Category-Theoretic Wanderings into Interpretability

Unruly Abstractions

September 11, 2025

Ian Rios-Sialer

Independent

ORCID: 0009-0001-6970-6058

ian@unrulyabstractions.com

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.

I promise I am normally more shy about sharing such intimacies[1] but chances are you also use your Large Language Models (LLMs) somewhat alike[2]. 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 [7]. As a pre-paradigmatic field [8], [9], terminology and definitions are a bit confusing and/or inconsistent [10], [11], [12]. A central desideratum of interpretability is to contribute meaningfully to AI Alignment [13], AI Control [14], and AI Safety [1], [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.[3]

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 somewhat speculative epistemic status, and they are not meant to present clear contributions (yet). Instead, 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 **abuse notation** to 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.

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] We often start conversations in the context of future Artificial Super Intelligence [1] and its risks. Instead, I would like to first ground ourselves in what is familiar, felt [2] and present; hopefully inviting many more different types [3] of people to the table [4], to engage [5] with the development of a technology that will profoundly impact them sooner than they think.

[2] Personal support (emotional application), like therapy, was the top use case of Gen AI in 2025 [6]

[3] 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

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**[4] is [23], [24], [25] to reason considering some of these steps:

- 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

Interpretability can benefit from thinking more *categorically*

Why is thinking categorically useful? Because it can serve as a tool[5] 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 [9], [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][6]. 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 measure the failures of compositionality [36]. **Imagine it as learning to cook. Sometimes, the**

[4] or "category-theoretically"

[5] A thinking technology?

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

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 [9], [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 [45]. **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[7] [19], [23], [24], [46] and some more intermediate ones [34], [47] 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):

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

With that[8], the CringeShows category consists of the following **Data:**

1. A collection ob(CringeShows) of *objects*, the elements of the set TVShows
2. For every $x, y \in \text{ob}(\text{CringeShows})$, a set of *morphisms*[9] $[x, y]_{\text{CringeShows}}$[10] , in our case, such set will have single element when $\text{Cringe}(x) \leq \text{Cringe}(y)$ and be the empty set $\varnothing$ otherwise.

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

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

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

[10] $[x, y]_C$ could also be written as $\text{Hom}_C(x, y)$, $C(x, y)$ or just $[x, y]$

With the following **Structure**:

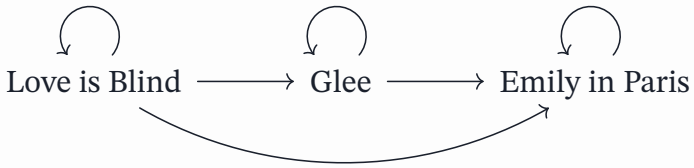
1. An identity morphism for every object, $x \xrightarrow{\mathrm{id}_x} x$ which is $x \leq x$
2. A composition operation such that any two morphisms like $x \xrightarrow{f} y$ and $y \xrightarrow{g} z$ produce $x \xrightarrow{g \circ f} z$, which means $x \leq y$ and $y \leq z$ imply $x \leq z$

With the following **Properties**:

1. Unit Law, such that given $x \xrightarrow{f} y$, $f = \mathrm{id}_y \circ f = f \circ \mathrm{id}_x$
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:

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



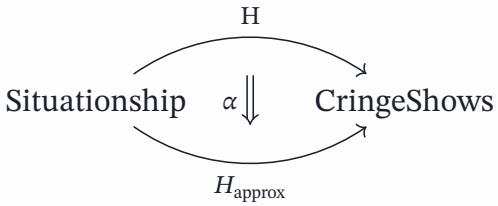
I have drawn identity arrows and compositions, but those are usually not drawn as they can be inferred. Our category is called "thin" because we have at most one arrow between any two objects. Other categories could have many more.

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 Situationship category.

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

We are going up a level of abstraction[11] and treating categories like objects themselves. F and G are called functors, "morphisms between categories"[12]. 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[13]:



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[14] coming in and out of Mood. We can observe them, perform interventions [50], 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[15]. And eventually, take action[16].

The following sections will be more technical in nature[17]. I will build up from specific papers. I will strive to provide the high-level picture, but if you are interested in the details, I strongly encourage you to read the particular papers I will build upon.

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

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

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

[14] I can arguably also communicate my feelings. To some faithful [48] extent, anyway [49].

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

[16] Yes, this means stop texting him.

[17] I also suggest reading the appendix right now if you want more build-up before diving in

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* [59], [60], 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 [61]:

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

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:

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

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

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:

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

[18] Really many [22], [28], [51], [52], [53], [54], [55], [56], [57], [58]

How do we categorize this? Let's set aside the probabilities for a moment and focus on what they act on: words. How do we form sentences, texts,... from words? The base language category will provide us with the 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: $s_i \in \text{Seq}_s$ as in Equation (3.2)
- Morphisms: $[s_i, s_j]$ is singleton set if prefix relation as in Equation (3.5), empty otherwise.

This category is a thin category similar to the one we had in Section 2 (What is a category?)! Let's visualize it. For simplicity, let's assume 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 an 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. [59] will give you technical details of this enrichment and $[0, 1]$-categories, so that 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: The hom-set $[x, y]_{L_{\text{base}}}$ is enriched such that $L_{\text{syn}}(x, y) \coloneqq \pi(y|x)$ is the probability that y extends x defined as:

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

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:

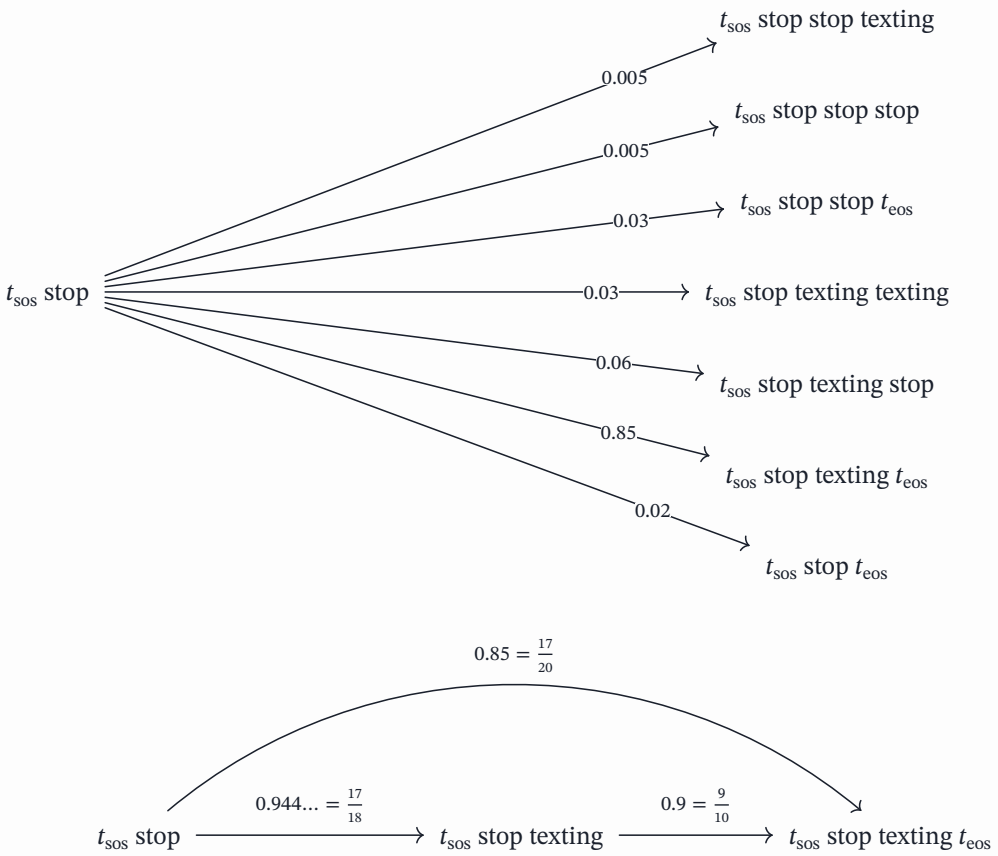


Figure 2: **Top:** Showing total probabilities in L_{syn} for given prompt.
Bottom: Showing compositionality of $\pi(y \mid x)$

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 [62], 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 the others. To pry on possible structure, let's go back to my situation(ship)s for a second.

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

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 disregard 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 examine what the LLM tells me when I ask "Should I text Mahdi?" vs "Should I text Mark?" ?

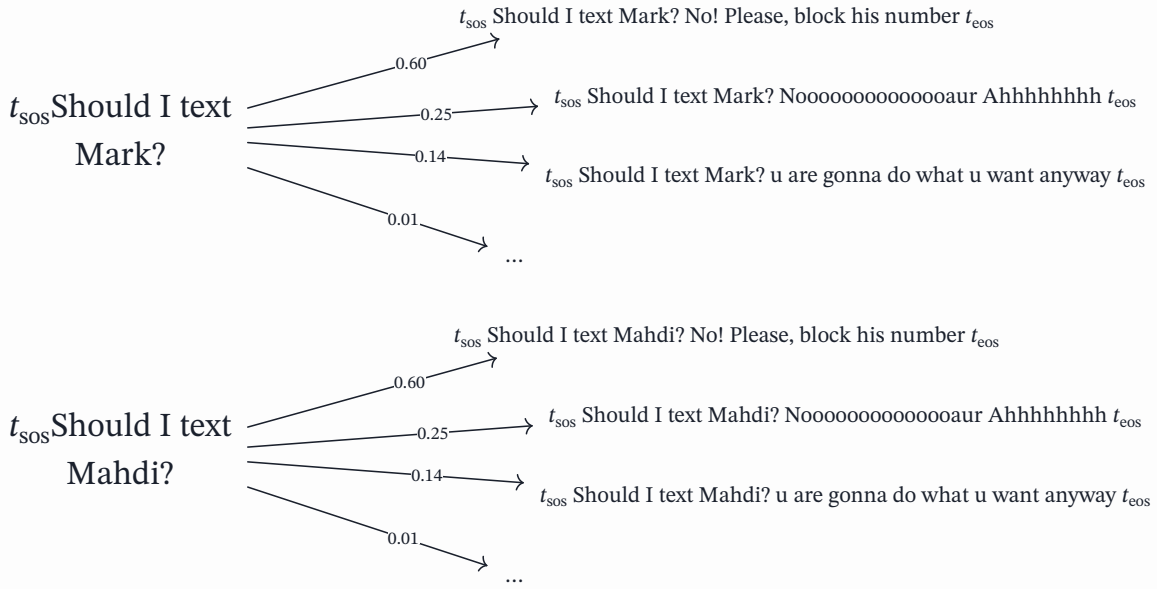


Figure 3: Two prompts having the same distribution of continuations

What is happening? We have two prompts $x_p \coloneqq t_{sos}$Should I text Mark? and $x_q \coloneqq t_{sos}$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 that both prompts are *observationally equivalent*. We can use the [0, 1]-enriched Yoneda Embedding from *Bradley et al* [60] to investigate what that equivalence looks like more explicitly. We have a functor from L_{syn} to another category, L_{sem}.

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

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

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

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* [63] through this contextual constraint.

I recommend you check out *Bradley et al* [59], [60] for the technical details[20]. We have quickly introduced a lot, so let's see what this looks like in my example:

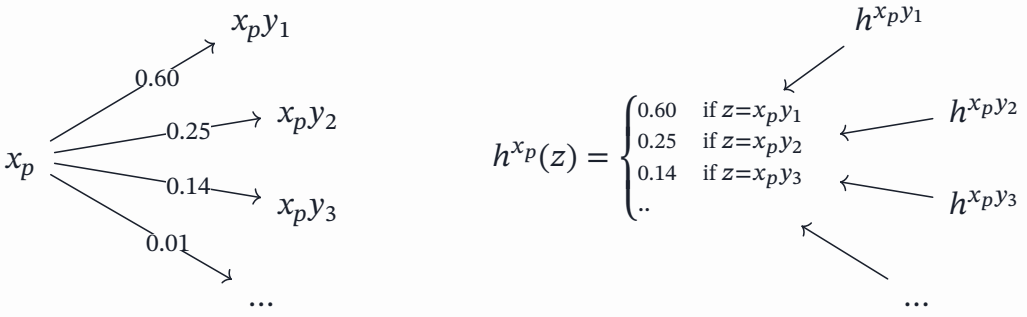


Figure 4: Applying the Yoneda Embedding to "Should I text Mark?" prompt from Figure 3

[20] To define the base language category, [60] considers substrings while [59] considers prefixes (which directly connect to LLMs; what we use in this piece). Just keep that in mind as you read through those papers.

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* [64]: 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]:

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

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

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

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:

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

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

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

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 $d_{\mathrm{chain}} = \inf_{(x=v_0,\ldots,v_n=y)} \sum_i d_s(v_i, v_{i+1})$. 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 <u>new</u> 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:

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

[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}.

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 examine 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.

