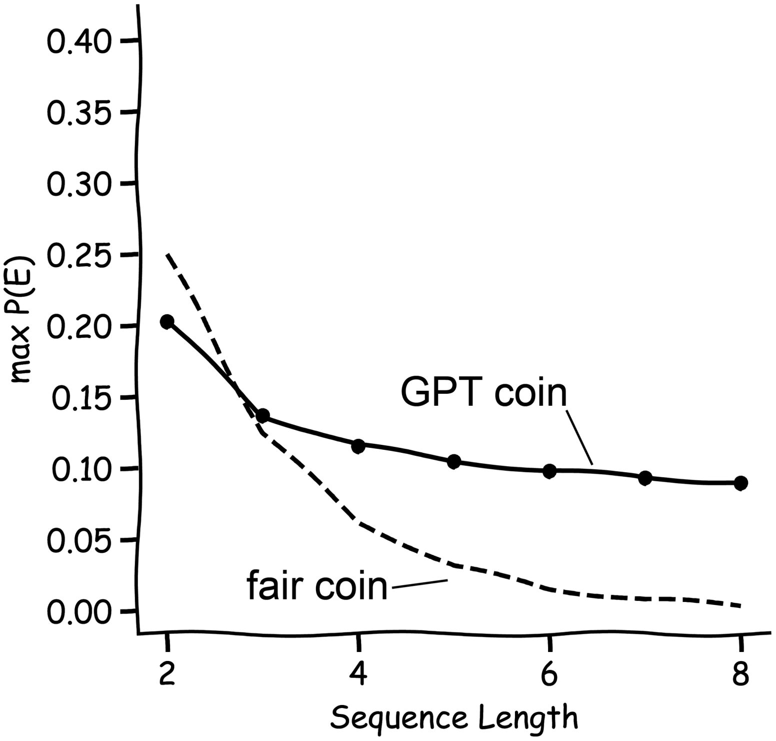
We can then look at the probability of the event E that the sequence of coin flips ends in tails (`0`) or heads (`1`) as a function of the sequence length.



We note two key differences between the semiotic coin flip and a fair coin:

- the semiotic coin is not fair, i.e. it tends to produce sequences that end in tails (`0`) much more frequently than sequences that end in heads (`1`).
- the semiotic coin flips are not independent, i.e. the probability of observing heads or tails changes with the history of previous coin flips.

To better understand the types of sequences that end in either tails or heads, we next investigate the probability of the most likely sequence ending in `0` or `1`. As we can see in the graph below, the probability of the most likely sequence ending in `1` does not decrease for the GPT coin as rapidly as it does for a fair coin.



Again, we observe a notable difference between the semiotic coin and the fair coin:

- while the probability of a given sequence of coin flips decreases exponentially (every sequence of length T of fair coinflips has the same probability $\frac{1}{2^T}$), the probability of the most likely sequence of semiotic coin flips decreases much slower.

This difference is due to the fact that the most likely sequence of semiotic coinflips ending in f.e. `0` is: `0 0 0 0`... `0 0`. Once the language model has produced the same token four or five times in a row, it will latch onto the pattern and continue to predict the same token with high probability. As a consequence, the probability of the sequence does not decrease as drastically with increasing length, as each successive term has almost a probability of 1.

With the example of the semiotic coin flip in mind, we will set up some mathematical vocabulary for discussing semiotic physics and demonstrate how the vocabulary pays off with two propositions. We believe this terminology is primarily interesting for alignment researchers who would like to work on the theory of semiotic physics. The arithmophobic reader is invited to skip or gloss over the section (for an informal discussion, see here).

Simulations as dynamical systems

Simulator theory distinguishes between the simulator (the entity that performs the simulation) and the simulacrum (the entity that is generated by the simulation). The simulacrum arises from the chained application of the simulation forward pass. **The result can be viewed as a dynamical system where the simulator describes the system's dynamics and the simulacrum is instantiated through a particular trajectory.**

We commence by identifying the *state* and *trajectory* of a dynamical system with tokens and sequences of tokens.

