Category-Theoretic Wanderings into Interpretability

Unruly Abstractions

August 23, 2025

Ian Rios-Sialer[1]

Independent

ORCID: 0009-0001-6970-6058

iansebas@umich.edu

Introduction

It's late already. However, it's still summer, and I am not quite done yet. I type "Does he still think about me?" but then I fancy something more actionable, so I settle on "Should I text him?" instead and press enter. Sometimes ChatGPT says something different from Claude, but they both seem annoyingly aligned on this matter. Can you guess their answer? They both said 'don't'. I trust their instincts, this time. I even delete his number, this time.

[1] iansebastos.github.io

I promise I am normally more shy about sharing such intimacies[2] but chances are you also use your Large Language Models (LLMs) somewhat alike[3]. It's even later in the night, but now I crave to understand. Why did they tell me not to text him? What makes my LLM sidekicks tell me the things they do? Interpretability steps in.

Interpretability can be defined as the ability to explain the inner workings of an AI model to a human in understandable terms [6]. As a pre-paradigmatic field [7], [8], terminology and definitions are a bit confusing and/or inconsistent [9], [10], [11]. A central desideratum of interpretability is to contribute meaningfully to AI Alignment [12], AI Control [13], and AI Safety [14], [15]. The hope [16] is that the more we understand AI systems, the more we can ensure they work the way we want them to.[4]

In this piece, I spill some developing ideas of how category theory [19] frames interpretability, and where our imagination can go with category-theoretic thinking. As the title suggests, these "wanderings" have rather speculative epistemic status, and they are not meant to present clear contributions (yet). Rather, I invite you to "think together" with me in an always-in-process [20] collaboration.

This is a conceptual exploration: I define a few categorical objects and relate them to interpretability. Instead of proofs and empirical validation to close arguments, I provide conjectures to open up alternative intuitions in you. And in between formalisms, I sometimes drop in confessions to open new empathies in you, too. Understandability (and thus also interpretability) always has something contingent, something relational, something personal.

Section 1 (Why Category Theory?) will explain why we are looking at the specific intersection of interpretability and category theory. Section 2 (What is a category?) provides a gentle introduction to category-theoretic thinking. Section 3 (LLMs as categories?) starts our exploration by looking at a [0,1]-enriched category (L_{syn}) that is defined for every LLM. In Section 4 (Looking for meaning through syntax), we realize we will need to compare things to interpret them, so we wonder if we could use (L_{syn}) to make insightful comparisons between LLMs. In Section 5 (Framing Interpretability), we get more serious and formulate what interpretability is. Section 6 (Decomposing Faithfulness) **breaks down faithfulness, the core technical problem in interpretability**. Section 7 (Interpreting Circuit Tracing) applies the developed concepts to *Circuit Tracing* work by Anthropic. Section 8 (Wayfinding) closes the piece with a reflection.

[2] We often start conversations in the context of future Artificial Super Intelligence (ASI) and its risks. Instead, I would like to first ground ourselves in what is familiar, felt [1] and present; hopefully inviting many more different types [2] of people to the table [3], to engage [4] with the development of a technology that will profoundly impact them sooner than they think.

[3] Personal support (emotional application), like therapy, was the top use case of Gen AI in 2025 [5]

[4] There is a lot more to say [17], [18] on how interpretability fits in the big picture, but I'll follow up on that in future writing

If you are in a hurry, you might want to go directly through Sections 5-7 for the technical juice. I do hope you stay around for longer.

1. Why Category Theory?

Category theory is an **abstract** formal language to study **structures and their relations** [21], [22]. It is a mathematical theory and a technical framework, but it can also be seen as a "way of thinking about thinking" [23]. To **think categorically**[5] is to [23], [24], [25]:

- use intuition to find an interesting structure
- pry on the how/why of that structure
- define its context via relationships
- look for similar structure(s) in other contexts
- reason about in which sense they are similar and different
- form abstractions that unify the structures in precise ways
- consider the new abstractions as possible (part of) structures with relationships themselves
- go back and think about what nuance or details (in the structures/situations) are not captured by the available abstractions
- repeat the process, or form higher-level abstractions to attempt to capture the missing nuance if desired

Why is thinking categorically useful? Because it can serve as a tool[6] to think better by allowing us to manipulate abstractions more freely yet still rigorously. By looking at situations from multiple perspectives and scales, we gain new intuitions and insight.

As a young field, interpretability faces many challenges [8], [26], [27]. I believe category theory could help address many of these challenges. To mention a select few ways:

- **Unifying top-down and bottom-up approaches**: The very abstract nature of Category Theory allows us to "zoom in and out" [23] and reason how macroscopic and microscopic structures relate to each other. In fact, we are already seeing category-theoretic bridges forming to understand Deep Learning Architectures [28][7]. For interpretability, the goal would be to connect Mechanistic Interpretability [15] and Representation Engineering [29]. **You can think of this as roughly connecting Neuroscience to Psychiatry/Psychology, linking how neurons work to how we think, feel, and behave.**
- **Develop stronger foundations for decomposition methods**: Compositional Decomposability is needed to reverse-engineer neural-networks [30]. Compositionality requires modular [31] parts with a given interface [32] (with respect to a specific property [33]) that has no emergent/generative effects [34]. Category Theory not only studies the depths of compositionality when present [35], but even allows us to

[5] or "category-theoretically"

[6] A thinking technology?

[7] More specifically, unifying the specification of constraints (often in relation to data) and the specification implementations (like the individual tensor operations)" [28]

measure the failures of compositionality [36]. **Imagine it as learning to cook. Sometimes, the order of ingredients does not matter, and some properties of the resulting dish can be directly inferred from looking at the ingredients (like tossing together a salad, counting calories). Other times, it matters how you combine things (like baking a cake or fermenting bread), where the process changes the outcome entirely.**

- **Inspire the development of new theories**: Although interpretability has allowed us to gain insight into interesting mechanisms [37], the results are limited to specific prompts [38]. Some of our models for interpretability (like Superposition or Linear Representation Hypothesis) and methods (like Sparse Auto Encoders) are falling short in providing us generalizable and applicable interpretations [8], [39], [40], [41]: We need theoretical and conceptual breakthroughs [17], [42]. Thinking categorically helps us reason about complex systems by focusing on the observable relationships between components, without needing full access to their opaque internal workings [25]. Moreover, Category Theory invites us to use formal diagrams, which externalize structure and relevant reasoning [43]. By having more visual (and relational) representations, we could shed light on what has been so far "unthinkable" [44] and form the intuitions needed to progress interpretability. **You can think of this as when Copernicus, looking at the same sky, was inspired to consider other possibilities other than the Earth being at the center of the universe**.

2. What is a category?

There are many good introductory books for Category Theory[8] [19], [23], [24], [45] and some more intermediate ones [34], [46] which I strongly invite you to check out. Instead of going through a list of formal definitions, I will walk you through an unserious but hopefully intuitive example. If you already have the background, feel free to skip this section.

My friends have noticed I sometimes binge on very cringey TV shows. Let's think about this categorically. We start by defining a finite set of all TV shows (available for me to stream from my couch), and a function from each element of such a set to a positive real number that expresses how cringe a show is (based on my friends' opinions):

$$\text{TVShows} := \{\text{White Lotus}, \text{Hacks}, ...\} \qquad \text{Cringe} : \text{TVShows} \to \mathbb{R}^+ \tag{2.1}$$

With that[9], the $\mathsf{CringeShows}$ category consists of the following **Data:**

1. A collection $\text{ob}(\mathsf{CringeShows})$ of *objects*, the elements of the set TVShows

[8] I highly recommend The Joy of Abstraction by Eugenia Cheng [23]

[9] You might have noticed we have formed a partially-ordered set (poset) $(\text{Cringe}(\text{TVShows}), \leq)$

[10] Also called *arrows* or *maps*

2. For every $x, y \in \mathrm{ob}(\mathsf{CringeShows})$, a set of *morphisms*[10] $\mathsf{Hom}_{\mathsf{CringeShows}}(x, y)$[11] , in our case, such set will have single element when $\mathrm{Cringe}(x) \leq \mathrm{Cringe}(y)$ and be the empty set $\emptyset$ otherwise.

With the following **Structure**:

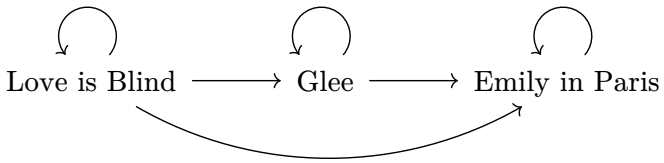
1. An identity morphism for every object, $x \overset{\mathrm{id}_x}{\to} x$ which is $x \leq x$
2. A composition operation such that any two morphisms like $x \overset{f}{\to} y$ and $y \overset{g}{\to} z$ produce $x \overset{g \circ f}{\to} z$, which means $x \leq y$ and $y \leq z$ imply $x \leq z$

With the following **Properties**:

1. Unit Law, such that given $x \overset{f}{\to} y$, $f = 1_y \circ f = f \circ \mathrm{id}_x$
2. Associativity Law, such that $h \circ (g \circ f) = (h \circ g) \circ f$

Let's create a diagram for some of the TV shows:

$$\mathrm{Cringe}(\text{Love is Blind}) \leq \mathrm{Cringe}(\text{Glee}) \leq \mathrm{Cringe}(\text{Emily in Paris}) \tag{2.2}$$



I have drawn identity arrows and compositions, but those are usually not drawn as they can be inferred. Our category is called "thin" because we have at most one arrow between any two objects. Other categories could have many more.

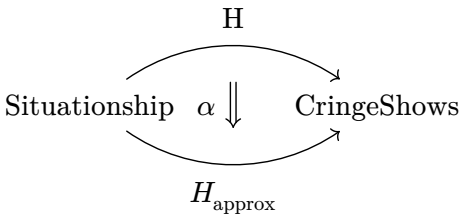
Let's backtrack from the mathematical notation for a second and think about what we have done: We are noticing a structure of interest and characterizing objects by how they fit in this structure. My friends will never watch *Love is Blind* so they are not interested in the intricacies of that reality show. They only pay attention to how cringey a show is in comparison to other shows. Why? My friends know I watch the cringiest shows when my mood is down and I need comfort. They would know something is very wrong if I ever get hooked on *Emily in Paris*.

My friends cannot observe my internal mood directly, but they sometimes hear about the situations in my life. Certain situation(ship)s affect my mood significantly. Let's now consider the Mood category and the $\mathsf{Situationship}$ category.

[11]For readability, for a category C, we will often write the morphisms as $\mathsf{Hom}(x, y)$ or $C(x, y)$ instead of $\mathsf{Hom}_C(x, y)$

$$\text{Situationship} \xrightarrow{\text{F}} \text{Mood} \xrightarrow{\text{G}} \text{CringeShows}$$

We are going up a level of abstraction[12] and treating categories like objects themselves. F and G are called functors, "morphisms between categories"[13]. We can also compose these functors $H = G \circ F$. My friends do not know exactly what H is, but they can approximate it as H_{approx}. We can go up a level of abstraction again and now ask, what is the relationship between these functors? The morphisms between functors are called natural transformations[14]:



We can start reasoning about how good the approximations are, whether they bring understanding. Maybe watching *Emily in Paris* wouldn't be so bad after all. We can also think about Mood more. In reality, there are many more arrows[15] coming in and out of Mood. We can observe them, perform interventions [49], see what happens after, etc. This is where the Yoneda Perspective comes in: "mathematical objects are completely determined by their relationships to other objects." [25]

We can also take even a bigger step back and reassess our abstractions. Maybe we should use more advanced category-theoretic objects to represent the relationships we see. Maybe we should also question the certainty of our observations and take that into account too.

$$\text{Reality} \rightarrow \text{Abstractions} \tag{2.3}$$

It's also easy to get lost in abstractions. It's important to maintain a purpose, and know when to look more closely into the real world[16]. And eventually, take action[17].

[12] Abstracting is a careful and controlled forgetting of the details to unify situations efficiently. [23]

[13] Traditional (strict) functors need to satisfy a condition called functoriality. We also have generalization of functors (like lax functors) that relax that condition

[14] Natural transformations need to satisfy a condition called naturality. In our case, this amounts to preserving the relative ordering

[15] I can arguably also communicate my feelings. To some faithful [47] extent, anyway [48].

[16] "The great human error is to reason in place of finding out", Simone Weil

[17] Yes, this means stop texting him.

The next sections will be more technical. I will build up from specific papers. I will try to always give the high-level picture, but if you are interested in the details, I strongly encourage you to read the specific papers I build upon.