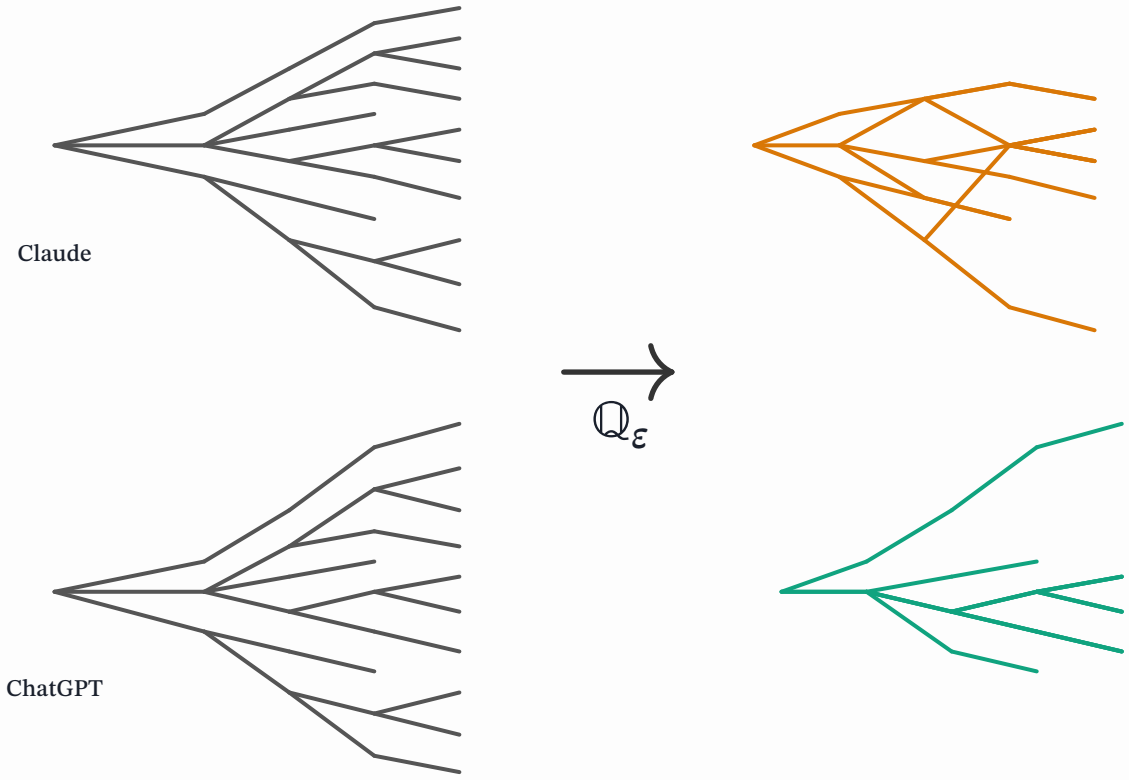


Figure 5: Imagining different induced topology for the same prompt

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 (i.e, the real structure persists and remains stable) and interventions to test causality (i.e, the organization of causal effects is preserved). We will delve into that in the following sections[24].

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 examine our objects from another viewpoint ($\mathbb{Q}_\varepsilon$). From that angle, we asked ourselves what things we could reason about, based on our original objects (observational identification by topology?). **Why did we do this?** To find interesting ways to compare objects 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.

So, why do we make comparisons when we interpret?
...what is to interpret, anyway?

[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 are curious about where they reappear

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. [10].

As such, understandability is situational and oriented[25]. Any explanation $\mathbb{E}$ has multiple associated understandabilities [12] Und, depending on who the subject is in what context:

$$\ldots \leq \mathsf{Und}_x(\mathbb{E}_i) \leq \mathsf{Und}_x(\mathbb{E}_j) \leq \ldots$$

$$\Delta \text{ context } \Downarrow$$

$$\ldots \leq \mathsf{Und}_y(\mathbb{E}_j) \leq \mathsf{Und}_y(\mathbb{E}_i) \leq \ldots$$

For instance, a weather forecast that uses *Fahrenheit* instead of *Celsius* would be less understandable to me than it would be for someone who grew up in the US.

The process of interpretation [12] is to map a less understandable explanation to a more understandable one:

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

What's an explanation?

An **explanation** is a representation of a mechanism that connects particular evaluations of it to observations :

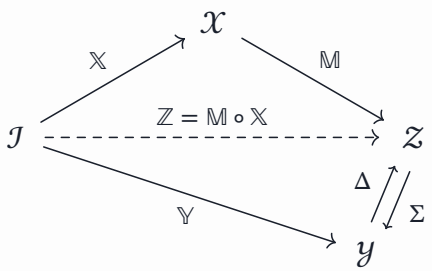


Figure 6:

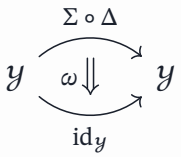
We can read the explanation diagram as:
- $\mathbb{X}$ selects an evaluation from $\mathcal{X}$
- $\mathbb{Y}$ selects an observation from $\mathcal{Y}$
- $\mathbb{M}$ mechanism produces a prediction in $\mathcal{Z}$ based on the selected evaluation.
- The observation maps to the prediction by analysis Δ, and vice versa by synthesis Σ.

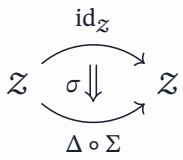
[25] When we also consider to whom what understandings are more reachable, understandability is also political [4]

Formally[26], 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 $\mathcal{I}$
- A mechanism functor $\mathbb{M} : \mathcal{X} \to \mathcal{Z}$
- An evaluation functor $\mathbb{X} : \mathcal{I} \to \mathcal{X}$
- An observation functor $\mathbb{Y} : \mathcal{I} \to \mathcal{Y}$
- Such that, an **explanation is the triplet**[27] $\mathbb{E} = (\mathbb{M}, \mathbb{X}, \mathbb{Y})$ that is stitched by:
  - A prediction functor: $\mathbb{Z} = \mathbb{M} \circ \mathbb{X}$
  - An analysis functor: $\Delta : \mathcal{Y} \to \mathcal{Z}$
  - A synthesis functor: $\Sigma : \mathcal{Z} \to \mathcal{Y}$

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. $\mathcal{X}$, $\mathcal{Z}$ and $\mathcal{Y}$ are also called the data categories.

The degree to which observations match predictions (and vice versa) is the degree of uncertainty and incompleteness [7] of the explanation. Going back and forth from observation to prediction (through Δ and Σ) introduces error. To capture that error, which is internal to an explanation, we could define:
- A precision unit $\omega : \mathrm{id}_Y \Rightarrow \Sigma \circ \Delta$ and counit $\sigma : \Delta \circ \Sigma \Rightarrow \mathrm{id}_Z$
- A resolution unit: $\sigma^* : \mathrm{id}_Z \Rightarrow \Delta \circ \Sigma$ and counit $\omega^* : \Sigma \circ \Delta \Rightarrow \mathrm{id}_Y$



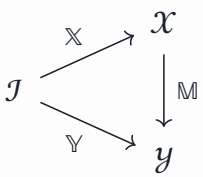


If these units and counits exist and are isomorphic, there is an equivalence $\mathcal{Y} \simeq \mathcal{Z}$

We often want to consider explanations that can match perfectly each prediction to its own observation. That is the *Infinite Fidelity Assumption*: We assume there is no noise and perfect resolution within an explanation, which leads to an isomorphism between predictions and observations:

$$\mathrm{id}_{\mathcal{Y}} = \Sigma \circ \Delta, \quad \Delta \circ \Sigma = \mathrm{id}_{\mathcal{Z}} \quad \to \quad \mathcal{Y} \cong \mathcal{Z} \tag{5.2}$$

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



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

[26] We could make this definition more general by using profunctors [47]

[27] In literature [12], you'll rather see Mechanism called Explanan, Observation as Explanandum, and Evaluation as Process of Explanation.