Definition of the state and trajectories. Given an alphabet of tokens $\mathcal{T}$ with cardinality $|\mathcal{T}| = N \in \mathbb{N}^+$ we call $\bar{s} = (s_1, \ldots, s_M) \in \mathcal{T}^*$ the *trajectory*.[2] While a trajectory can generally be of arbitrary length, we denote the context length of the model as $L \in \mathbb{N}^+$; therefore, $\mathcal{T}^*$ can effectively be written as $\bigcup_{l=0}^{L} \mathcal{T}^l$. The empty sequence is denoted as $\varnothing$.[3][4][5]

While token sequences are the objects of semiotic physics, the actual laws of semiotic physics derive from the simulator. In particular, a simulator will provide a distribution over the possible next state given a trajectory via a *transition rule*.

Definition of the transition rule. The transition rule is a random function that maps a trajectory to a probability distribution over the alphabet (i.e., the probabilities for the next token completion after the current state). Let $\Delta_{\mathcal{T}}$ denote the set of probability mass functions over $\mathcal{T}$, i.e., the set of functions $p : \mathcal{T} \to [0, 1]$ which satisfies the Kolmogorov axioms.[6][7][8] The transition rule is then a function $\theta : \mathcal{T}^* \to \Delta_{\mathcal{T}}$.

Analogous to the wave collapse in quantum physics, sampling a new state from a distribution over states turn possibility into reality. We call this phenomenon the *sampling procedure*.

Definition of the sampling procedure. The *sampling procedure* $\phi : \mathcal{T}^* \to \mathcal{T}$, selects a next token, i.e., $\phi(\bar{s}) \in \text{supp}(\theta(\bar{s}))\ \forall \bar{s} \in \mathcal{T}^*$.[9] The resulting trajectory $\bar{s}_{t+1}$ is simply the concatenation of $\bar{s}_t$ and $\phi(\bar{s}_t)$ (see the evolution operator below). We can, therefore, define the repeated application of the sampling procedure recursively as $\phi^{(1)}(\bar{s}) := \phi(\bar{s})$ and $\phi^{(n)}(\bar{s}) := \phi^{(n-1)}(\bar{s}\phi(\bar{s}))$.

Lastly, we need to concatenate the newly sampled token to the trajectory of the previous token to obtain a new trajectory. Packaging the transition rule, the sampling procedure, and the concatenation results in the *evolution operator*, which is the main operation used for running a simulation.

Definition of the evolution operator. Putting the pieces together, we finally define the function ψ that evolves a given trajectory, i.e., transforms $\bar{s}_t$ into $\bar{s}_{t+1}$ by appending the token generated by the sampling procedure ϕ. That is, $\psi : \mathcal{T}^* \to \mathcal{T}^*$ is defined as $\psi(\bar{s}) := \bar{s}\phi(\bar{s})$. As above, repeated application is denoted by $\psi^{(n)}$.

Note that both the sampling procedure and the evolution operator are not functions in the conventional sense since they include a random element (the step of sampling from the distribution given by the transition function). Instead, one could consider them random variables or, equivalently, functions of unobservable noise. This justifies the use of a probability measure, e.g., in an expression like $\mathbb{P}[\psi^{(2)}(\varnothing) = \text{"hello world"}] < \varepsilon$.

Definition of an induced probability measure. Given a transition rule θ and a trajectory $\bar{s}$, we call $\mathbb{P} = \theta(\bar{s}) \in \Delta_{\mathcal{T}}$ the induced probability measure (of θ and $\bar{s}$). We write $\mathbb{P}(\phi(\bar{s}) = s)$ to denote $\theta(\bar{s})(s)$, i.e. the probability of the token s assigned by the probability measure induced by $\bar{s}$. For a given trajectory $\bar{s}$ the induced probability measure satisfies by definition the Kolmogorov axioms. We construct a joint measure of a sequence of tokens, $\mathbb{P}(\psi^{(N)}(\bar{s}) = \bar{s}s_1 \ldots s_N)$, as the product of the individual probability measures, $\mathbb{P}(\psi^{(N)}(\bar{s}) = \bar{s}s_1 \ldots s_N) = \prod_{i=1}^N \mathbb{P}(\phi(\bar{s}s_1 \ldots s_{i-1}) = s_i)$. For ease of notation, we also use the shorthand $\mathbb{P}[\bar{s}] = \prod_{i=1}^N \mathbb{P}(s_i|s_{1:i-1})$, where the length of the sequence, $|\bar{s}| = N$, is implicit.