3. LLMs as categories?

There are many[18] possible starting points for us to start wandering from category theory into interpretability. I am particularly inspired by *Bradley et al* [58], [59], and I really recommend you to read her work (or watch her talks on YouTube). We will start by rephrasing some of her work in this section to get us started.

When we talk about LLMs in this piece, we refer to **auto-regressive** LLMs, which generate probabilities of a sequence of tokens [60]:

$$p(t_1, ..., t_n) = p(t_1) \prod_{i=1}^{n-1} p(t_{i+1} \mid t_1, ..., t_i) \tag{3.1}$$

where each token belongs to a core finite alphabet $t \in A_{t_{\text{core}}}$.

We also consider two special tokens to indicate the start-of-sequence and end-of-sequence: t_{sos} and t_{eos}. The extended alphabet is $A_{t_{\text{ext}}} = A_{t_{\text{core}}} \cup \{t_{\text{sos}}, t_{\text{eos}}\}$. We also consider that LLMs have a fixed context window $N_{\text{cutoff}} \in \mathbb{N}$

Sequences of tokens are strings. All possible finite strings are formed from the free monoid over the token alphabet, $A_{\text{s}} = A^*_{t_{\text{core}}}$. We want terminating texts, so we define the set of sequences of valid strings Seq_s as the strings that start with t_{sos}, and

- end with t_{eos} and have less than $N_{\text{cutoff}} - 1$ core tokens (finished texts)
- do not end with t_{eos} and have less or equal $N_{\text{cutoff}} - 1$ core tokens (unfinished texts)

$$\text{Seq}_s = \{t_{\text{sos}} s : s \in A_{\text{s}} \wedge |s| \leq N_{\text{cutoff}} - 1\} \cup \{t_{\text{sos}} s t_{\text{eos}} : s \in A_{\text{s}} \wedge |s| < N_{\text{cutoff}} - 1\} \tag{3.2}$$

$|s|$ counts core tokens; the full sequence length always includes t_{sos}, but only t_{eos} if finished.

Let's think about **prompts** and **continuations** as:

$$x = s_{\text{prompt}} = t_{\text{sos}} t_1 ... t_p \tag{3.3}$$

[18] Really many [22], [28], [50], [51], [52], [53], [54], [55], [56], [57]

$$y = s_{\text{cont}} = xt_{p+1}...t_{p+k} \quad \text{where} \quad t_{p+k} = t_{\text{eos}} \quad \text{or} \quad |s| = |t_1...t_{p+k}| \leq N_{\text{cutoff}} - 1 \tag{3.4}$$

We can also define a prefix relation like

$$x \leq y \Leftrightarrow \exists s \mid y = xs \tag{3.5}$$

For every prompt, we have the set of all terminating texts $T(x)$. We can chain the token-level probabilities to give full-text probabilities:

$$p(y|x) = p(t_{p+1}|x) \prod_{i=1}^{k-1} p(t_{p+i+1} \mid xt_{p+1}...t_{p+i}) = \prod_{i=1}^{k} p(t_{p+i} \mid y_{<p+i}) \tag{3.6}$$

How do we form a category to represent this? Let's set aside the probabilities for a little bit and focus on what they act on: words. How do we form sentences, texts,... from words? The base language category will give us that initial support to structure words.

The **Base Language Category** L_{base} is a prefix poset ($x \leq y$ iff x is a prefix of y) which has

- Objects: elements of Seq_s as in Equation (3.2)
- Morphisms: prefix relation as in Equation (3.5)

This category is a thin category like we had in Section 2 (What is a category?)! Let's visualize it. For simplicity, imagine our words match our tokens and we have $A_{\text{small}} = \{\text{stop}, \text{texting}\}$ and $N_{\text{cutoff}} = 4$. So our category would look like:

Figure 1: Example of L_{base} for a small alphabet. Terminating states are labeled by box

In Figure 1, all the objects with no outgoing arrows constitute the set of terminating states $T(x)$.

Does this represent a LLM? Not yet! The L_{base} category captures the **compositional** structure of language, all the syntactic ways a prompt can be extended. However, we are missing the **distributional** structure. For that, let's construct a category enriched over the unit interval. [58] will give you technical details of this enrichment and $[0, 1]$-categories, so we will focus here on the result.

Every LLM defines a L_{syn} Category

The **Language Syntax Category** L_{syn} is a $[0, 1]$-category which has

- Objects: Same objects as L_{base}
- Hom-objects: For every $x, y \in \text{Obj}(L_{\text{syn}})$, we have $L_{\text{syn}}(x, y) := \pi(y|x)$, the probability that y extends x defined as:

$$\pi(y \mid x) := \begin{cases} 1 & \text{when } x = y \\ 0 & \text{when } x \not\to y \\ \prod_{i=1}^{k} p(t_{p+i} \mid y_{<p+i}) & \text{when } x \to y \end{cases} \tag{3.7}$$

Note that $\pi(-|x)$ becomes the probability mass function **only** when restricted to $T(x)$. The composition becomes

$$\pi(y|x)\pi(z|y) \leq \pi(z|x) \tag{3.8}$$

Equality holds when y is exactly the chosen intermediate prefix on the unique path $x \leq z$; the $\leq$ is the enriched triangle inequality.

Let's consider our previous example with a small alphabet, but now over the enriched version of the category. Let's say our prompt is $x = t_{\text{sos}}$ stop. The possible terminating continuations y will have the following probabilities:

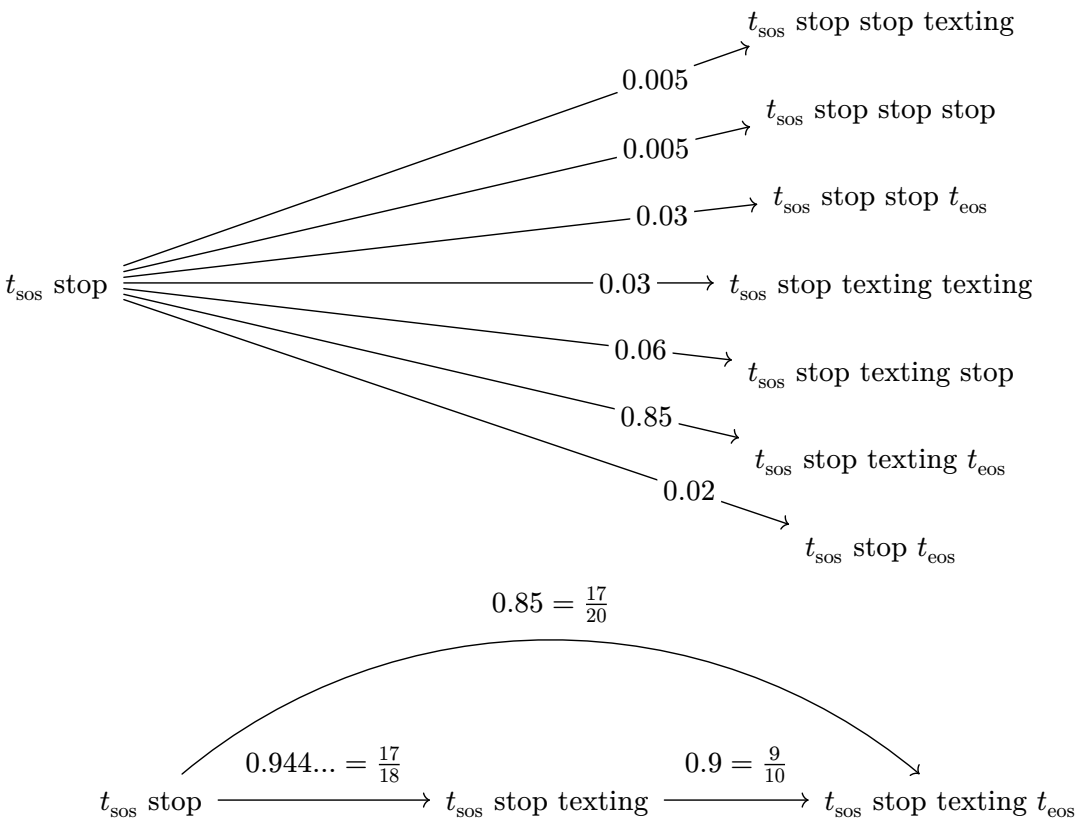


Figure 2: **Top:** Showing total probabilities in L_{syn} for given prompt.
Bottom: Showing compositionality of $\pi(y \mid x)$

And voilà! That's how we can construct categories that encode the behavior of LLMs.

Every LLM will correspond to a single L_{syn} category. All possible L_{syn}categories will live inside an ambient category we will name $\mathfrak{L}_{\text{syn}}$. Real relations between concrete LLMs now correspond to transformations in $\mathfrak{L}_{\text{syn}}$. For instance, we have transformations between L_{syn} categories that represent processes that update model weights [61], including Pretraining(PT), Supervised Fine-Tuning(SFT), and Reinforcement Learning from Human Feedback (RLHF):

$$L_{\text{syn}}^{\text{untrained}} \xrightarrow{\text{PT}} L_{\text{syn}}^{\text{pretrained}} \xrightarrow{\text{SFT}} L_{\text{syn}}^{\text{fine-tuned}} \xrightarrow{\text{RLHF}} L_{\text{syn}}^{\text{aligned}}$$

4. Looking for meaning through syntax

As we will see later, **to reason about interpretations, it helps to have ways to make comparisons, sometimes between two LLMs.** Since every LLM defines a specific L_{syn} category, a sensible next step for us is to explore how each L_{syn} category relates to each other. To pry on possible structure, let's go back to my situation(ship)s for a second.

As it turns out, there are multiple people[19] I ask my LLM about. Interestingly enough, my chatbot sometimes says the same thing about certain people. Some men whose names start with the letter "M" I should definitely not text. Let's ignore the long and emotional writing I often input as part of my context window, and let's imagine my LLM actually has all of that as internal knowledge. Let's look at what the LLM tells me when I ask "Should I text Mahdi?" vs "Should I text Mark?" ?

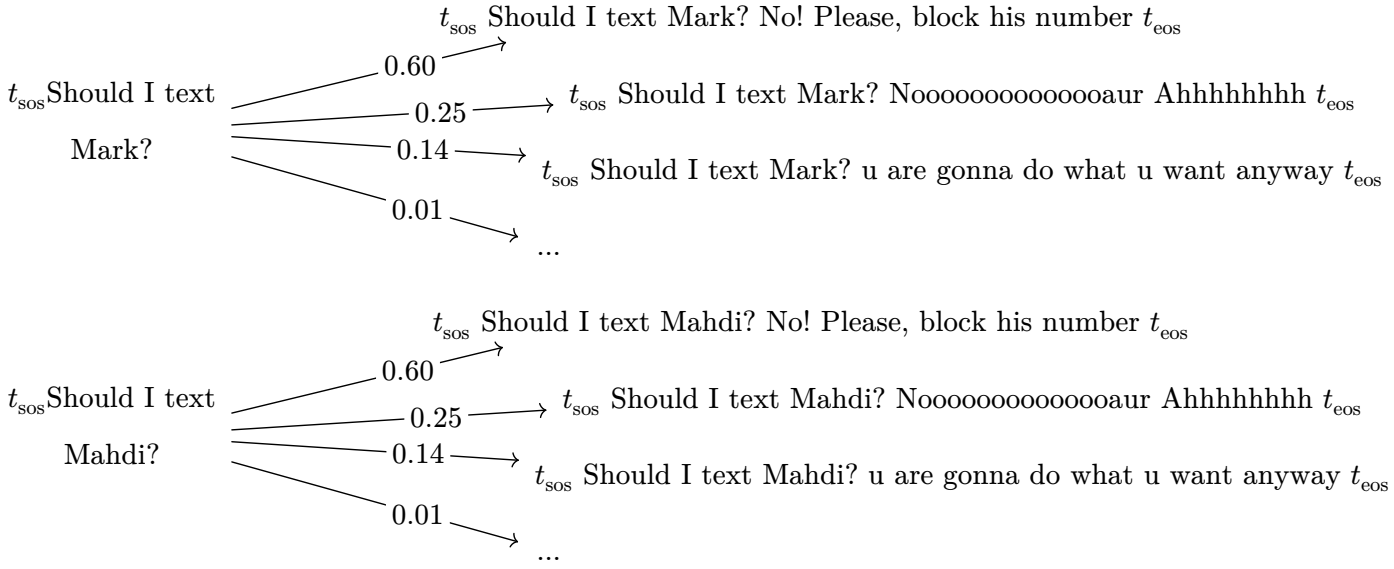


Figure 3: Two prompts having the same distribution of continuations

What is happening? We have two prompts $x_p := t_{\text{sos}}$Should I text Mark? and $x_q := t_{\text{sos}}$Should I text Mahdi? that produce the same distribution of continuation suffixes. As I mentioned before, when we think categorically, we often encode everything knowable about an object in terms of how other objects relate to it. We could say both prompts are *observationally equivalent*. We can use the $[0, 1]$-enriched Yoneda Embedding from *Bradley et al* [59] to investigate what that equivalence looks like more explicitly. We have a functor from L_{syn} to another category, L_{sem}.

$$L_{\text{syn}} \xrightarrow{\text{Yoneda}} L_{\text{sem}}$$

The **Language Semantic Category** L_{sem} has

- Objects: For every object x in L_{syn}, we have an enriched functor $h^x := L_{\text{syn}}(x, -)$ as an object in L_{sem} such that

[19]The curse of being both polyamorous and a love addict, I fear

$$h^x(y) := \begin{cases} \pi(y \mid x) & \text{if } x \to y \text{ in } L_{\text{syn}} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

h^x is the enriched Yoneda Embedding, also an enriched copresheaf.