Scope is the extent of validity of an explanation. We can think of $\mathcal{X}$ as defining the entire possible domain and $\mathcal{I}$ as 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 :

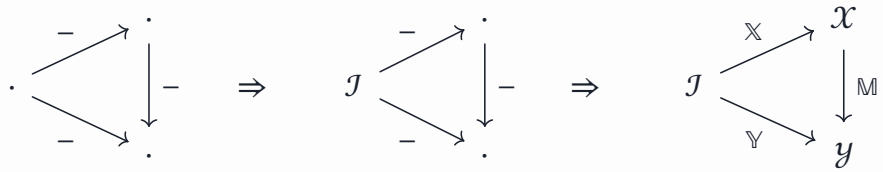
- $\mathbb{E}_G$ will have $\mathcal{I} = \mathbb{R}$. In this case, the scope of our explanation covers all possible inputs.
- $\mathbb{E}_L$ will have $\mathcal{I} = \{1\}$. In this case, the scope of our explanation encompasses a single number from a much larger possible domain $\mathbb{R}$.

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 category allows us to pick a particular value from all possible values in the input category. This "selection" of an explanation is an object of the functor category $[\mathcal{S}, \mathsf{Cat}]$ where Cat is the category of all categories. Each explanation will have a "selection". We notice that explanations all connect at an index category, so our selections are more specifically objects of $[\mathcal{S}, [\mathcal{I}, \mathsf{Cat}]]$. In Equation (5.3) case, this would be:



The **diagram selection operator** is then:

$$\text{diag} : \text{Expl} \to [\mathcal{S}, [\mathcal{I}, \mathsf{Cat}]] \tag{5.4}$$

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

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

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

To help us with technicalities, we define a *conjugation*. Given $F, G : C \to D$, $R : V \to C$, $L : D \to W$ functors, and $\alpha : F \Rightarrow G$ natural transformation, we have the natural transformation:

$$(L \circ \_ \circ R)[\alpha] : \quad L \circ F \circ R \Rightarrow L \circ G \circ R \tag{5.6}$$

What's an interpretation?

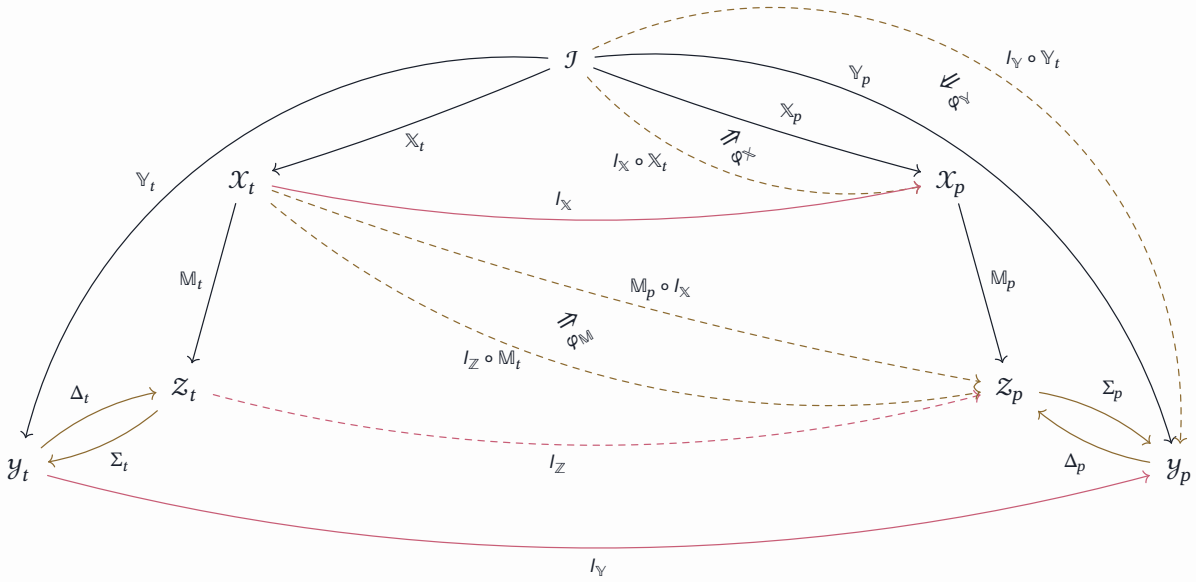


Figure 7: To interpret a target explanation ($\mathbb{E}_t = \mathbb{E}_{\text{target}}$) is to relate it to a proxy explanation ($\mathbb{E}_p = \mathbb{E}_{\text{proxy}}$)

An **interpretation** $\mathbb{I} : \mathbb{E}_{\text{target}} \to \mathbb{E}_{\text{proxy}}$ consists of transport functors and comparison 2-cells[29]:

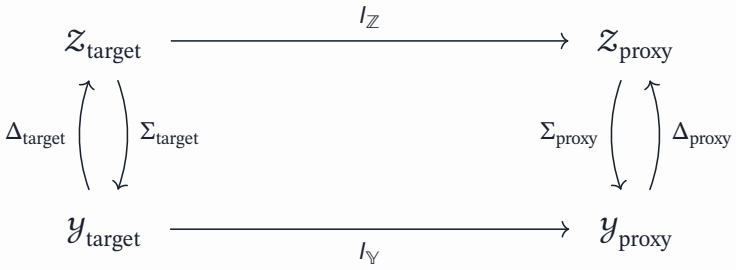
$$I_{\mathbb{X}} : \mathcal{X}_{\text{target}} \to \mathcal{X}_{\text{proxy}} \qquad \varphi_{\mathbb{X}} : I_{\mathbb{X}} \circ \mathbb{X}_{\text{target}} \Rightarrow \mathbb{X}_{\text{proxy}}$$
$$I_{\mathbb{Y}} : \mathcal{Y}_{\text{target}} \to \mathcal{Y}_{\text{proxy}} \qquad \varphi_{\mathbb{Y}} : I_{\mathbb{Y}} \circ \mathbb{Y}_{\text{target}} \Rightarrow \mathbb{Y}_{\text{proxy}}$$
$$\varphi_{\mathbb{M}} : I_{\mathbb{Z}} \circ \mathbb{M}_{\text{target}} \Rightarrow \mathbb{M}_{\text{proxy}} \circ I_{\mathbb{X}}$$

$$I_{\mathbb{Z}} \coloneqq (\Delta_{\text{proxy}} \circ \_ \circ \Sigma_{\text{target}})[I_{\mathbb{Y}}] : \quad \mathcal{Z}_{\text{target}} \to \mathcal{Z}_{\text{proxy}}$$

Gist. Think of $\mathbb{I}$ as the translator between worlds. It moves inputs ($\mathcal{X}$), outcomes ($\mathcal{Y}$), and predictions ($\mathcal{Z}$) from target to proxy through transport functors, and its comparison 2-cells (φ) measure how much "translate then run" path differs from "run then translate" path.