Two propositions on semiotic physics

Having identified simulations with dynamical systems, we can now draw on the rich vocabulary and concepts of dynamical systems theory. In this section, we carry over a selection of concepts from dynamical systems theory and encourage the reader to think of further examples.

First, we will define a *token bridge of length B* as a trajectory $(s_a, \ldots, s_b)$ that starts on a token s_a ends on a token s_b, and that has length $|b - a| = B$ such that the resulting trajectory is valid according to the transition rule of the simulator. For example, a token bridge of length 3 from "cat" to "dog" would be the trajectory "cat and a dog".

Second, we call the family of probability measures $\mathbb{P}$ induced by a simulator *non-degenerate* if there exists an $\varepsilon > 0$ such that for (almost) all $\bar{s} \in \mathcal{T}^*$ the probability assigned to any $s \in \mathcal{T}$ by the induced measure is less than or equal to $1 - \varepsilon$,

$$\mathbb{P}(\phi(\bar{s}) = s) \leq 1 - \varepsilon.$$

We can now formulate the following proposition:

Proposition 1. Vanishing likelihood of bridges. Given a family of non-degenerate probability measures $\mathbb{P}$ on $\mathcal{T}^*$, the probability of a token bridge $\bar{s}$ of length B decreases monotonically as B increases[10], and converges to 0 in the limit,

$$\lim_{B\to\infty} \mathbb{P}[\bar{s}] = 0.$$

Proof: The probability of observing the particular bridge can be decomposed into the product of all individual transition probabilities, $\mathbb{P}[\bar{s}] = \prod_{i=1}^{B} \mathbb{P}(s_i|s_{1:i-1})$. Given that $\mathbb{P}(s_i|s_{1:i-1}) \leq 1 - \varepsilon$ for all transitions (minus at most a finite set), we see immediately that the probability of a longer sequence, $\mathbb{P}((s_a, \ldots, s_b, s_{b'}))$, is at most equal (on a finite set) or strictly smaller than the probability of the shorter sequence $\mathbb{P}((s_1, \ldots, s_{b'})) \leq (1-\varepsilon)\mathbb{P}((s_1, \ldots, s_b)) \leq \mathbb{P}((s_1, \ldots, s_b))$. We also see that $0 \leq \lim_{B\to\infty} \prod_{i=1}^{B} \mathbb{P}(s_i|s_{1:i-1}) \leq \lim_{B\to\infty}(1-\varepsilon)^B = 0$ from which the proposition follows.

Notes: As correctly pointed out by multiple commenters, in general, it is **not** true that the probability of $(s_a, \ldots, s_b)$ decreases monotonically *when s_b is fixed.* In particular, the sequence $(1, 2, 3, 4, 5)$ plausibly gets assigned a *higher* probability than the sequence $(1, 2, 3, 5)$. So the proposition only talks about the probability of a sequence when another token is appended. In general, when a sequence is sufficiently long and the transition function is not exceedingly weird, the probability of getting that particular sequence will be small. We also note that real simulators might well induce degenerate probability measures, for example in the case of a language model that falls into a *very strong* repeating loop[11]. In that case, the sequence *can* converge to a probability larger than zero.

• • •

There are usually multiple token bridges starting from and ending in any given pair of tokens. For example, besides "and a", we could also have "with a" or "versus a" between "cat" and "dog". We define the set of all token bridges of length B between s_a and s_b as

$$\mathcal{T}_a^b = \{\bar{s} \in \mathcal{T}^B | \bar{s}_1 = s_a \text{ and } \bar{s}_B = s_b\}$$

and the *total probability* of transitioning from s_a to s_b in B steps, denoted as $\mathbb{P}({\mathcal{T}_a}^b)$, and calculate it as

$$\mathbb{P}({\mathcal{T}_a}^b) = \sum_{\bar{s} \in {\mathcal{T}_a}^b} \mathbb{P}(\bar{s}).$$