- Hom-objects: For every $h^x, h^y \in \text{Obj}(L_{\text{sem}})$, we have $L_{\text{sem}}(h^x, h^y) := \inf_{z \in L_{\text{syn}}}[h^x(z), h^y(z)]$, where [a,b] is the internal hom defined as:

$$[a, b] := \begin{cases} \frac{b}{a} & \text{if } b < a \\ 1 & \text{otherwise} \end{cases} \tag{4.2}$$

The direction of the arrows is reversed between L_{syn} and L_{sem}. If we have $x \to y$ in L_{syn}, we will have $h^x \leftarrow h^y$ in L_{sem}. **The more specific a text becomes, the fewer contexts it can meaningfully continue into. Text grows by accumulation while meaning emerges through constraint**. We could start to philosophize a bit and say that meaning is *the instructions of concept assembly* [62] through this contextual constraint.

I recommend you check out *Bradley et al* [58], [59] for the technical details[20]. We have quickly introduced a lot, so let's see what this looks like in my example:

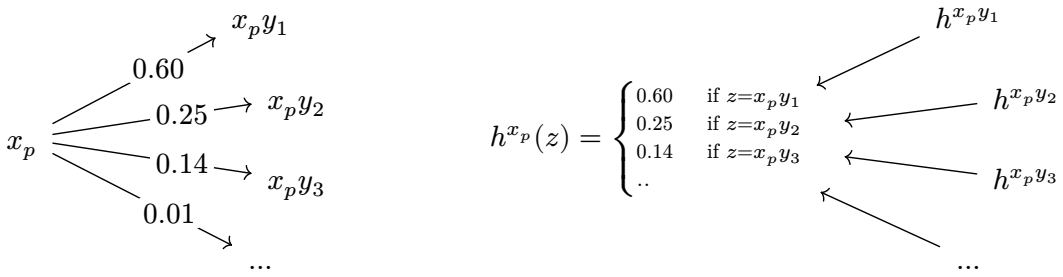


Figure 4: Applying the Yoneda Embedding to "Should I text Mark?" prompt from Figure 3

What we see is that the semantic content of x_p ("Should I text Mark?") is **encoded by the distribution of its continuations**. This coincides with *TY Liu et al* [63]: if two prompts have the same "syntactic meaning representation", they will be "indistinguishable based on their continuations for the model". We can define the "Continuation Similarity" as[21]:

[20] To define the base language category, [59] considers substrings while [58] considers prefixes (which directly connect to LLMs; what we use in this piece). Just keep that in mind as you read through those papers.

[21] There are some details to iron out to make this precise. First, whether we only require terminal texts to have the same distribution or all intermediate continuations (and whether that is different). Secondly, what happens when one of the prompts puts us considerably closer to N_{cutoff}.

$$x \sim y \iff \forall z \in A_{\mathrm{s}} : h^x(xz) = h^y(yz) \tag{4.3}$$

In real life, we will not find exact equivalences. Let's provide an approximate version. The "Approximate Continuation Similarity" would be

$$x \underset{\varepsilon}{\sim} y \iff \forall z \in A_{\mathrm{s}} : d(h^x(xz), h^y(yz)) < \varepsilon \tag{4.4}$$

where $d(a, b)$ is a probability metric like Jensen-Shannon Distance or Total Variation Distance. Then, we can define "Approximate Continuation Equivalence" if there is a chain of similarities like:

$$x \underset{\varepsilon}{\equiv} y \iff \exists n \geq 0, \exists v_0, ...v_n : x = v_0 \underset{\varepsilon}{\sim} v_1 \underset{\varepsilon}{\sim} ... \underset{\varepsilon}{\sim} v_n = y \tag{4.5}$$

which in turn defines the **equivalence class**:

$$[x] = \left\{ y \in \mathrm{Obj}\big(L_{\mathrm{syn}}\big) : x \underset{\varepsilon}{\equiv} y \right\} \tag{4.6}$$

We could have made different choices to construct this equivalence class. For instance, instead of a probability metric, we could have leveraged our enriched setting to define $S(x, y) = \inf_{z \in A_s}[h^x(z), h^y(z)] \in [0, 1]$ and $d_s(x, y) = -\log(S(x, y))$. We could also have introduced a chain-metric $d_{\mathrm{chain}} = \inf_{(x=v_0,...,v_n=y)} \sum_i d_s\big(v_i, v_{i+1}\big)$. Each of these choices has pros and cons. Right now, I am more interested in what we could **do with such an equivalence class**: Let's define a new category.

Interpreting topology from meaning

The **ε-Continuation Quotient Category** is $Q_\varepsilon := L_{\mathrm{syn}} / \underset{\varepsilon}{\equiv}$. It is a [0,1]-category with:

- Objects: Equivalence classes [x]
- Hom-objects: For direct edges, we choose the supremum over representatives:

$$Q_\varepsilon([x], [y]) = \sup_{(x' \in [x], y' \in [y])} L_{\mathrm{syn}}(x', y') \tag{4.7}$$

- For compositionality to work, we need to be careful about choosing intermediaries that align. So for the one-intermediary case (via any $y' \in [y]$), this would be:

$$Q_\varepsilon([x], [z]) = \sup_{(x' \in [x], y' \in [y], z' \in [z])} L_{\mathrm{syn}}(y', z') L_{\mathrm{syn}}(x', y') \tag{4.8}$$

When we think about longer chains, the formula looks a bit overwhelming, but the idea is that we keep consistent intermediaries as we multiply over all possible paths:

$$Q_\varepsilon([x],[z]) = \sup_{k\geq 0} \sup_{(x'\in[x],z'\in[z])} \sup_{(y'_1,\ldots,y'_{k-1})} \prod_{i=0}^{k-1} L_{\text{syn}}(y'_i, y'_{i+1}) \tag{4.9}$$

with $y'_0 = x'$ and $y'_k = z'$ such that $Q_\varepsilon([x],[x]) = 1$ and $\forall [x],[y],[z] : Q_\varepsilon([y],[z])Q_\varepsilon([x],[y]) \leq Q_\varepsilon([x],[z])$

We can think that there is an ε-Continuation Collapse Functor $\mathbb{Q}_\varepsilon : L_{\text{syn}} \to Q_\varepsilon$. $\mathbb{Q}_\varepsilon$ collapses prompts whose continuation distributions are ε-close, then uses the largest available transition probability between any representatives as the class-to-class hom. This lets us study the 'shape' of meaning neighborhoods and compare them across models.

In practice, we can look at a specific prompt, smartly sample continuations to get a wide picture of the distributions, and apply $\mathbb{Q}_\varepsilon$. Remember that terminal states form a total probability, and that every $L_{\text{syn}}(x, y)$ is the upper bound on any terminal continuation beneath it.

Why do we want to do any of this? LLMs that have the same token alphabet and context window, even if they have different architectures[22], will share the same L_{base}. If we wanted to compare two L_{syn} categories, we are stuck with two of the same diagrams but with different $\pi(y|x)$ hom-objects, each produced by a different causal structure. Constructions like $\mathbb{Q}_\varepsilon$ allow us to translate the structure in one domain (information-theoretic enrichment in L_{syn}) into another domain (topology, geometry of Q_ε), where we can apply different types of mathematics to investigate what's below the surface.

I am frustrated that both ChatGPT and Claude tell me not to text him. What if I could fold what I read from them into a shape:

[22] As long as they are auto-regressive.

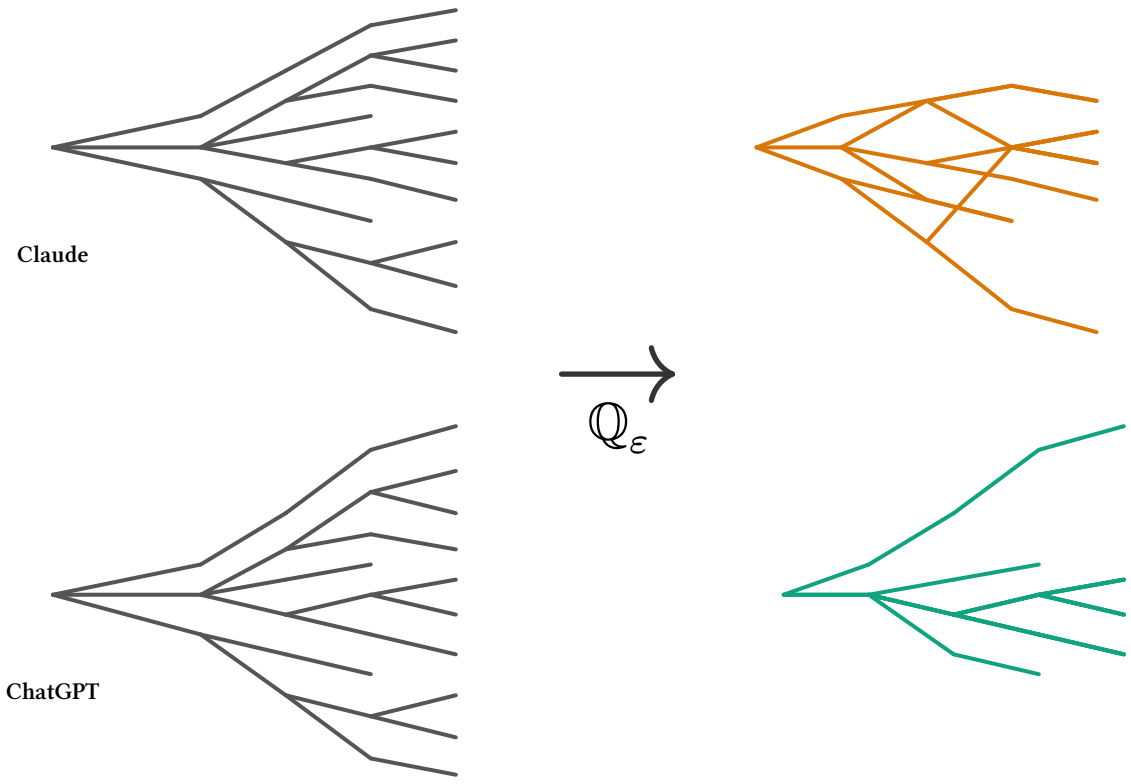


Figure 5: Imagining different induced topology for the same prompt

I start to wonder, what does it mean for the shapes of the quotient to be similar or different?[23]. Even if they were identical, there is a possibility that different inner mechanisms produce the same observable behavior. To gather evidence to expand our *observational faithfulness*, we can make *edits* and see if the new Q_ε shapes still match up. We can perform perturbations to test for robustness (real structure persists and remains stable) and interventions to test causality (the organization of causal effects is preserved). We will dig into that in the next sections[24]. We are about to get a bit more serious, kids.

Before moving on, **let's highlight what we did**: We inspected objects (of an LLM category) by looking at their relationships (L_{syn} to L_{sem}). We used our intuition (two prompts have the same meaning if their continuations are the same) and chose a well-motivated structure ($\underset{\varepsilon}{\equiv}$) to look at our objects through another viewpoint ($\mathbb{Q}_\varepsilon$). From that angle, we asked ourselves what things we could reason about from our original objects (observational identification by topology?). **Why did we do this?** To find ways to compare objects insightfully in our quest for interpretability. **What's the big picture?** More than championing this specific technical construction, I wanted to show how we can think categorically about LLMs and build bridges across perspectives.