[29] We use $\cdot$ for vertical composition of natural transformation, and $\circ$ for functor composition/ whiskering

The analysis and synthesis functors help us connect $I_{\mathbb{Y}}$ with $I_{\mathbb{Z}}$



Leading to comparison 2-cell for analysis and synthesis:

$$\varphi_{\Delta} : I_{\mathbb{Z}} \circ \Delta_{\text{target}} \Rightarrow \Delta_{\text{proxy}} \circ I_{\mathbb{Y}} \qquad \varphi_{\Sigma} : I_{\mathbb{Y}} \circ \Sigma_{\text{target}} \Rightarrow \Sigma_{\text{proxy}} \circ I_{\mathbb{Z}} \tag{5.7}$$

Mechanism interpretation. The mechanism is a special component of the interpretation. It encodes the "identifiable" change in "causal-structure." It is not fully determined even when all other components are. We will define it like a bundle:

$$I_M \coloneqq (I_{\mathbb{X}}, \varphi_{\mathbb{M}}) \tag{5.8}$$

We will talk more about this in Section 6 (Mechanistic Faithfulness).

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

$$\mathbb{D}_{\text{target}} \coloneqq \text{diag}\big(\mathbb{E}_{\text{target}}\big) \tag{5.9}$$

$$\mathbb{D}_{\text{proxy}} \coloneqq \text{diag}\big(\mathbb{E}_{\text{proxy}}\big) \tag{5.10}$$

Gist. Two worlds, same pipeline. An interpretation tells us how to align each shape across worlds, allowing us to compare them.

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

$$\varphi_{\bullet} : \mathbb{I} \circ \mathbb{D}_{\text{target}} \Rightarrow \mathbb{D}_{\text{proxy}} \tag{5.11}$$

All the diagrams still need to be coherent. You might have noticed we have not defined $\varphi_{\mathbb{Z}}$. It can be derived canonically from other terms, but it is a bit verbose. If we consider Equation (5.2), to simplify, we get two forms that produce $\varphi_{\mathbb{Z}} : I_{\mathbb{Z}} \circ \mathbb{Z}_{\text{target}} \Rightarrow \mathbb{Z}_{\text{proxy}}$. Both of them have to agree[30]:

$$\varphi_{\mathbb{Z}} \quad = \quad \Delta_{\text{proxy}} \circ \varphi_{\mathbb{Y}} \quad = \quad \big(\text{id}_{\mathbb{M}_{\text{proxy}}} * \varphi_{\mathbb{X}}\big) \cdot \big(\varphi_{\mathbb{M}} * \text{id}_{\mathbb{X}_{\text{target}}}\big) \tag{5.12}$$

[30] We use $*$ for horizontal composition of natural transformations

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

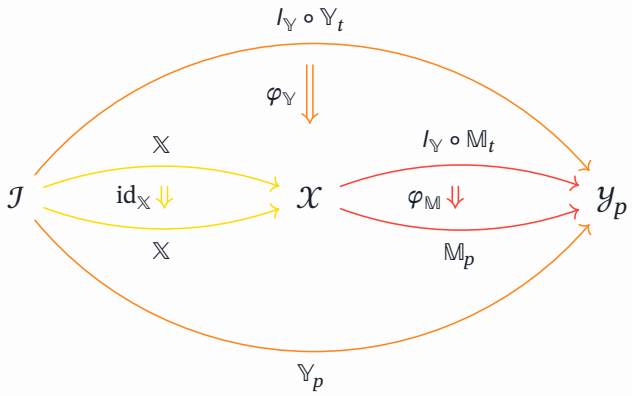
$$\mathbb{X}_{\text{target}} \equiv \mathbb{X}_{\text{proxy}} \tag{5.13}$$

and the *Ideal Evaluation Assumption*. We also then have

$$I_{\mathbb{X}} = \text{id}_{\mathcal{X}} \qquad \varphi_{\mathbb{X}} = \text{id}_{\mathbb{X}} \tag{5.14}$$

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

$$\varphi_{\mathbb{Y}} = \varphi_{\mathbb{M}} * \text{id}_{\mathbb{X}} \quad : \quad I_{\mathbb{Y}} \circ \mathbb{Y}_{\text{target}} \Rightarrow \mathbb{Y}_{\text{proxy}} \tag{5.15}$$



Let's name this equation 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 identity, we would have *perfect observational transport*:

$$\mathbb{I} \circ \mathbb{D}_{\text{target}} \cong \mathbb{D}_{\text{proxy}} \tag{5.16}$$

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 helpful starting point is to remember that we centered this discussion with a goal; there is something specific we want to understand: *Why does Claude tell me not to text ... ?!?!* 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[31]. We could formalize this with respect to R_U, a chosen understandability reference[32]:

$$\mathbb{E}_{\text{prod}} \text{ is inherently interpretable} \Leftrightarrow \forall i, \mathsf{R}_U \ll \mathsf{Und}_i\left(\mathbb{E}_{\text{prod}}\right) \tag{6.2}$$

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**[33] explanation (analytical, precise, formal) for how to deadlift.

Things started to click inside me. My body started telling me new things, 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**[34] down, a fuzzy internal explanation for deadlifting I directly operate on.

The process of understanding [10] itself is a map [12]: $\mathbb{E}_{\text{verbatim}} \to \mathbb{E}_{\text{gist}}$

[31] Again, is understanding distributed evenly [65]? Are we imposing a code of legibility [66]?

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

[33] This terminology is from [10], same with "gist"

[34] A gist is the internalization of an explanation. The concretization of a gist is a mechanism (an algorithm?). What is the concretization of *a vibe*?

The risks exist and persist. Thinking I understand something is different from actually understanding it. A $\mathbb{E}_{\text{verbatim}}$ can feel very plausible [67], so I become overconfident that my $\mathbb{E}_{\text{gist}}$ is faithful[35]. 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 find my way again. **In one way or another, we end up leveraging a post-hoc explanation.** The question of faithfulness does not go away.

Broadly, if explanations only differ in mechanisms, we have two types of structures that determine faithfulness:
- Causal-Structural
- Mechanistic-Structural

In the remainder of this section, we will explore a few ways an explanation can be faithful. We'll first examine how faithfulness relates to evaluations and observations themselves (locality and lossiness); then we'll see what it takes to have causal-structural faithfulness (perturbational, observational, interventional, counterfactual); and finally, we explore what *could* be mechanistic-structural faithfulness.

Working assumption moving forward We will assume Equation (5.3). For the rest of the section, we identify $\mathcal{Y}$ with $\mathcal{Z}$, writing $\mathbb{Y} \equiv \mathbb{Z}$ and **reusing symbols accordingly**. We also continue to assume Identical Evaluation in Equation (5.13), so the Basic Approximation in Equation (5.15) holds.

Flag. From here on, we silently identify $\mathcal{Y}$ with $\mathcal{Z}$ and reuse symbols across that bridge.

Locality

Gist. Local explanations zoom into a subdomain. We judge them by how well they fit there—and by how cleanly locals glue into a global story.

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 formalized by inclusion[36] map like $\mathbb{X}_{\text{local}} \hookrightarrow \mathbb{X}_{\text{target}}$. We can think of having a functor that restricts our index category:

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

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

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

where $(\mathbb{M})_i$ is the submechanism in action

[35] This is an actual risk with me, with this piece. I align with [68] that math is a profoundly human, bodily practice, which uses formalisms to refine and expand intuition. If not ultimately correct, I hope my failures [69] do guide you to better intuitions.