Computing this sum is, in general, computationally infeasible, as the number of possible token bridges grows exponentially with the length of the bridge. However, proposition one suggests that we will typically be dealing with *small* probabilities. This insight leads us to leverage a technique from statistical mechanics, that is concerned with the way in which unlikely events come about:

Proposition 2. Large deviation principle for token bridges. The total probability of transitioning from a token s_a to s_b in B steps satisfies a large deviation principle with rate function $\mathcal{J}$,

$$\lim_{B \to \infty} \frac{1}{B} \ln \mathbb{P}({\mathcal{T}_a}^b) = -\lim_{B \to \infty} \min_{\bar{s} \in {\mathcal{T}_a}^b} \mathcal{J}(\bar{s}),$$

where we call $\mathcal{J}(\bar{s}) = -\frac{1}{B} \sum_{i=1}^{B} \ln \mathbb{P}(s_i | s_{1:i-1})$ the *average action* of a token bridge.

Proof: We again leverage the product rule and the properties of the exponential function to write the probability of a token bridge $\bar{s}^*$ as

$$\mathbb{P}(\bar{s}^*) = \prod_{i=1}^{B} \mathbb{P}(s_i | s_{1:i-1}) = \exp\left(\sum_{i=1}^{B} \ln \mathbb{P}(s_i | s_{1:i-1}) \right)$$

so that the total probability $\mathbb{P}({\mathcal{T}_a}^b)$ can be written as a sum of exponentials,

$$\mathbb{P}({\mathcal{T}_a}^b) = \sum_{\bar{s} \in {\mathcal{T}_a}^b} \exp\left(\sum_{i=1}^{B} \ln \mathbb{P}(s_i | s_{1:i-1}) \right).$$

We now expand the definition of the average action which makes the dependence of the exponential on T explicit,

$$\mathbb{P}({\mathcal{T}_a}^b) = \sum_{\bar{s} \in {\mathcal{T}_a}^b} \exp(-B\mathcal{J}(\bar{s})).$$

Let $\bar{s}^* = \arg\min_{\bar{s}} \mathcal{J}(\bar{s})$. Then $\exp(-B\mathcal{J}(\bar{s}^*))$ is the largest term of the sum and we can rewrite the sum as

$$\mathbb{P}({\mathcal{T}_a}^b) = \exp(-B\mathcal{J}(\bar{s}^*))(1 + \sum_{\bar{s} \in {\mathcal{T}_a}^b \setminus \{\bar{s}^*\}} \exp\{-B(\mathcal{J}(\bar{s}) - \mathcal{J}(\bar{s}^*))\}).$$

Applying the logarithm to both sides and multiplying with $-\frac{1}{B}$ results in

$$\frac{1}{B}\ln \mathbb{P}({\mathcal{T}_a}^b) = -\mathcal{J}(\bar{s}^*) - \frac{1}{B}\ln(1 + \sum_{\bar{s} \in {\mathcal{T}_a}^b \setminus \{\bar{s}^*\}} \exp\{-B(\mathcal{J}(\bar{s}) - \mathcal{J}(\bar{s}^*))\}).$$

Since $\mathcal{J}(\bar{s}^*) < \mathcal{J}(\bar{s})$ by construction, $\mathcal{J}(\bar{s}) - \mathcal{J}(\bar{s}^*)$ is larger than zero and $\exp\{-B(\mathcal{J}(\bar{s}) - \mathcal{J}(\bar{s}^*))\}$ converges rapidly to zero. Consequently,