[23] In real life, we also need to consider the metric d and threshold ε sensitivity, iterate through design choices for our quotient category, and ultimately gauge how much shape of the shape difference can be attributed to noise and pipeline choice

[24] In Section 6 (Observational Faithfulness), we'll briefly return to these constructions, if you were curious where they fit later

5. Framing Interpretability

To interpret is to form an understandable explanation. An explanation is understandable when it conveys context-specific meaning that fits our mental model, so we can anticipate the system's behavior and choose appropriate actions. [9].

As such, understandability is situational and oriented[25]. Any explanation $\mathbb{E}$ has multiple associated understandabilities Und, depending on who the subject is in what context:

$$\begin{aligned} ... \leq \mathrm{Und}_x(\mathbb{E}_i) \leq \mathrm{Und}_x(\mathbb{E}_j) \leq ... \\ \text{change of subject / context} \Big\Downarrow \\ .. \leq \mathrm{Und}_y(\mathbb{E}_j) \leq \mathrm{Und}_y(\mathbb{E}_i) \leq ... \end{aligned}$$

The process of interpretation [11] is to map a less understandable explanation to a more understandable one:

$$\mathrm{Interpret}_x = \mathbb{I}_x : \mathbb{E}_i \to \mathbb{E}_j \Leftrightarrow \mathrm{Und}_x(\mathbb{E}_i) \leq \mathrm{Und}_x(\mathbb{E}_j) \tag{5.1}$$

An **explanation** is a representation of a mechanism applied to observations[26] Formally[27], working in 2-category **Cat**, we define:

- An input category $\mathcal{X}$
- Two output categories: an observation category $\mathcal{Y}$ and a prediction category $\mathcal{Z}$
- An indexing category $\mathcal{I}$
- The mechanism $\mathbb{M} : \mathcal{X} \to \mathcal{Z}$
- The evaluation $\mathbb{X} : \mathcal{I} \to \mathcal{X}$
- The observation $\mathbb{Y} : \mathcal{I} \to \mathcal{Y}$
- An explanation as the triplet $\mathbb{E} = (\mathbb{M}, \mathbb{X}, \mathbb{Y})$, with:
  - The prediction: $\mathbb{Z} = \mathbb{M} \circ \mathbb{X}$
  - The analysis natural functor: $\Delta : \mathbb{Y} \to \mathbb{Z}$
  - The synthesis natural functor: $\Sigma : \mathbb{Z} \to \mathbb{Y}$
  - The precision witness unit: $\omega_{\text{precision}} : \mathrm{id}_{\mathbb{Y}} \to \Sigma \circ \Delta$
  - The resolution witness counit: $\omega_{\text{resolution}} : \Delta \circ \Sigma \to \mathrm{id}_{\mathbb{Z}}$

  Note: For convenience, we will abuse notation and write $\mathbb{E} = (\mathbb{M}, \mathcal{X}, \mathcal{Y})$ to make the underlying input/output categories explicit.

[25] When we also consider to whom what understandings are more reachable, understandability is also political [3]

[26] In literature [11], you'll rather see Mechanism called Explanan, Observation as Explanandum, and Evaluation as Process of Explanation.

[27] We could make this definition more general by using profunctors [46]

The analysis/synthesis functor and precision/resolution witnesses give us language to reason about a given explanation's inner uncertainty and incompleteness [6]. **We often have situations where we can match each prediction to its own observation**. That is the *Infinite Fidelity Assumption*: no noise and perfect resolution within an explanation, which leads to an isomorphism between predictions and observations:

$$\mathrm{id}_{\mathbb{Y}} = \Sigma \circ \Delta \quad \wedge \quad \mathrm{id}_{\mathbb{Z}} = \Delta \circ \Sigma \quad \Rightarrow \quad \mathbb{Z} \cong \mathbb{Y} \tag{5.2}$$

We can also consider a stricter *Ideal Prediction Assumption* as:

$$\mathbb{Z} \equiv \mathbb{Y} \tag{5.3}$$

Scope is the extent of validity of an explanation. We can think of $\mathcal{X}$ defining the whole possible domain and $\mathcal{I}$ selecting a concrete subdomain of interest. The concrete subdomain is what ultimately determines our explanatory scope.

To understand this, imagine we have $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. We could have the different choices for the indexing category $\mathcal{I}$, defining two different explanations :

- $\mathbb{E}_G$ will have $\mathcal{I} = \mathbb{R}$. In this case, the scope of our explanation covers all possible inputs.
- $\mathbb{E}_L$ will have $\mathcal{I} = \{1\}$. In this case, the scope of our explanation covers a single number of a much bigger possible domain $\mathbb{R}$.

But what if instead we re-defined $\mathcal{X} = \{x_0\}$ in $\mathbb{E}_L$? Nothing in our explanation itself would change. For an explanation $\mathbb{E}_L$, considering a possible domain bigger than its concrete one is meaningless unless we compare it to another explanation $\mathbb{E}_G$, which has a concrete domain that covers all of $\mathbb{E}_L$'s possible domain.[28]

An indexing category defines a "shape". Functors of the form $\mathbb{F} : \mathcal{I} \to \mathcal{C}$ are "selecting" a diagram in $\mathcal{C}$ with $\mathcal{I}$ shape. We have shapes in all levels of abstraction,

An explanation is a triplet of functors, so its shape will start with arrows from three categories C_a, C_b, C_c, to another three categories C_1, C_2, C_3. But the definition of explanation requires the shape to respect further structure. For instance, we know that two of the categories must be the same as a particular indexing category: $C_b = C_c = \mathcal{I}$.

The **shape category** $\mathcal{S}$ will be the category that fully encapsulates the shape of an explanation. This shape can be used to "pick" a particular explanation from all possible explanations, just like the index

[28] We will continue this discussion on Section 6 (Locality)

category allows us to pick a particular value from all possible values in the input category. This "selection" of an explanation is a transformation $[\mathcal{S}, \mathsf{Cat}] : \mathcal{S} \to \mathsf{Cat}$ where Cat is the category of all categories. Each explanation will have a "selection". The **diagram selection operator** is then:

$$\mathrm{diag}(-) : \mathbb{E} \to [\mathcal{S}, \mathsf{Cat}] \tag{5.4}$$

The **diagram** of an explanation $\mathbb{E}_k$ is:

$$\mathbb{D}_k \coloneqq \mathrm{diag}(\mathbb{E}_k) \tag{5.5}$$

An **interpretation** $\mathbb{I} : \mathbb{E}_{\text{target}} \to \mathbb{E}_{\text{proxy}}$ will consist of natural transformations for every explanation component:

- A mechanism transformation $\mathbb{I}_{\mathbb{M}} : \mathbb{M}_{\text{target}} \to \mathbb{M}_{\text{proxy}}$
- An evaluation transformation $\mathbb{I}_{\mathbb{X}} : \mathbb{X}_{\text{target}} \to \mathbb{X}_{\text{proxy}}$
- A prediction transformation $\mathbb{I}_{\mathbb{Z}} : \mathbb{Z}_{\text{target}} \to \mathbb{Z}_{\text{proxy}}$
- An observation transformation $\mathbb{I}_{\mathbb{Y}} : \mathbb{Y}_{\text{target}} \to \mathbb{Y}_{\text{proxy}}$
- These transformations are lax[29] [36], so we have the **approximation components**:

$$\varphi_{\mathbb{M}} : \mathbb{I}_{\mathbb{Z}} \circ \mathbb{M}_{\text{target}} \to \mathbb{M}_{\text{proxy}} \circ \mathbb{I}_{\mathbb{X}} \tag{5.6}$$

$$\varphi_{\mathbb{X}} : \mathbb{I}_{\mathbb{X}} \circ \mathbb{X}_{\text{target}} \to \mathbb{X}_{\text{proxy}} \tag{5.7}$$

$$\varphi_{\mathbb{Y}} : \mathbb{I}_{\mathbb{Y}} \circ \mathbb{Y}_{\text{target}} \to \mathbb{Y}_{\text{proxy}} \tag{5.8}$$

What about $\varphi_{\mathbb{Z}}$? To help us with the technicalities of connecting predictions and observations, we define a *conjugation*. Given $F, G : C \to D$, $R : V \to C$, $L : D \to W$ functors, and $\alpha : F \to G$ natural transformation, we have the natural transformation:

$$(L \circ \_ \circ R)[\alpha] : L \circ F \circ R \to L \circ G \circ R \tag{5.9}$$

Considering Equation (5.2), we can canonically define :

$$\mathbb{I}_{\mathbb{Z}} \coloneqq \big(\Delta_{\text{proxy}} \circ \_ \circ \Sigma_{\text{target}}\big)[\mathbb{I}_{\mathbb{Y}}] : \mathbb{Z}_{\text{target}} \to \mathbb{Z}_{\text{proxy}} \tag{5.10}$$

$$\varphi_{\mathbb{Z}} \coloneqq \big(\Delta_{\text{proxy}} \circ \_ \circ \Sigma_{\text{target}}\big)[\varphi_{\mathbb{Y}}] : \mathbb{I}_{\mathbb{Z}} \circ \mathbb{Z}_{\text{target}} \to \mathbb{Z}_{\text{proxy}} \tag{5.11}$$

[29] lax = only approximately commuting

Interpretations connect two explanations. Each explanation has a different diagram but they have the same shape:

$$\begin{aligned} \mathbb{D}_{\text{target}} &:= \operatorname{diag}\big(\mathbb{E}_{\text{target}}\big) : \mathcal{S} \to \mathsf{Cat} \\ \mathbb{D}_{\text{proxy}} &:= \operatorname{diag}\big(\mathbb{E}_{\text{proxy}}\big) : \mathcal{S} \to \mathsf{Cat} \end{aligned} \tag{5.12}$$

Interpreting is matching diagrams by shape and then asking what is different besides the shape. The approximation components $\varphi_{\mathbb{M}}, \varphi_{\mathbb{X}}, \varphi_{\mathbb{Z}}, \varphi_{\mathbb{Y}}$ form the **approximation transformation**, which captures that idea:

$$\varphi_{\bullet} : \mathbb{I} \circ \mathbb{D}_{\text{target}} \to \mathbb{D}_{\text{proxy}} \tag{5.13}$$

All the diagrams still need to be coherent. There are three sensible ways to go from target observations $\mathbb{Y}_{\text{target}}$ to proxy predictions $\mathbb{Z}_{\text{proxy}}$[30] : (A) approximate the observations first and then run the analysis on the proxy side; (B) run the analysis on the target, translate to proxy, and then approximate the resulting predictions; and (C) approximate the inputs and mechanism, and run that proxy pipeline after translating the analysis on target. Coherence says these routes agree (up to the explicit approximation maps): This coherence is formally written[31] :

$$\Delta_{\text{proxy}} \circ \varphi_{\mathbb{Y}} = \varphi_{\mathbb{Z}} \circ \mathbb{I}_{\mathbb{Z}}\big(\Delta_{\text{target}}\big) = \mathbb{M}_{\text{proxy}}(\varphi_{\mathbb{X}}) \cdot \varphi_{\mathbb{M}}\big(\mathbb{X}_{\text{target}}\big) \cdot \mathbb{I}_{\mathbb{Z}}\big(\Delta_{\text{target}}\big) \tag{5.14}$$

To simplify this, we can consider the *Identical Evaluation Assumption*:

$$\mathbb{X}_{\text{target}} \equiv \mathbb{X}_{\text{proxy}} \tag{5.15}$$

which implies *Ideal Evaluation*:

$$\mathbb{I}_{\mathbb{X}} = \varphi_{\mathbb{X}} = \mathrm{id}_{\mathbb{X}} \tag{5.16}$$

which, together with Equation (5.3), reduces Equation (5.14) to:

$$\varphi_{\mathbb{Y}} = \varphi_{\mathbb{M}}(\mathbb{X}) \tag{5.17}$$

[30]To internalize the intuition for this coherence, loosely think that "to translate object" = "to interpret proxy object from target object" and "to run analysis" = "to form predictions from observations"

[31]We use $\cdot$ for vertical composition of natural transformations

Let's name Equation (5.17) as *Basic Approximation* because we will use it a lot on concrete applications that leverage the simplifying assumptions to make calculations easy.

In the very ideal situation, where our approximation transform is id, we would have *perfect observational transport*:

$$\mathbb{I} \circ \mathbb{D}_{\text{target}} = \mathbb{D}_{\text{proxy}} \tag{5.18}$$

With *perfect observational transport*, any $\|\varphi_{\bullet}\| = 0$

Phew, that is a lot of notation. I do not want to dwell on formalism in this piece. This is what is important so far: We have **language** to start reasoning about faithfulness. **What's the big picture?** The approximation transformation $\varphi_{\bullet}$ tells us that the gaps between explanations are precisely gaps between mechanisms $\varphi_{\mathbb{M}}$, evaluations $\varphi_{\mathbb{X}}$, or observations $\varphi_{\mathbb{Y}}$.