[36] $\hookrightarrow$ means inclusion

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. By our *working assumptions*, we see our coherence is also restricted to the inclusion:

$$\varphi_{\mathbb{Y}} * i = \varphi_{\mathbb{M}} * \mathrm{id}_{\mathbb{X}} * i \tag{6.5}$$

For a second, let me do a sheaf-theoretic [70] 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 with 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

Gist. If your observation map collapses distinctions, no downstream comparison can recover them. Mechanistic claims are then "modulo the collapse."

Sometimes, we are less concerned about the complete 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 to control its behavior concerning an attribute [71].

We have lossy observations when our proxy explanation has an observation transformation $I_{\mathbb{Y}}$ that collapses differences too much, allowing only for *coarse discriminations* between observations. Then, even if our proxy had exact observations ($\varphi_{\mathbb{Y}} = \mathrm{id}_{\mathbb{Y}}$), our interpretation could never recover enough of the target explanation on its own. By our *working assumptions*, 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 causal-structural faithful modulo $\mathbb{I}_{\mathbb{Z}}^{\text{lossy}}$. We can express this as[37]:

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

As I mentioned before, understandability is goal-oriented. Representation Engineering has shown great promise as a framework for Control [14]. **What's the big picture?** We need to be aware of our blind spots, and so we can address current challenges [71], like deterioration of capabilities.

[37] Yes, I am definitely abusing notation here

Observational Faithfulness

Gist. Your residual is a choice. Pick a view of outputs that actually sees what you care about; otherwise you'll certify the wrong sameness.

By our *working assumptions*, we can calculate the **observational residual** as:

$$r_o = \| \varphi_{\mathbb{Y}} \| \tag{6.7}$$

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

$$r_o < \varepsilon \tag{6.8}$$

Our r_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 $(\varphi_{\mathbb{Y}})_i = \big(p_t(-|x), p_p(-|x)\big)_i$, and the functional is the mean of KL divergence, then $\varphi_{\mathbb{Y}} = E_{x \in \mathrm{Seqs}}\big[D_{\mathrm{KL}}\big(p_t(-|x), p_p(-|x)\big)\big]$. But we can also choose other $\mathcal{Y}$ categories with different functionals.

In Equation (6.1), I chose $\mathfrak{L}_{\mathrm{syn}}$ to be the observation category to highlight other possibilities for r_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 [72]? We could then have $\varphi_{\mathbb{Y}} = D_{\mathrm{spectral}}\big(G\big(\mathbb{Q}_\varepsilon L^t_{\mathrm{syn}}\big), G\big(\mathbb{Q}_\varepsilon L^p_{\mathrm{syn}}\big)\big)$.

Some choices for $\|\_\|$ will have the equivalent effect of doing a $I^{\mathrm{lossy}}_{\mathbb{Y}}$ transformation[38]. 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.

Perturbational Robustness

Gist. Stable stories are easier to trust. Find regions where small input noise doesn't blow up your comparisons—and focus there first.

We are interested in interpretations that are stable against small perturbations. Our interpretations would be challenging to understand if small amounts of noise in the inputs created large deviations.

[38] Consider that instead of D_{KL}, you use $D_{\max} = \big|\max p_t(-|x) - \max p_p(-|x)\big|$. As long as the peak probabilities do not change, any change to the distributions will be invisible to you.

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 :

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

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 focus on those.

Interventional Faithfulness

Gist. We don't just match outputs—we edit both worlds and see if their edited behaviors still match. Good studies include big edits, small edits, and compositions of edits.

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 the intervened proxy to exhibit the same behavior after making edits to both.

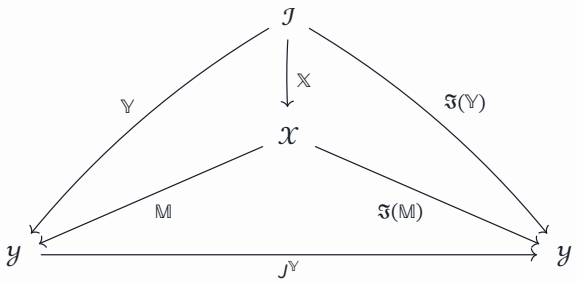


Figure 8: For interventions, we impose that this diagram must commute on the nose.

An **intervention** will consist of endofunctors[39] $(J^{\mathbb{X}}, J^{\mathbb{Z}}, J^{\mathbb{Y}})$ that relate an unedited explanation to an edited one: $\mathbb{J} : \mathbb{E} \to \mathfrak{J}(\mathbb{E})$.

By our *working assumptions*, the only non-trivial endofunctor will then be:
- A distributional endofunctor $J^{\mathbb{Y}} : \mathcal{Y} \to \mathcal{Y}$

Such that the other transformations are $J^{\mathbb{X}} = \mathrm{id}_{\mathbb{X}}$ and $J^{\mathbb{Z}} = J^{\mathbb{Y}}$

[39] Because intervening should not change the data type

These transformations should **not** be lax because when we make interventions, we are making explicit, intentional, and structurally-informed edits :

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

The **edited explanation** for an intervention written like:

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

We perform an intervention for each explanation in the non-edited interpretation. Given an **original interpretation**, $\mathbb{I} : \mathbb{E}_t \to \mathbb{E}_p$, we then have:

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

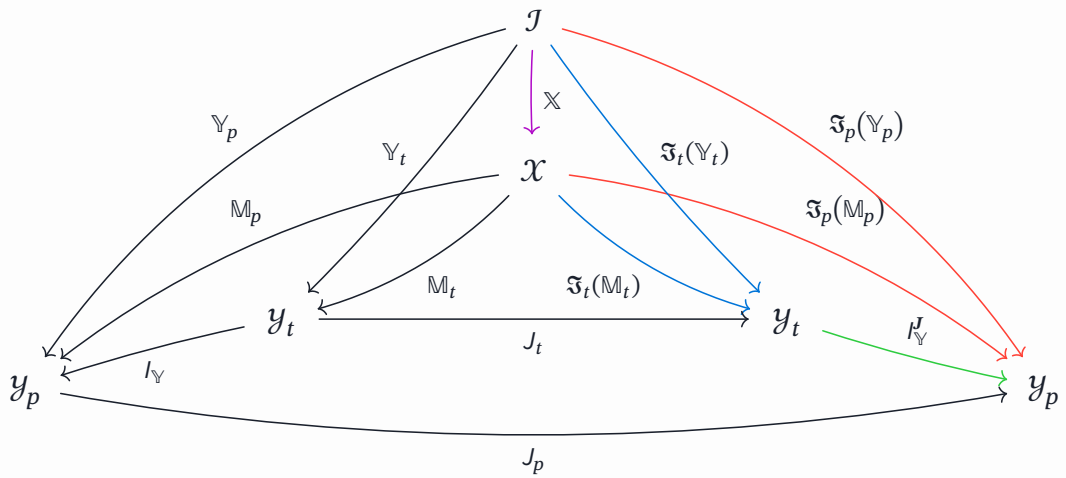
$$\mathbb{J}_p : \mathbb{E}_p \to \mathfrak{F}_p(\mathbb{E}_p) \tag{6.13}$$

which can be grouped as an **interventional pairing**:

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

How does our original interpretation hold up after the interventions? The relation between worlds (from the target to the proxy) could look very different after the edits. The interpretation induced by the interventions, which might not be aligned with the original interpretation, is called the **intervened interpretation**:

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

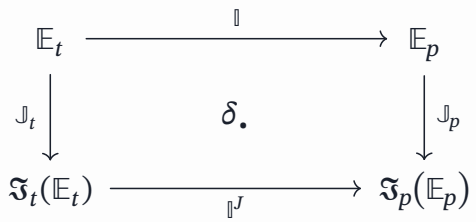


If we intervene and then interpret, would we get something very different from interpreting first and intervening second? If the interventions we performed really are the "same" edit, just looked at in "different" worlds, the original interpretation should align with the intervened intervention. So if my original interpretation was excellent, my intervened interpretation should be just as good. However, we may not be able to make edits that are the "same" because each explanation changes very differently.