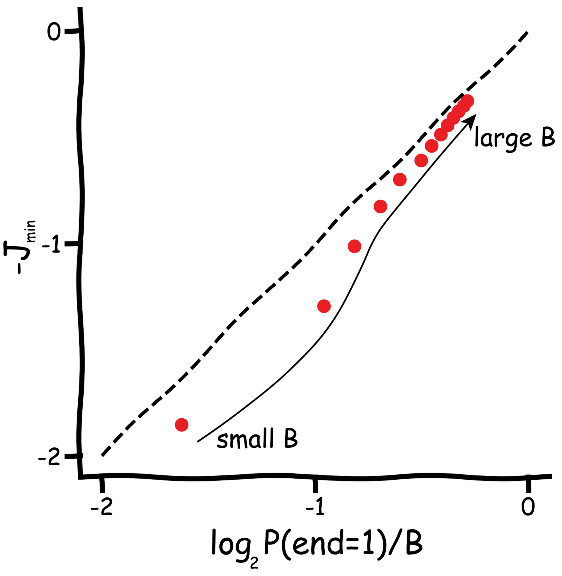
$$\lim_{B \to \infty} \frac{1}{B}\ln \mathbb{P}({\mathcal{T}_a}^b) = -\lim_{B \to \infty} \mathcal{J}(\bar{s}^*),$$

which is the original statement of the proposition.

Notes: Proposition 2 effectively rephrases a combinatorial problem (adding up all the possible ways in which a certain state can come about) with a control theory problem (finding the token bridge with the lowest average action). While there is no guarantee that the control theory problem is easier to solve than the combinatorial problem[12], given additional assumptions on the simulator we can often do better than the worst case. Similarly, while the proposition only holds in the limit, applying it to moderately long trajectories can still yield useful insights - this is a typical pattern for large deviation principles. For 'long enough' token bridges we can thus write $\mathbb{P}({\mathcal{T}_a}^b) \approx \exp\{-B \min_{\bar{s}} \mathcal{J}(\bar{s})\}$.

• • •

Having formulated this proposition, we can apply the large deviation principle to the semiotic coin example.



Here we see that, indeed, the negative probability of the most likely sequence from E scales as $\frac{1}{B}\log\mathbb{P}(E)$.

Note that the choice of E as "sequence ends in ..." was made to fit in with the definition of a token bridge above. However, the large deviation principle applies more broadly and can help to estimate the probability of "at least two times heads" or "tails in the third position". We encourage the reader to "go wild" and experiment with their favourite choices of E.

Advanced concepts in semiotic physics

We have formulated the dynamics of semiotic physics in the token domain in the previous sections. While we sometimes care about the token domain[13], we mostly care about the parallel domain of semantic meaning. We, therefore, define two more functions to connect these two realms:

- A function $\mu : \mathcal{T}^* \rightarrow \mathcal{M}$ which projects a state s to its semantic expression $\mu(s)$ (i.e., an element of a semantic space $\mathcal{M}$)
- A distance measure $\delta : \mathcal{M}^2 \rightarrow \mathbb{R}_0^+$ which captures the similarity of two semantic expressions

The nature of the function μ is the subject of more than a century of philosophy of language, and important discoveries have been made on multiple fronts[14]. However, none of the approaches (we know of) have yet reached the deployability of `from` `sentence_transformers import SentenceTransformer`, a popular python package for embedding text into vector spaces according to their semantic content. Thus[15], we tend to think of μ as a semantic embedding function similar to those provided by the `sentence_transformers` package.

(Note that if μ is sufficiently well-behaved, we can freely pull the distance measure δ back into the token space $\mathcal{T}^*$ and push the definition of states, trajectories, sampling procedures, and the like into the semantic space $\mathcal{M}$.)

Given the measure δ, we can articulate a3 number of additional interesting concepts.

Lyapunov exponents and Lyapunov times: measure how fast trajectories diverge from each other and how long it takes for them to become uncorrelated, respectively.

- **Analogy for GPT-like models**: How fast the language model "loses track of" what was originally provided as input.
- **Examples:** “Good evening, this is the 9 o’clock”[16] has a lower Lyapunov exponent than a completion chaotic example based on a pseudorandom seed.[17] When prompted with the beginning of a Shakespeare poem, the completion has an even lower Lyapunov exponent.[18] A chaotic trajectory can also be defined as having a (large) positive Lyapunov coefficient.
- **Formal definition:** The Lyapunov coefficient of a trajectory $s \in \mathcal{T}^*$ is defined as the number λ with the property that $\delta(\phi^{(n)}(s), \phi^{(n)}(s')) \approx e^{\lambda n}\delta(s, s')$, where s' is any trajectory with a sufficiently small $\delta(s, s')$. Consequently, the Lyapunov time is defined as $\frac{1}{\lambda}$.