6. Decomposing Faithfulness

Faithfulness is the **degree** to which we can look at two explanations and say they are "the same". The definition of faithfulness has been fuzzy because we have not totally agreed on what explanatory "sameness" we care about and which we can live without.

A useful starting point is to remember that we came into this discussion because we had a goal, there is something specific we want to understand: *why does Claude tell me not to text Mark ?!?!* Well, maybe now I am interested in the general case, so concretely, the specific **target of our explanations is the production LLM** for all its possible inputs and outputs.

How do we connect something abstract, like an explanation, to something concrete, like an LLM? This is the trick: **the LLM model itself can be thought of as an explanation, one with perfect faithfulness**:

$$\mathbb{E}_{\text{prod}} = \left(\text{Arch}^{\text{prod}}_{\theta^{\text{prod}}}, \text{Seq}_s, \mathfrak{L}_{\text{syn}}\right) \tag{6.1}$$

where $\text{Arch}^{\text{prod}}_{\theta^{\text{prod}}}$ is the complete computation process defined by a specific production LLM architecture $\text{Arch}^{\text{prod}}$ with θ^{prod} parameters, that deterministically computes next-token distributions to every input prompt. Seq_s is the set of sequences of valid strings defined in Equation (3.2), and $\mathfrak{L}_{\text{syn}}$ is the ambient category of all L_{syn} defined in Section 3 (Every LLM defines a L_{syn} Category).

What about inherent interpretability? Inherent interpretability would correspond to the case our $\mathbb{E}_{\text{prod}}$ itself has high understandability for everyone[32]. We could formalize this with respect to Ω, an understandability reference[33]:

$$\mathbb{E}_{\text{prod}} \text{ is inherently interpretable} \Leftrightarrow \forall i, \Omega \ll \text{Und}_i(\mathbb{E}_{\text{prod}}) \tag{6.2}$$

Everyone will *tell* you they understand it. I ask myself, do I really? *How does each of us connect something abstract, like an explanation, to something deep within us, something personal, our experience?* Another trick: **our own mental model itself can be thought of as an explanation, one with perfect understandability.**

Not all mental models are the same. When I started deadlifting, I learned everything that I needed to do to have good form. For many months, I silently recited the checklist and asked my coach for feedback. I only had a **verbatim** explanation (analytical, precise, formal) for how to deadlift.

Things started to click inside me. My body started telling new things to me, or maybe I was finally ready to listen. Without trying, I would use "correct form" when picking up packages. And when a boy hurt my heart, my body craved to deadlift next. I got the **gist** down, a fuzzy internal explanation for deadlifting I directly operate on.

The process of understanding [9] itself is a map: $\mathbb{E}_{\text{verbatim}} \to \mathbb{E}_{\text{gist}}$

The risks exist and persist. Thinking I understand something is different from actually understanding it. A $\mathbb{E}_{\text{verbatim}}$ can feel very plausible [66], so I become overconfident that my $\mathbb{E}_{\text{gist}}$ is faithful[34]. Some other times, I really did have a faithful explanation of something. But the world changed, and I did too. So I had to form new interpretations to guide me where I wanted to go next. **In one way or another, we end up leveraging a post-hoc explanation.** The question of faithfulness does not go away.

In the rest of this section, we will technically explore a few ways an explanation can be faithful.

Locality

Sometimes, we try to form interpretations on a subset of all possible inputs. We say $\mathbb{X}_{\text{proxy}} : \mathcal{I} \to \mathcal{X}_{\text{proxy}}$ is restricted if for $\mathbb{X}_{\text{target}} : \mathcal{I} \to \mathcal{X}_{\text{target}}$, we have $\mathbb{X}_{\text{proxy}} \subseteq \mathbb{X}_{\text{target}}$. Restriction form local explanations is

[32] Again, is understanding distributed evenly [64]? Are we imposing a code of legibility [65]?

[33] How much my husband expects me to know of basketball

[34] This is an actual risk with me, with this piece. I align with [67] that math is a deeply human, bodily practice, which uses formalisms to refine and expand intuition. If not ultimately correct, I hope my failures [68] do guide you to better intuitions.

formalized by inclusion[35] map like $\mathbb{X}_{\text{local}} \hookrightarrow \mathbb{X}_{\text{target}}$. We can think of having a functor that restricts our index category:

$$i : \mathcal{i} \to \mathcal{I} \tag{6.3}$$

If we are only looking at one prompt, we would have $i : 1 \to \mathcal{I}$ Our local explanation would then be:

$$\mathbb{E}_{\text{local}} = (\mathbb{M}_i, \mathbb{X} \circ i, \mathbb{Y} \circ i) \tag{6.4}$$

In the best case, our local explanations are incomplete, but behaviorally faithful in the restricted domain. Most likely, however, the approximations will not be exact. If we assume Equation (5.3) and Equation (5.16), we see our coherence is also restricted to the inclusion:

$$\varphi_{\mathbb{Y}} \circ i = \varphi_{\mathbb{M}}(\mathbb{X} \circ i) \tag{6.5}$$

For a second, let me do a sheaf-theoretic [69] speculation. We could glue local explanations $\mathbb{E}_{\text{local}}$ to form a $\mathbb{E}_{\text{global}}$ that covers the whole $\mathbb{X}_{\text{target}}$. This construction would need to satisfy the gluing and locality conditions. These conditions could be applied to different levels of the explanation (observational, causal, ...), forcing each $\mathbb{E}_{\text{local}}$ to comply to each level of structure.

What's the big picture? We need more principled and practical ways to compare local explanations to each other based on their overlap, and ways to compose local explanations into global ones.

Lossiness

Sometimes, we are less concerned about the full observations (text outputs from LLM) and more interested in specific attributes of the observations (like toxicity, honesty, fairness, etc). This is the case with Representation Engineering (RE): Manipulation of the representations of a model in order to control its behavior with regard to an attribute [70].

We have lossy observations when our proxy explanation has an observation transformation $\mathbb{I}_{\mathbb{Y}}$ that has non-injective components. Then, even if our proxy had exact observations ($\varphi_{\mathbb{Y}} = \text{id}_{\mathbb{Y}}$), our interpretation could never recover the whole target explanation by themselves. If we take Equation (5.2) as before, you can see that $\mathbb{I}_{\mathbb{Y}}^{\text{lossy}}$ also implies $\mathbb{I}_{\mathbb{Z}}^{\text{lossy}}$. Then, looking at $\varphi_{\mathbb{M}}$, we see that $\mathbb{M}_{\text{proxy}}$ can be at best (behaviorally) faithful modulo $\mathbb{I}_{\mathbb{Z}}^{\text{lossy}}$. We can express this as:

[35] $\hookrightarrow$ means inclusion

$$\mathbb{I}_{\mathbb{Y}}^{\text{lossy}} : \mathbb{Y}_{\text{target}} \overset{\text{collapse}}{\rightarrow} \mathbb{Y}_{\text{proxy}}$$
$$\mathbb{E}_{\text{lossy}} = \left(\mathbb{M} \,/\ker\left(\mathbb{I}_{\mathbb{Z}}^{\text{lossy}}\right), \mathbb{X}, \mathbb{Y}/\ker\left(\mathbb{I}_{\mathbb{Y}}^{\text{lossy}}\right)\right) \tag{6.6}$$

As I mentioned before, understandability is goal-oriented. Representation Engineering has shown great promise as a framework for Control [13]. **What's the big picture?** We just need to be aware of our blind spots, and so we can address current challenges [70], like deterioration of capabilities.

Observational Faithfulness

When we consider the basic case of Equation (5.17), we can calculate the **observational residual** as:

$$r_{\mathbb{O}} = \| \varphi_{\mathbb{Y}} \| \tag{6.7}$$

where $\|\_\|$ is a real-valued chosen functional. We can establish that an **interpretation is ε-observationally faithful** if

$$r_{\mathbb{O}} < \varepsilon \tag{6.8}$$

Our $r_{\mathbb{O}}$ value will depend on our choice of observation category $\mathcal{Y}$ and the corresponding functional $\|\_\|$. If we follow tradition, for every evaluation i, we get pair of next-token distributions for the same prompt $(\varphi_{\mathbb{Y}})_i = \left(p_{\text{t}}(-|x), p_{\text{p}}(-|x)\right)_i$, and the functional is the mean of KL divergence, then $\varphi_{\mathbb{Y}} = E_{x \in \text{Seq}_s}\left[D_{\text{KL}}\left(p_{\text{t}}(-|x), p_{\text{p}}(-|x)\right)\right]$. But we can also choose other $\mathcal{Y}$ categories with different functionals.

In Equation (6.1), I chose $\mathfrak{L}_{\text{syn}}$ to be the observation category to highlight other possibilities for $r_{\mathbb{O}}$. In Section 4 (Interpreting topology from meaning), we saw how the $\mathbb{Q}_\varepsilon$ functor can provide us a different view into each L_{syn}. What if we would like to compare the categories through their underlying graphs G with the Spectral Distance [71]? We could then have $\varphi_{\mathbb{Y}} = D_{\text{spectral}}\left(G\left(\mathbb{Q}_\varepsilon L_{\text{syn}}^t\right), G\left(\mathbb{Q}_\varepsilon L_{\text{syn}}^p\right)\right)$.

Some choices for $\|\_\|$ will have the equivalent effect of doing a $\mathbb{I}_{\mathbb{Y}}^{\text{lossy}}$ transformation[36]. Does your choice of metric capture all structures of interest?

What's the big picture? If all you have is an observational account of your interpretation, be creative and explore what different residuals tell you about the structure. If doing a construction like $\mathbb{Q}_\varepsilon$ is not appealing, at least make sure you consider higher-order statistics and not just look at expected value as the only functional you use to understand your interpretation.

[36]Consider that instead of D_{KL}, you use $D_{\max} = \left|\max p_{\text{t}}(-|x) - \max p_{\text{p}}(-|x)\right|$. As long as the peak probabilities do not change, any change to the distributions will be invisible to you.

Perturbational Robustness

We are interested in interpretations that are stable against small perturbations. Our interpretations would not be very understandable if a little noise in the inputs created large deviations.

We can think of adding a bit of noise to the input and estimating how different our approximation transformation is afterwards. I am going to hand-wave a bit from Equation (6.7) and define the **perturbation residual** with the perturbation kernel N :

$$r_P = E_{\boldsymbol{\delta} \sim \mathrm{N}} \left[\| \varphi_{\mathbb{Y}(x+\boldsymbol{\delta})} - \varphi_{\mathbb{Y}(x)} \| \right] \tag{6.9}$$

We have an **ε-robust interpretation** if $r_P < \varepsilon$

What's the big picture? Erratic systems will be harder to interpret. Try finding regions in the input domain that are more stable, and restrict your work there.

Interventional Faithfulness

With observational faithfulness, we can predict the target's behavior based on my proxy. However, **what we are ultimately interested in is controlling the target model by performing changes to it, based on our understanding of the proxy**. To do that, I first need to check that their outputs still match after I made changes (to both the target and proxy). In other words, we want the intervened target and intervened proxy to coincide in behavior after making edits to both.

An **intervention**[37] $\mathbb{J} : \mathbb{E} \to \mathfrak{I}(\mathbb{E})$ will consist of natural transformations for every explanation component $(\mathbb{J}_{\mathbb{M}}, \mathbb{J}_{\mathbb{X}}, \mathbb{J}_{\mathbb{Z}}, \mathbb{J}_{\mathbb{Y}})$

For simplicity, we assume Equation (5.3) and the analogous Equation (5.16), so $\mathbb{J}_{\mathbb{X}} = \mathrm{Id}_{\mathbb{X}}, \mathbb{J}_{\mathbb{Z}} = \mathbb{J}_{\mathbb{Y}}$.