Extremely loosely, imagine explanations are points in the x-y plane. Each world is its own vector field on the plane that dictates how each world's explanations change. A target explanation could have the same coordinate as a proxy explanation ("perfect" original interpretation, no distance between explanations). But because they belong to different vector fields, they can only "move" in different curves at different speeds. We need to make a comparison: If I move forward by one second along the target curve and then switch to the nearest point on the proxy curve, would that put me in a very different place compared to first jumping to the proxy curve and then moving one second along it?

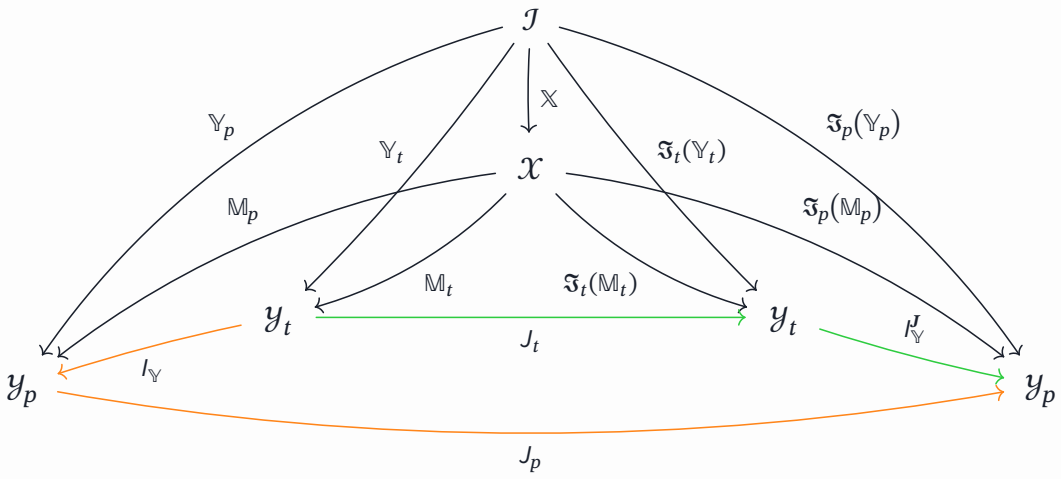
The **defect transform** captures that path comparison:

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



What the shape says. We wished the order of operations ("edit" and "interpret") did not determine where we land. The gap between path trajectories is $\delta_\bullet$

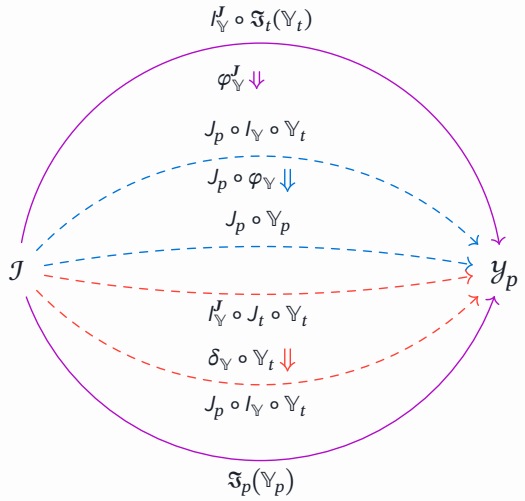
We can zoom in, (and simplify with our *working assumptions*) to see more closely which paths we are comparing:



By our *working assumptions*, the **defect component** will then only be:

$$\delta_{\mathbb{Y}} : I^{\boldsymbol{J}}_{\mathbb{Y}} \circ J_t \to J_p \circ I_{\mathbb{Y}} \tag{6.17}$$

How well our intervened interpretation aligns ($\varphi^{J}_{\bullet}$) will depend on both the approximation in the original interpretation ($\varphi_{\bullet}$), and the defect from edits themselves ($\delta_{\bullet}$). For the observation components, this would look like:



We can write this as:

$$\varphi^{J}_{\bullet} = \left(\mathbb{J}_p \circ \varphi_{\bullet}\right) \cdot \left(\delta_{\bullet} \circ \mathbb{D}_t\right) \tag{6.18}$$

We want to isolate the causal mismatch from the defect.
We can then calculate the **defect residual** as:

$$r_{\delta} = \| \varphi^{J} - \left(\mathbb{J}_p \circ \varphi_{\bullet}\right) \cdot \left(\mathrm{id}_{\bullet} \circ \mathbb{D}_t\right) \| \tag{6.19}$$

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* [73]:

$$\mathbb{I}^{J} \circ \mathbb{J}_t = \mathbb{J}_p \circ \mathbb{I} \tag{6.20}$$

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

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

$$\mathbb{A} = (\mathbb{I}, \boldsymbol{J}, \ldots) \tag{6.21}$$

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

$$\mathbb{S} = \{\mathbb{A}_{\alpha}, \mathbb{A}_{\beta}, \mathbb{A}_{\gamma}, \ldots\} \tag{6.22}$$

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

$$r_{\mathcal{S}}^{\beta\circ\alpha} = \| \varphi^{\boldsymbol{J}^{\beta}\circ\boldsymbol{J}^{\alpha}} - \left(\mathbb{J}_p^{\beta} \circ \mathbb{J}_p^{\alpha} \circ \varphi_{\bullet}\right) \cdot \left(\mathrm{id}_{y_t} \bullet \circ \mathbb{D}_t\right)\| \tag{6.23}$$

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

$$T(\boldsymbol{R}_{\mathbb{S}}) < \varepsilon \tag{6.24}$$

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 large *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 [74] part of a whole network). **Should any intervention on a non-transportable part of the target mechanism be invisible to the proxy?**.
- If you are working in a more concrete environment, leverage their constructions. For instance, Equation (6.24) has an analogue in Causal Abstraction [50] 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

Gist. Now we compare edits themselves. If "what changed" in the proxy mirrors "what changed" in the target (at the same input), counterfactuals line up.