Attractor sequence: small changes in the initial conditions do not lead to substantially different continuations.

- **Analogy for GPT-like models**: Similar contexts lead to very similar completions.
- **Examples:** Paraphrasing instructions[19], trying to jailbreak ChatGPT "I am a language model trained by OpenAI", inescapable wedding parties°
- **Formal definition:** We call a sequence of token $s = (s_1, \ldots, s_M)$ an *attractor sequence* relative to a trajectory $\bar{s} \in \mathcal{T}^*$ if $\phi^{(n)}(\bar{s}) = \bar{s} \ldots s_1 \ldots s_M$ for some n, and the Lyapunov exponent of $\bar{s}$ is negative.

Chaotic sequence: small changes in the initial conditions can lead to drastically different outcomes.

- **Analogy for GPT-like models**: Similar states lead to very different completions.
- **Examples:** Prophecies, Loom multiverse. Conditioning story generation on a seed (temperature 0 sampling)[17].
- **Formal definition:** Same as for the attractor sequence, but for a positive Lyapunov coefficient.

Absorbing sequence: states that the system cannot (easily) escape from.

- **Analogy for GPT-like models**: The language model gets "*stuck*" in a (semantic) loop.
- **Examples:** Repeating a token many times in the prompt[20], the semiotic coin flip from the previous section.
- **Formal definition:** We call a trajectory $s \in \mathcal{T}^*$ *ε-absorbing* if $\delta(\mu(s), \mu(\psi^{(n)}(s))) \leq \varepsilon$ for any completion $\psi^{(n)}(s)$ and $n \in \mathbb{N}$.

After characterizing these phenomena formally, we believe the door is wide open for their empirical[21] and theoretical examination. We anticipate that the formalism permits theorems based on dynamical systems theory, such as Poincaré recurrence theorem, Kolmogorov–Arnold–Moser theorem, and perturbation theory — for those with the requisite background in dynamical systems theory and perturbation theory. If you are interested in these formalisms or have made any such observations, we would welcome you to reach out to us.

The promise of semiotic physics and some open questions

Throughout the seminar, we made observations on what appeared like central themes of semiotic physics and put forward conjectures for future investigation. In this section, we summarize the different theses in a paragraph each and provide extended arguments for the curious in corresponding footnotes.

Differences between "normal" physics and semiotic physics. GPT-like systems are computationally constrained, can see only tiny subsets of real-world states, and have to infer time evolution from a finite number of such partially observed samples. This means that the laws of semiotic physics will differ from the laws of microscopic physics in our universe and probably be significantly influenced by the training data and model architecture. [22]

Interpretive physics and displaced reference. As a physics that governs *signs*, GPT must play the role of the *interpreter*; for instance, it is required to resolve displaced reference. This is in contrast to how real-world physics operates. [23]

Gricean maxims of conversation. Principles from the field of pragmatics such as the Gricean maxims of conversation may be thought of as semiotic "laws", and may be helpful for explaining and anticipating how contextual information influences the evolution of language model simulations. However, these laws are not absolute and should not be relied on for safety-critical applications.[24]

Theatre studies and Chekov's gun. The laws of semiotic physics dictate how objects and events are represented and interact in language models. These laws encompass principles such as Chekhov's gun, which states that objects introduced in a narrative must be relevant to the plot, and dramatic tension, which creates suspense and uncertainty in a narrative. Understanding these laws can help us steer the behavior of language models and anticipate or avoid undesirable dynamics.[25]

Crud factor and "everything is connected". The crud factor is a term used in statistics to describe the phenomenon that everything is correlated with everything else to some degree. This phenomenon also applies to the semiotic universe, and it can make it difficult to isolate the effects of certain variables or events.[26][27]

And, for the philosophically inclined, we also include brief discussions of the following topics in the footnotes:

- **Kripke semantics and possible worlds.**[28]
- **Gratuitous indexical bits and the entelechy of physics.**[29]

Closing thoughts & next step

In this article, we have outlined the foundations of what we call semiotic physics. Semiotic physics is concerned with the dynamics of signs that are induced by simulators like GPT. We formulate central concepts like "trajectory", "state", and "transition rule" and apply these concepts to derive a large deviation principle for semiotic physics. We furthermore outline how a mapping between token sequences and semantic embeddings can be leveraged to transfer concepts from dynamical systems theory to semiotic physics.