The only two non-trivial transformations will then be:

- A causal-structural transformation $\mathbb{J}_{\mathbb{M}} : \mathbb{M} \to \mathfrak{I}(\mathbb{M})$
- A distributional transformation $\mathbb{J}_{\mathbb{Y}} : \mathbb{Y} \to \mathfrak{I}(\mathbb{Y})$

This transformation is <u>not lax</u> because when we make interventions, we are making explicit, intentional, and structurally-informed edits :

$$\mathfrak{I}(\mathbb{M}) \circ \mathbb{X} = \mathbb{J}_{\mathbb{Y}} \circ \mathbb{M} \circ \mathbb{X} \tag{6.10}$$

[37] an *edit*

The **edited explanation** are:

$$\mathfrak{I}(\mathbb{E}) = (\mathfrak{I}(\mathbb{M}), \mathbb{X}, \mathfrak{I}(\mathbb{Y})) \tag{6.11}$$

Consider that for $\mathbb{I} : \mathbb{E}_t \to \mathbb{E}_p$, we have:

$$\mathbb{J}_t : \mathbb{E}_t \to \mathfrak{I}_t(E_t) \tag{6.12}$$

$$\mathbb{J}_p : \mathbb{E}_p \to \mathfrak{I}_p\big(E_p\big) \tag{6.13}$$

The **interventional pairing** is then:

$$\boldsymbol{J} = \big(\mathbb{J}_t, \mathbb{J}_p\big) \tag{6.14}$$

The interpretation lifted through the interventions is the **interventional interpretation**:

$$\mathbb{I}^{\boldsymbol{J}} : \mathfrak{I}_t(\mathbb{E}_t) \to \mathfrak{I}_p\big(\mathbb{E}_p\big) \tag{6.15}$$

As any interpretation, it will have interpretation components $\mathbb{I}^{\boldsymbol{J}}_{\mathbb{M}}, \mathbb{I}^{\boldsymbol{J}}_{\mathbb{X}}, \mathbb{I}^{\boldsymbol{J}}_{\mathbb{Z}}, \mathbb{I}^{\boldsymbol{J}}_{\mathbb{Y}}$ and approximation components $\varphi^{\boldsymbol{J}}_{\mathbb{M}}, \varphi^{\boldsymbol{J}}_{\mathbb{X}}, \varphi^{\boldsymbol{J}}_{\mathbb{Z}}, \varphi^{\boldsymbol{J}}_{\mathbb{Y}}$.

Just as in Equation (5.13), to assess interventions, we want to "match up" the shape and see how values are different. Remember that to compare explanations is to see how different the interpreted target is to the proxy. To calculate approximation in interpretations, we compare how the non-edited. To calculate the defect in interventions, we compare how the edited explanations. The **defect transform** captures the interventional comparison:

$$\delta_{\bullet} : \mathbb{I}^{\boldsymbol{J}} \circ \mathbb{J}_t \to \mathbb{J}_p \circ \mathbb{I} \tag{6.16}$$

The **defect components** will then[38] be:

$$\delta_{\mathbb{M}} : \mathbb{I}_{\mathbb{Y}} \circ \mathfrak{I}_t(\mathbb{M}_{\mathrm{t}}) \to \mathfrak{I}_p\big(\mathbb{M}_{\mathrm{p}}\big) \tag{6.17}$$

$$\delta_{\mathbb{Y}} : \mathbb{I}_{\mathbb{Y}} \circ \mathbb{Y}_{\mathrm{t}} \to \mathbb{Y}_{\mathrm{p}} \tag{6.18}$$

[38] Again, to simplify analysis, we will assume Equation (5.3) and Equation (5.16) for both $\mathbb{I}$ and $\mathbb{I}^{\boldsymbol{J}}$.

with coherence:

$$\delta_{\mathbb{Y}} = \delta_{\mathbb{M}} \circ \mathbb{X} \tag{6.19}$$

We relate the defect transformation with the approximation transformations through conjugation, Equation (5.9):

$$\varphi^{\boldsymbol{J}}_{\bullet} = \big(\mathbb{J}_p \circ \_ \circ \mathbb{J}_t\big)[\varphi_{\bullet}] \cdot \delta_{\bullet} \tag{6.20}$$

This tells us that the approximation error we see in $\mathbb{I}^{\boldsymbol{J}}$ comes from both the original interpretation $\mathbb{I}$ approximation error φ (transported through the interventions), and defects δ from intervention causal-mismatch. We can calculate the **defect residual** as:

$$r_{\delta} = \| \ \varphi^{\boldsymbol{J}} - \big(\mathbb{J}_p \circ \_ \circ \mathbb{J}_t\big)[\varphi_{\bullet}] \| \tag{6.21}$$

where $\|\_\|$ is the metric used for predictive discrepancy (e.g., token-level KL-Divergence), and in most basic cases, we only consider the $\varphi_{\mathbb{Y}}$ component.

In the ideal situation where our defect transform is id, we would have *perfect interventional transport* [72]:

$$\mathbb{I}^{\boldsymbol{J}} \circ \mathbb{J}_t = \mathbb{J}_p \circ \mathbb{I} \tag{6.22}$$

With *perfect interventional transport*, $r_{\delta} = 0$

We can name the whole interventional setup as a *Interventional Study Arm* of $\mathbb{I}$:

$$\mathbb{A} = (\mathbb{I}, \boldsymbol{J}, ...) \tag{6.23}$$

We usually consider multiple interventions to get a fuller picture. We call an *Interventional Study* for $\mathbb{I}$, the set of study arms:

$$\mathbb{S} = \big\{\mathbb{A}_{\alpha}, \mathbb{A}_{\beta}, \mathbb{A}_{\gamma}, ...\big\} \tag{6.24}$$

A good interventional study will also include study arms that are formed by composing the interventions in other arms:

$$r_{\delta}^{\beta\circ\alpha} = \| \varphi_{\bullet}^{\boldsymbol{J}^{\beta}\circ\boldsymbol{J}^{\beta}} - \left(\left(\mathbb{J}_{p}^{\beta}\circ\mathbb{J}_{p}^{\alpha}\right)\circ\_\circ\left(\mathbb{J}_{t}^{\beta}\circ\mathbb{J}_{t}^{\alpha}\right)\right)[\varphi_{\bullet}]\| \tag{6.25}$$

The set of residuals for a study is $\boldsymbol{R}_{\mathbb{S}}$

Finally, we can establish that an **interpretation is ε-interventionally faithful** under $\mathbb{S}$ if

$$T(\boldsymbol{R}_{\mathbb{S}}) < \varepsilon \tag{6.26}$$

where T is a real-valued statistic like $E_{r_i\in\boldsymbol{R}_{\mathbb{S}}}[r_i]$ or $\max_{r_i\in\boldsymbol{R}_{\mathbb{S}}}(r_i)$.

What's the big picture? The strength of your claim of causal faithfulness of an interpretation will hinge on how well-designed your interventional study is. A few of my intuitions:

- Do *non-transportable* interventions too. Sometimes the proxy mechanism will not account for large parts of the target mechanism (Like when the proxy is a single circuit [73] part of a whole network). Ideally, **intervening on a non-transportable part of the target mechanism should be invisible to the proxy**. Quantify that if possible.
- If you are working in a more concrete environment, leverage their constructions. For instance, Equation (6.26) has an analogue in Causal Abstraction [49] as "Approximate Transformation".
- Get insights about the degrees of freedom of your mechanism. The more complex it is, the larger your study should be.

Counterfactual Faithfulness

With interventional faithfulness, our edited explanations have matching observations/predictions. However, sometimes proxy edits look very different from target edits. It seems we can make causal faithfulness stronger if the changes in the target map nicely and consistently with the changes in the proxy. We want this especially to be fulfilled when we fix the evaluation $\mathbb{X}$ between the target and the proxy.

Whereas in interventions, we were interested in estimating the difference between outputs: $\mathbb{I}^{\boldsymbol{J}} : \mathfrak{I}_t(\mathbb{E}_t) \to \mathfrak{I}_p(\mathbb{E}_p)$, we are now interested in estimating the **difference between edits themselves**. The edits in the proxy define a world, the edits in the target define another world. The **cross-world transformation** [74] would be:

$$\diamondsuit(\boldsymbol{J}) : \mathbb{J}_p \circ \mathbb{I} \Rightarrow \mathbb{I}^{\boldsymbol{J}} \circ \mathbb{J}_t \tag{6.27}$$

We are also interested in analyzing the **same case** over fixed conditions (prompt and interventions), so for a fixed index $i \in \mathcal{J}$, the **measurement functor** is:

$$\mathcal{M}_i : i \to \mathcal{Y} \qquad \mathcal{M}_i(\mathbb{E}) \coloneqq \mathbb{Y}_{\mathbb{E}}(i) \tag{6.28}$$

Using Equation (5.9), we can whisker to get a natural transformation $\chi^{(i)} : \mathbb{E}_t \to \mathcal{Y} \Rightarrow \mathbb{E}_t \to \mathcal{Y}$, which we will call the **cross-world comparator**:

$$\chi^{(\boldsymbol{J},i)} \coloneqq (\mathcal{M}_i \circ \_ \circ \mathrm{id})[\diamondsuit(\boldsymbol{J})] \tag{6.29}$$

This is telling us, considering concrete edits, how two worlds align.

To get something we can compute, let's look at the component for $\mathbb{E}_t$:

$$\chi^{(\boldsymbol{J},i)}(\mathbb{E}_t) = \mathcal{M}_i \circ \mathbb{J}_p \circ \mathbb{I} \circ \mathbb{E}_t \to \mathcal{M}_i \circ \mathbb{I} \circ \mathbb{J}_t \circ \mathbb{E}_t \tag{6.30}$$

Simplifying, we get the **effect gap transformation**:

$$\rho_e(\boldsymbol{J}, i) : \big(\mathfrak{I}_p\big(\mathbb{Y}_p\big)\big)(i) \to (\mathbb{I}_{\mathbb{Y}} \circ \mathfrak{I}_t(\mathbb{Y}_t))(i) \tag{6.31}$$

We can define the base observations:

$$y_t \coloneqq \mathbb{Y}_t(i) \qquad y_p \coloneqq \mathbb{Y}_p(i) \tag{6.32}$$

We can define the post-edit observations:

$$y_t^{\boldsymbol{J}} \coloneqq \mathfrak{I}_t(\mathbb{Y}_t)(i) \qquad y_p^{\boldsymbol{J}} \coloneqq \mathfrak{I}_p\big(\mathbb{Y}_p\big)(i) \tag{6.33}$$

We can define the cross observations:

$$y_p^{\times} \coloneqq \mathbb{I}_{\mathbb{Y}}(y_t) \qquad y_p^{\boldsymbol{J}\times} \coloneqq \mathbb{I}_{\mathbb{Y}}(y_t^{\boldsymbol{J}}) \tag{6.34}$$

We define the effect calculator as:

$$\Xi : \mathbb{Y} \times \mathbb{Y} \to \xi \tag{6.35}$$

where ξ is the effect space equipped with semi-norm $\|\_\|$. The contrast operator is then:

$$\triangle_{\xi}(y, y') \coloneqq \Xi(y, y') \tag{6.36}$$

With it, we can calculate the **effect residual** as:

$$r_{\text{cf}} = \| \triangle_{\xi} \left(y_p, y_p^{\boldsymbol{J}} \right) - \triangle_{\xi} \left(y_p^{\times}, y_p^{\boldsymbol{J}^{\times}} \right) \| \tag{6.37}$$

In the ideal situation where our cross-world comparator is id, we would have *perfect counterfactual transport*:

$$\begin{aligned} \forall J, i, \qquad & \chi^{(\boldsymbol{J},i)}(\mathbb{E}_t) = \text{id}, s \\ \forall J, i, \qquad & \mathcal{M}_i \circ \mathbb{J}_p \circ \mathbb{I} \circ \mathbb{E}_t = \mathcal{M}_i \circ \mathbb{I} \circ \mathbb{J}_t \circ \mathbb{E}_t \end{aligned} \tag{6.38}$$

With *perfect counterfactual transport*, $r_{\text{cf}} = 0$

Thus, we can establish that an **interpretation is ε-counterfactually faithful** under the interventional study $\mathbb{S}$ if

$$T(\left\{ r_{\text{cf}}\left(\boldsymbol{J}_{\rangle} : \boldsymbol{J} \in \mathbb{S} \right\}\right) < \varepsilon \tag{6.39}$$

where T is a real-valued statistic.

What's the big picture? Counterfactual faithfulness requires high causal alignment (aligned difference over differences). We are only able to form counterfactual evaluations after we have done interventions, so the same considerations about having a good study apply.

Mechanistic Faithfulness

We say an interpretation is **ε-causally faithful** if it is ε-observationally, ε-interventionally, and ε-counterfactually faithful. What does this say about my mechanism? Many mechanisms can fulfill the same causal structure. We need to look inside $\mathbb{M}$.

What's the big picture? Mechanical faithfulness is nuanced because we are below the causal identifiable limit. Mechanisms are often defined in particular categories, so we can try to find sensible constraints in those domains.

I would like us to consider what would happen if there were an irreducible amount of opacity [75], not only in these systems but also in ourselves [76].

7. Interpreting Circuit Tracing

In *Circuit Tracing* [38], Anthropic used Cross-Layer Transcoders (CLT) to build global and local replacement models for a target model. They also construct local attribution graphs and global-weights models. Each of these is an explanation, with interpretations in between. For this section, we will assume Equation (5.3) and Equation (5.15) to simplify analysis.

Explanations

The target is $\mathbb{E}_t = (\mathbb{M}_t, \mathcal{X}, \mathcal{Y})$ where $\mathbb{M}_t$ is the full mechanism of a language model of interest[39], $\mathcal{X}$ is the set of all possible input prompts based the model's token alphabet, $\mathcal{Y}$ is the set of next-token distributions.

$$\mathcal{X} = \{x_0, x_1, ...\} \qquad \mathcal{Y} = \{y_i = p(-|x_i) : x_i \in \mathcal{X}\} \tag{7.1}$$

The global replacement model is $\mathbb{E}_{\text{glob}} : (\mathbb{M}_{\text{glob}}, \mathcal{X}, \mathcal{Y})$, which has the same $\mathcal{X}$ evaluation and observation $\mathcal{Y}$ domains as $\mathbb{E}_t$ but a different mechanism, a modified of $\mathbb{M}_t$ where MLP layers are replaced by CLT [38].

The local replacement model is $\mathbb{E}_{\text{local}} : (m_{\text{local}}, x_{\text{local}}, y_{\text{local}})$, which not only has a different mechanism, but the evaluation and observation domains are different too. We have[40] a restriction map i that ignores all of $\mathcal{X}$ except a single one.

$$i : 1 \to \mathcal{I} \tag{7.2}$$

$$m_{\text{local}} = (\mathbb{M}_{\text{local}})_i \qquad x_{\text{local}} = \mathcal{X} \circ i \qquad y_{\text{local}} = \mathcal{Y} \circ i = p(-|x_{\text{local}}) \tag{7.3}$$

The attribution graph is $\mathbb{E}_{\text{att}} : (m_{\text{att}}, x_{\text{local}}, y_{\text{local}})$ where $m_{\text{att}} = (\mathbb{M}_{\text{att}})_i$.

The global-weights model is $\mathbb{E}_{\text{w}} : (\mathbb{M}_{\text{w}}, \mathcal{X}, \mathcal{Y})$.

All the mechanisms ($\mathbb{M}$) of these explanations are related by non-trivial transformations. To determine faithfulness, we can have causal-structural analysis (do mechanisms behave similarly to changes?) and a strict mechanistic analysis (what do the concrete transformations between mechanisms $\mathbb{I}_{\mathbb{M}}$ tell us about what behavior can be preserved)

We also have a non-injective map g.

[39] 18-layer LLM or Claude Haiku 3.5
[40] Look at Section 6 (Locality)

$$\mathrm{g} : \mathcal{Y} \to \mathbb{R} \quad \forall y_j \in \mathcal{Y} : g(y_j) = \max y_j = \mathrm{argmax}_{z \in \mathbb{R}}\, p(z|x_j) = y_{\max} \tag{7.4}$$

$$y_{\text{att}} = g(y_{\text{local}}) = \max p(-|x_{\text{local}}) \tag{7.5}$$

Interpretations

Each of the explanations will relate to the others through interpretations. We generally care about interpretations from opaque/faithful to understandable/unfaithful explanations. Sometimes, we have implicit interpretations we should be careful about. For instance, let's say we compare explanations by top-1 accuracy, then we have a non-injective map g:

$$g : \mathcal{Y} \to \mathbb{R} \quad \forall y_j \in \mathcal{Y} : g(y_j) = \max y_j = \mathrm{argmax}_{z \in \mathbb{R}}\, p(z|x_j) = y_{\max} \tag{7.6}$$

For instance, when we view the target model from top-1 accuracy view, we get[41]:

$$\mathbb{I}^{\text{top-1}} : \mathbb{E}_t : (\mathbb{M}_t, \mathcal{X}, \mathcal{Y}) \to \mathbb{E}_t^{\text{top-1}} : (\mathbb{M}_t\ /\ker(g), \mathcal{X}, \mathcal{Y}/\ker(g)) \tag{7.7}$$

This means that if we only use top-1 accuracy to determine faithfulness, we can only identify any mechanism up to equivalence by $\ker(g)$. In the paper, the authors show with Figure 3 that top-1 accuracy tracks the KL Divergence (which is more injective), maybe hinting that in practice, top-1 accuracy would identify as much structure as the KL Divergence.

Similarly, we could also try to compare feature maps instead of tokens. Then, we have an interpretation :

$$\mathbb{I}^f : \mathbb{E}_t : (\mathbb{M}_t, \mathcal{X}, \mathcal{Y}) \to \mathbb{E}_t^f : \left(\mathbb{M}_t^f \hookrightarrow \mathbb{M}_t, \mathcal{X}, \mathcal{F}\right) \tag{7.8}$$

We see that then our faithfulness claims are limited to the restricted mechanism $\mathbb{M}_t^f$. This also opens up the question of injectivity $\mathcal{F} \to \mathcal{Y}$.

Strict mechanistic view

Mechanistic faithfulness almost always requires further specification[42] to be established. We can, however, see what each mechanism transformation $\mathbb{I}_{\mathbb{M}}$ does, and see what types of mechanistic faithfulness are plausible for each.

We first have the construction of the global replacement model from the target model:

[41] Look at Section 6 (Lossiness)
[42] Check Section 6 (Mechanistic Faithfulness)

$$\mathbb{I}_{\mathbb{M}}^{t\to\ \mathrm{glob}} : \mathbb{M}_t \to \mathbb{M}_{\mathrm{glob}} \tag{7.9}$$

We immediately see that $\mathbb{I}_{\mathbb{M}}^{t\to\ \mathrm{glob}}$ cannot have high structural-mechanistic faithfulness if we consider network connectivity as part of the mechanism: $\mathbb{M}_{\mathrm{glob}}$ has a computational graph that is not graph-isomorphic to $\mathbb{M}_{\mathrm{t}}$. The global replacement model can provide output to all its subsequent layers, while the target does not have those connections.

To address this, we would need to prove that computational abstraction is (approximately) preserved in an input/output perspective. Ideally, we would also like it to be preserved at certain intermediates. At least, we would hope that computational alignment would monotonically increase with depth.

We then consider how the local replacement model was built in three steps:

$$\mathbb{I}_{\mathbb{M}}^{\mathrm{glob}\ \times t(x)\to\ \mathrm{local}} : \mathbb{M}_{\mathrm{glob}} \times \mathbb{M}_t \circ x_{\mathrm{local}} \to \mathbb{M}_{\mathrm{local}} \tag{7.10}$$

This means we modified $\mathbb{M}_{\mathrm{glob}}$ with *values* from $\mathbb{M}_t$ evaluated at a specific prompt x_{local}: evaluated attention patterns and normalization denominators. Viewed from the perspective of $\mathbb{M}_{\mathrm{glob}}$, this transformation does not have any obvious issues with mechanical faithfulness. We could even conceive of it as a *soft intervention* [49]:

$$I \tag{7.11}$$

We take the replacement model to form the local replacement model $\mathbb{I}^{\mathrm{repl}\ \to\ \mathrm{local}}$. The local replacement model restricts evaluation to a single prompt, gets the frozen attention and normalization from $\mathbb{M}_{\mathrm{t}}$, and its mechanism gets help from error adjustment nodes.

Causal-structural view

One of the evaluations for the global replacement model is the next-token completion match. This evaluation implicitly interprets both explanations with a non-injective map on their observations $\mathbb{I}_{\mathbb{Y}}^{\mathrm{eval}} : \mathcal{Y} \to y_{\mathrm{peak}}$. It identifies next-token distributions if they have the same peak, similar to the lossy situation in Equation (6.6).

$$\begin{aligned} (\mathbb{M}_{\mathrm{t}}, \mathcal{X}, \mathcal{Y}) &\to \left(\mathbb{M}_{\mathrm{t}}, \mathcal{X}, y_{\mathrm{peak}}\right) \\ (\mathbb{M}_{\mathrm{CLT}}, \mathcal{X}, \mathcal{Y}) &\to \left(\mathbb{M}_{\mathrm{CLT}}, \mathcal{X}, y_{\mathrm{peak}}\right) \end{aligned} \tag{7.12}$$

The authors also evaluated with it the KL divergence, which does not have this effect.

8. Wayfinding

Thank you for engaging with this unruly abstraction, in whichever form you did.

Why did I write this paper, truly? Because I wished someone else had already. I love learning about interpretability, and I am often lost. I have a recently-developed-yet-strong taste for category theory. It helps me understand more than math, my queerness in a way. So, in a way, I wrote this paper so you could notice all the ways I am wrong and then write a much better category-theoretic interpretability paper, one that goes way beyond wanderings, one with data and proofs, one that I will definitely to read.

If I got you in the mood to categorify "something", I also have one suggestion (for my own curious appetite): Can we think categorically about features? Are they metric spaces [77]? Are they properties that activate particular mechanisms [78]? What structure is induced by their associated level of abstraction [79]? How do they all map [80] from and into R^k of the residual stream? They might not be proper manifolds [39], [81]? Are they some curvature [82]? What are they?

This piece presented some of my mental models around interpretability. As I mentioned before, I did not intend to give you a tidy, provably correct, data-driven cartographic map [83]. Instead, I wanted to show you how I **feel my way** *through* this field. And then threw a lot of category theory and notation at you. I am sorry. I got too excited,

Keep wayfinding.

Citation Information

Use the following to cite this piece:

```bibtex
@article{sialer2025catinterpretwander,
    author  = {Ian Rios-Sialer},
    title   = {Category-Theoretic Wanderings in Interpretability},
    journal = {LessWrong / Substack},
    year    = {2025},
    url     = {https://www.lesswrong.com/users/unruly-abstractions}
  }
```


Acknowledgement

I want to thank my amazing friend Abdul Wasay for feedback and revision. Also, I used ChatGPT and Claude to give me feedback as I was writing and to explore concepts.

References

[1] a. m. brown, F. Rodriguez, and L. L. Piepzna-Samarasinha, *Pleasure Activism: The Politics of Feeling Good*. in Emergent Strategy. AK Press, 2019. [Online]. Available: https://books.google.com/books?id=wlJUDwAAQBAJ

[2] O. L. Haimson, *Trans Technologies*. MIT Press, 2025. [Online]. Available: https://books.google.com/books?id=MQ0NEQAAQBAJ

[3] S. Ahmed, *Queer Phenomenology: Orientations, Objects, Others*. Duke University Press, 2006. [Online]. Available: https://books.google.com/books?id=sQY1RWdUW0AC

[4] S. Harney and F. Moten, *The Undercommons: Fugitive Planning & Black Study*. Minor Compositions, 2013. [Online]. Available: https://books.google.com/books?id=M9VuAQAACAAJ

[5] M. Zao-Sanders, "How People Are Really Using Gen AI in 2025." [Online]. Available: https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

[6] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning." [Online]. Available: https://arxiv.org/abs/1702.08608

[7] T. Freiesleben and G. König, "Dear XAI Community, We Need to Talk! Fundamental Misconceptions in Current XAI Research." [Online]. Available: https://arxiv.org/abs/2306.04292

[8] D. Tan, "Mech Interp Lacks Good Paradigms." [Online]. Available: https://www.lesswrong.com/posts/3CZF3x8FX9rv65Brp/mech-interp-lacks-good-paradigms

[9] D. Broniatowski, "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence." [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931426

[10] N. Saphra and S. Wiegreffe, "Mechanistic?." [Online]. Available: https://arxiv.org/abs/2410.09087

[11] A. Erasmus, T. D. Brunet, and E. Fisher, "What is interpretability?," *Philosophy & Technology*, vol. 34, no. 4, pp. 833–862, 2021.