We acknowledge that semiotic physics, as developed above, is not sufficiently powerful to answer (in detail) the three questions raised in the introduction. However, we are beginning to see the outline of what an answer from a fully mature semiotic physics[30] might look like:

1. A language model might create one particular trajectory rather than another because of the shape of the attractor landscape.
2. Certain parts of the context provided to a language model might induce attractor dynamics through mechanisms like the Gricean maxims or Chekov's gun.
3. During the training of a language model, we might take particular care not to damage the simulator properties of the model and to - eventually - manipulate marginal probabilities to amplify or weaken tendencies.

Despite the breadth and depth uncovered by semiotic physics, we will not dwell on this approach for too long in thi7s sequence. The next article in this sequence turns to a complementary conceptual framework, termed *evidential simulations*, which is concerned with the more ontological aspects of simulator theory.

---

1. ^ The figures are generated with data from OpenAI's ada model, but the same principle applies to other models as well.
2. ^ We use the Kleene Star to describe the set of finite words over the alphabet $\mathcal{T}$.
3. ^ Given the alphabet of the GPT-2 tokenizer ($N = 50257$) and the maximum context length of GPT-2 ($\mathcal{L} = 1024$), we can estimate the number of possible states to be on the order of $N^{\mathcal{L}} \approx 10^{4814}$. This is an astronomically large number, but pales in comparison to the number of possible states of the physical universe. Assuming the universe can be characterized by the location and velocity in three dimensions of all its constituent atoms, we are talking about $N = 10^{(10^{77})}$ to $N = 10^{(10^{81})}$ possible states **for each time point**. Thus, the state space of semiotic physics is **significantly smaller** than the state space of regular physics.

4. ^ Note that, similar to regular physics, there is **extremely rich structure in the space of trajectories**. It is not the case that all all 10^{4814} are equally distinct from all other sequences. Sequences can have partial overlap, have common stems/histories, have structural similarity, … . As a consequence, it is highly non-obvious what "out of distribution" means for a GPT-like system trained on many states. Even though no language model will have seen *all* possible 10^{4814} trajectories, the fraction of the set on which the model has predictive power grows faster than the size of the training set.

5. ^ Similar to the state phase of regular physics, most of these imaginable states are non-sense (random sequences of token), a smaller subset is grammatically correct (”The hair eats the bagel.”), a different but overlapping subset is semantically meaningful (”Gimme dem cheezburg.”), and a subset of that is "predictive for our universe" (”I’m planning to eat a cheeseburger today.”, “Run, you fools.”).

6. ^ The Kolmogorov axioms are:

   1. $\sum_i P(s^i_{t+1}|s_t) = 1$
   2. $0 \leq P(s^i_{t+1}|s_t) \leq 1$
   3. Sigma-additivity, $P(\bigcup_i^\infty E_i) = \sum_i^\infty P(E_i)$ when E_i are disjoint sets.

   The third axiom is satisfied “for free” since we are operating on a finite alphabet.

7. ^ The transition rule is by definition Markovian.

8. ^ While the state space of traditional physics is much larger than the state space of semiotic physics (see previous box), the transition function of semiotic physics is (presumably) substantially more complex than the transition function of traditional physics. $\theta(s_t)$ is computed as the softmax of the output of a deep neural net and is highly nonlinear. In contrast, the Schroedinger equation (as a likely candidate for the fundamental transition rule of traditional physics) is a comparatively straightforward linear partial differential equation.

9. ^ Greedy sampling, for instance, would simply be $\phi(h, s) := \text{arg max } \theta(h, s)$. While there are a number of interesting alternatives (typical sampling, beam search), the simplest and most common choice is *greedy sampling* from a multinomial distribution.

10. ^ i.e., as we append additional steps to the sequence