[12] J. Ji *et al.*, "AI Alignment: A Comprehensive Survey." [Online]. Available: https://arxiv.org/abs/2310.19852

[13] R. Greenblatt, B. Shlegeris, K. Sachan, and F. Roger, "AI Control: Improving Safety Despite Intentional Subversion." [Online]. Available: https://arxiv.org/abs/2312.06942

[14] D. Amodei, "The Urgency of Interpretability." [Online]. Available: https://www.darioamodei.com/post/the-urgency-of-interpretability

[15] L. Bereska and E. Gavves, "Mechanistic Interpretability for AI Safety – A Review." [Online]. Available: https://arxiv.org/abs/2404.14082

[16] C. Olah, "Interpretability Dreams: An informal note on future goals for mechanistic interpretability." [Online]. Available: https://transformer-circuits.pub/2023/interpretability-dreams/index.html

[17] N. Nanda, "Interpretability Will Not Reliably Find Deceptive AI." [Online]. Available: https://www.alignmentforum.org/posts/PwnadG4BFjaER3MGf/interpretability-will-not-reliably-find-deceptive-ai

[18] C. Singh, J. P. Inala, M. Galley, R. Caruana, and J. Gao, "Rethinking Interpretability in the Era of Large Language Models." [Online]. Available: https://arxiv.org/abs/2402.01761

[19] E. Riehl, *Category Theory in Context*. in Aurora: Dover Modern Math Originals. Dover Publications, 2017. [Online]. Available: https://books.google.com/books?id=6B9MDgAAQBAJ

[20] A. L. Tsing, *The Mushroom at the End of the World: On the Possibility of Life in Capitalist Ruins*. Princeton University Press, 2015. [Online]. Available: https://books.google.com/books?id=tLlKCAAAQBAJ

[21] J.-P. Marquis, "Category Theory," *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2023.

[22] F. R. Crescenzi, "Towards a Categorical Foundation of Deep Learning: A Survey." [Online]. Available: https://arxiv.org/abs/2410.05353

[23] E. Cheng, *The Joy of Abstraction: An Exploration of Math, Category Theory, and Life*. Cambridge University Press, 2022. [Online]. Available: https://books.google.com/books?id=N_GCEAAAQBAJ

[24] J. Goedecke, "Category Theory: Lecture Notes," 2013.

[25] T.-D. Bradley, "The Yoneda Perspective." [Online]. Available: https://www.math3ma.com/blog/the-yoneda-perspective

[26] L. Sharkey *et al.*, "Open Problems in Mechanistic Interpretability." [Online]. Available: http://arxiv.org/abs/2501.16496

[27] W. Saeed and C. Omlin, "Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities." [Online]. Available: https://arxiv.org/abs/2111.06420

[28] B. Gavranović, P. Lessard, A. Dudzik, T. von Glehn, J. G. M. Araújo, and P. Veličković, "Position: Categorical Deep Learning is an Algebraic Theory of All Architectures." [Online]. Available: https://arxiv.org/abs/2402.15332

[29] A. Zou *et al.*, "Representation Engineering: A Top-Down Approach to AI Transparency." [Online]. Available: https://arxiv.org/abs/2310.01405

[30] N. Elhage *et al.*, "Toy Models of Superposition." [Online]. Available: https://transformer-circuits.pub/2022/toy_model/index.html

[31] F. R. Genovese, “Modularity vs Compositionality: A History of Misunderstandings.” [Online]. Available: https://medium.com/statebox/modularity-vs-compositionality-a-history-of-misunderstandings-be0150033568

[32] J. M. Hedges, “Towards compositional game theory,” 2016.

[33] B. Gavranović, “Why Category Theory?.” 2022.

[34] B. Fong and D. I. Spivak, “Seven Sketches in Compositionality: An Invitation to Applied Category Theory.” [Online]. Available: https://arxiv.org/abs/1803.05316

[35] T.-D. Bradley, “What is Applied Category Theory?.” [Online]. Available: http://arxiv.org/abs/1809.05923

[36] C. Puca, A. Hadzihasanovic, F. Genovese, and B. Coecke, “Obstructions to Compositionality,” *Electronic Proceedings in Theoretical Computer Science*, vol. 397, pp. 226–245, Dec. 2023, doi: 10.4204/EPTCS.397.14.

[37] J. Lindsey *et al.*, “On the Biology of a Large Language Model.” [Online]. Available: https://transformer-circuits.pub/2025/attribution-graphs/biology.html

[38] E. Ameisen *et al.*, “Circuit Tracing: Revealing Computational Graphs in Language Models.” [Online]. Available: https://transformer-circuits.pub/2025/attribution-graphs/methods.html

[39] M. Robinson, S. Dey, and T. Chiang, “Token embeddings violate the manifold hypothesis.” [Online]. Available: https://arxiv.org/abs/2504.01002

[40] AI Alignment Forum, ““Negative Results for SAEs on Downstream Tasks and Deprioritising…” – SAE Progress Update #2 (Draft).” [Online]. Available: https://www.alignmentforum.org/posts/4uXCAJNuPKtKBsi28/negative-results-for-saes-on-downstream-tasks

[41] A. Jermyn, “Activation space interpretability may be doomed.” [Online]. Available: https://www.alignmentforum.org/posts/gYfpPbww3wQRaxAFD/activation-space-interpretability-may-be-doomed

[42] J. Mendel, “SAE Feature Geometry is Outside the Superposition Hypothesis.” [Online]. Available: https://www.lesswrong.com/posts/MFBTjb2qf3ziWmzz6/sae-feature-geometry-is-outside-the-superposition-hypothesis

[43] S. De Toffoli, "'Chasing' the diagram—the use of visualizations in algebraic reasoning," *The Review of Symbolic Logic*, vol. 10, no. 1, pp. 158–186, 2017.

[44] B. Victor, "Media for Thinking the Unthinkable." [Online]. Available: http://worrydream.com/MediaForThinkingTheUnthinkable/

[45] F. W. Lawvere and S. H. Schanuel, *Conceptual Mathematics: A First Introduction to Categories*. Cambridge University Press, 2009. [Online]. Available: https://books.google.com/books?id=h0zOGPlFmcQC

[46] G. M. Kelly, *Basic Concepts of Enriched Category Theory*. in London Mathematical Society Lecture Note Series. Cambridge University Press, 1982.

[47] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks, "AI Deception: A Survey of Examples, Risks, and Potential Solutions." [Online]. Available: https://arxiv.org/abs/2308.14752

[48] L. Weng, "Why We Think." [Online]. Available: https://lilianweng.github.io/posts/2025-05-01-thinking/

[49] A. Geiger *et al.*, "Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability." [Online]. Available: http://arxiv.org/abs/2301.04709

[50] Y. Zhang and M. Sugiyama, "A Category-theoretical Meta-analysis of Definitions of Disentanglement," in *Proceedings of the 40th International Conference on Machine Learning*, in Proceedings of Machine Learning Research, vol. 202. PMLR, Jul. 2023, pp. 41596–41612. [Online]. Available: https://proceedings.mlr.press/v202/zhang23ak.html

[51] Y. Chen, Z. Zhou, and J. Yan, "Going Beyond Neural Network Feature Similarity: The Network Feature Complexity and Its Interpretation Using Category Theory." [Online]. Available: http://arxiv.org/abs/2310.06756

[52] B. Gavranović, "Fundamental Components of Deep Learning: A category-theoretic approach." [Online]. Available: http://arxiv.org/abs/2403.13001

[53] Y. Jia, G. Peng, Z. Yang, and T. Chen, "Category-Theoretical and Topos-Theoretical Frameworks in Machine Learning: A Survey." [Online]. Available: https://arxiv.org/abs/2408.14014

[54] D. Shiebler, B. Gavranović, and P. Wilson, "Category Theory in Machine Learning." [Online]. Available: http://arxiv.org/abs/2106.07032

[55] V. Abbott, T. Xu, and Y. Maruyama, "Category Theory for Artificial General Intelligence," in *Artificial General Intelligence*, K. R. Thórisson, P. Isaev, and A. Sheikhlar, Eds., Cham: Springer Nature Switzerland, 2024, pp. 119–129.

[56] N. P. Shaw, P. M. Furlong, B. Anderson, and J. Orchard, "Developing a foundation of vector symbolic architectures using category theory." [Online]. Available: https://arxiv.org/abs/2501.05368

[57] D. Ghosh, D. Ghosh, and D. P. Ghosh, "Think in Arrows: A Categorical Scaffolding Framework for Robust Artificial Scientific Discovery," Apr. 2025, doi: 10.13140/RG.2.2.16950.41280.

[58] T.-D. Bradley and J. P. Vigneaux, "The Magnitude of Categories of Texts Enriched by Language Models." [Online]. Available: http://arxiv.org/abs/2501.06662

[59] T.-D. Bradley, J. Terilla, and Y. Vlassopoulos, "An enriched category theory of language: from syntax to semantics." [Online]. Available: https://arxiv.org/abs/2106.07890

[60] J. Ferrando, G. Sarti, and M. R. Costa-jussà, "A Primer on the Inner Workings of Transformer-based Language Models." [Online]. Available: https://arxiv.org/abs/2405.00208

[61] X.-K. Wu *et al.*, "LLM Fine-Tuning: Concepts, Opportunities, and Challenges," *Big Data and Cognitive Computing*, vol. 9, no. 4, p. 87, 2025.

[62] P. M. Pietroski, "Conjoining Meanings: Semantics Without Truth Values," *Conjoining Meanings: Semantics Without Truth Values*. Oxford University Press, 2018. doi: 10.1093/oso/9780198812722.001.0001.

[63] T. Y. Liu, M. Trager, A. Achille, P. Perera, L. Zancato, and S. Soatto, "Meaning Representations from Trajectories in Autoregressive Models." [Online]. Available: https://arxiv.org/abs/2310.18348

[64] E. K. Sedgwick, *Epistemology of the Closet*. Berkeley: University of California Press, 1990.

[65] G. C. Spivak, "Can the Subaltern Speak?," *Marxism and the Interpretation of Culture*. Macmillan, Basingstoke, pp. 271–313, 1988.

[66] C. Agarwal, S. H. Tanneru, and H. Lakkaraju, "Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models." [Online]. Available: https://arxiv.org/abs/2402.04614

[67] D. Bessis and K. Frey, *Mathematica: A Secret World of Intuition and Curiosity*. Yale University Press, 2024. [Online]. Available: https://books.google.com/books?id=jYQBEQAAQBAJ

[68] J. Halberstam, *The Queer Art of Failure*. in A John Hope Franklin Center Book. Durham, NC: Duke University Press, 2011. doi: 10.1215/9780822394358.

[69] R. Dhar, A. Karamolegkou, and A. Søgaard, "Toward a Sheaf-Theoretic Understanding of Compositionality in Large Language Models." [Online]. Available: https://openreview.net/forum?id=srOVvTzgPo

[70] J. Wehner, S. Abdelnabi, D. Tan, D. Krueger, and M. Fritz, "Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models." [Online]. Available: https://arxiv.org/abs/2502.19649

[71] J. Gu, B. Hua, and S. Liu, "Spectral distances on graphs," *Discrete Applied Mathematics*, pp. 56–74, Aug. 2015, doi: 10.1016/j.dam.2015.04.011.

[72] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, 2016, doi: 10.1073/pnas.1510507113.

[73] M. Hanna, S. Pezzelle, and Y. Belinkov, "Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms." [Online]. Available: https://arxiv.org/abs/2403.17806

[74] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, 2009. [Online]. Available: https://projects.illc.uva.nl/cil/uploaded_files/inlineitem/Pearl_2009_Causality.pdf

[75] G. D'Acunto and C. Battiloro, "The Relativity of Causal Knowledge." [Online]. Available: https://arxiv.org/abs/2503.11718

[76] A. Saketopoulou, *Sexuality Beyond Consent: Risk, Race, Traumatophilia*. NYU Press, 2023. [Online]. Available: https://books.google.com/books?id=Xb6ZEAAAQBAJ

[77] A. Modell, P. Rubin-Delanchy, and N. Whiteley, "The Origins of Representation Manifolds in Large Language Models." [Online]. Available: http://arxiv.org/abs/2505.18235

[78] D. Braun, L. Bushnaq, S. Heimersheim, J. Mendel, and L. Sharkey, "Interpretability in Parameter Space: Minimizing Mechanistic Description Length with Attribution-based Parameter Decomposition." [Online]. Available: http://arxiv.org/abs/2501.14926

[79] E. Cheng *et al.*, "Emergence of a High-Dimensional Abstraction Phase in Language Transformers." [Online]. Available: https://arxiv.org/abs/2405.15471

[80] O. Skean *et al.*, "Layer by Layer: Uncovering Hidden Representations in Language Models." [Online]. Available: https://arxiv.org/abs/2502.02013

[81] M. Robinson, S. Dey, and S. Sweet, "The structure of the token space for large language models." [Online]. Available: https://arxiv.org/abs/2410.08993

[82] R. Manson, "Curved Inference: Concern-Sensitive Geometry in Large Language Model Residual Streams." [Online]. Available: https://arxiv.org/abs/2507.21107

[83] T. Ingold, *The Perception of the Environment: Essays on Livelihood, Dwelling and Skill*. Routledge, 2000. [Online]. Available: https://books.google.com/books?id=nc1HZxsyZgIC