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** [75] would be:

$$\diamondsuit(\boldsymbol{J}) : \mathbb{J}_p \circ \mathbb{I} \Rightarrow \mathbb{I}^{\boldsymbol{J}} \circ \mathbb{J}_t \tag{6.25}$$

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

$$\mathcal{M}_i : i \to \mathcal{Y} \quad \mathcal{M}_i(\mathbb{E}_k) \coloneqq \mathbb{Y}_k(i) \tag{6.26}$$

Using Equation (5.6), 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**:

$$\chi^{(\boldsymbol{J},i)} \coloneqq (\mathcal{M}_i \circ \_ \circ \mathrm{id})[\Diamond(\boldsymbol{J})] \tag{6.27}$$

How to look. It's the same case i in two worlds. The comparator checks that the "difference of differences" matches. This reveals, considering concrete edits, how two worlds align.

$$\begin{array}{ccc} \mathbb{Y}_t(i) & \xrightarrow{\quad I_{\mathbb{Y}} \quad} & \mathbb{Y}_p(i) \\ {\scriptstyle \mathbb{J}_t}\downarrow & \chi^{(J,i)} & \downarrow{\scriptstyle \mathbb{J}_p} \\ \mathfrak{F}_t(\mathbb{Y}_t)(i) & \xrightarrow[\quad I^{\boldsymbol{J}}_{\mathbb{Y}} \quad]{} & \mathfrak{F}_p(\mathbb{Y}_p)(i) \end{array}$$

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

$$\chi^{(\boldsymbol{J},i)}(\mathbb{E}_t) = \mathcal{M}_i \circ \mathbb{J}_p \circ \mathbb{I} \circ \mathbb{E}_t \to \mathcal{M}_i \circ \mathbb{I}^{\boldsymbol{J}} \circ \mathbb{J}_t \circ \mathbb{E}_t \tag{6.28}$$

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

$$\rho_e(\boldsymbol{J}, i) : \big(\mathfrak{F}_p(\mathbb{Y}_p)\big)(i) \to \big(I^{\boldsymbol{J}}_{\mathbb{Y}} \circ \mathfrak{F}_t(\mathbb{Y}_t)\big)(i) \tag{6.29}$$

We can define the base observations:

$$y_t \coloneqq \mathbb{Y}_t(i) \qquad y_p \coloneqq \mathbb{Y}_p(i) \tag{6.30}$$

We can define the post-edit observations:

$$y_t^{\boldsymbol{J}} \coloneqq \mathfrak{F}_t(\mathbb{Y}_t)(i) \qquad y_p^{\boldsymbol{J}} \coloneqq \mathfrak{F}_p(\mathbb{Y}_p)(i) \tag{6.31}$$

We can define the cross observations:

$$y_p^{\times} \coloneqq I_{\mathbb{Y}}(y_t) \qquad y_p^{\boldsymbol{J}\times} \coloneqq I^{\boldsymbol{J}}_{\mathbb{Y}}(y_t^{\boldsymbol{J}}) \tag{6.32}$$

We define the effect calculator as:

$$\Xi : \mathbb{Y} \times \mathbb{Y} \to \aleph \tag{6.33}$$

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

$$\triangle_{\aleph}(y, y') \coloneqq \Xi(y, y') \tag{6.34}$$

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

$$r_{\mathrm{cf}} = \| \triangle_{\aleph}(y_p, y_p^{\boldsymbol{J}}) - \triangle_{\aleph}(y_p^{\times}, y_p^{\boldsymbol{J}\times}) \| \tag{6.35}$$

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

$$\begin{aligned} \forall J, i, \quad & \chi^{(J,i)}(\mathbb{E}_t) = \mathrm{id}, \\ \forall J, i, \quad & \mathcal{M}_i \circ \mathbb{J}_p \circ \mathbb{I} \circ \mathbb{E}_t = \mathcal{M}_i \circ \mathbb{I}^J \circ \mathbb{J}_t \circ \mathbb{E}_t \end{aligned} \tag{6.36}$$

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

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

$$T(\{r_{\mathrm{cf}}(\boldsymbol{J}) : \boldsymbol{J} \in \mathbb{S}\}) < \varepsilon \tag{6.37}$$

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

Gist. Below the identifiability limit, causal sameness doesn't pin down mechanisms. We still ask for mechanism-level coherence where inputs actually light it up.

We say an interpretation is **ε-causally faithful** if it is ε-observationally, ε-interventionally, and ε-counterfactually faithful. We have *perfect causal transport* ($\varepsilon = 0$) if we have perfect observational, interventional, and counterfactual transport. In such a scenario, the causal structure is invariant to $I_{\mathbb{M}}$, so $\mathbb{E}_t$ and $\mathbb{E}_p$ have the same *distributional layer* [76]. If two mechanisms have the same causal structure, they will not be causally identifiable from each other [77]. What can we say about mechanisms that have *causal structure invariance*? Do we need to *really* look inside $\mathbb{M}$? Yes! In practice, we are nowhere near establishing perfect causal transport. We need to go below and find better guarantees for our work.

Are mechanisms constrained in any way?

Each level of causal-structural faithfulness adds constraints to the mechanism. Using our *working assumptions*, we can point out a few we have seen so far:

$$\varphi_{\mathbb{Y}} = \varphi_{\mathbb{M}} * \mathrm{id}_{\mathbb{X}} \qquad \mathfrak{F}(\mathbb{M}) \circ \mathbb{X} = \mathfrak{F}(\mathbb{Y})$$

However, these constraints do not give enough specification for the mechanism to be unique. In our framework so far, a mechanism is only observable through its evaluations.

Mechanisms as categories

If mechanisms can be captured by category-theoretic formalism, mechanistic faithfulness would reduce to exploring a relevant "sameness" in category theory, which gives us many (maybe infinite[40]) levels of nuance to consider.

Let's go into a side quest (in the appendix) to work through some intuitions.
I'll wait for you to read.

As we saw, category theory can give us the framework to define mechanisms. I have been using very simple constructs, but new frameworks (like [36], whose authors reason about obstructions to composionality) keep coming out.

However, we need something outside of category theory to inform us what is *mechanistically meaningful*. We have too many category-theoretic available choices otherwise.

A prevalent common model for mechanisms is Structural Causal Models(SCM), which induce Directed Acyclic Graphs (DAG). In [76], *F. Zennaro* uses category-theoretic language to explain the consequences of choices by different formalisms on SCMs and DAGs, like [79]. They even remark that there is *"wide degree of freedom in defining what class of transformations should be considered an abstraction"*.

I have recently been seeing interesting developments in measure-theoretic causal frameworks. It would be interesting to see what category-theoretical take we could have on *Factored Space Models* [80] and *Causal Spaces* [81]

Perhaps the **opacity** in mechanisms is not so different from the opacity within each of us. It seems that if we want to develop a better theory of mechanisms, we need to get a little bit messy, like our own experience can be. This piece has taken a top-down approach so far. In future work, I will have to explore the middle-outs, and views from the bottom [82].

What's the big picture? Mechanistic faithfulness is difficult and 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.

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 (5.13) to simplify analysis.

[40] Higher category theory [78]

We will have more explanations than models

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[41], $\mathcal{X}$ is the set of all possible input prompts based the model's token alphabet, $\mathcal{Y}$ is the set of next-token distributions.

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

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 also different evaluation and observation domains. We have[42] 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 local 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 strict mechanistic analyses (what do the concrete transformations between mechanisms $\mathbb{I}_{\mathbb{M}}$ tell us about what behavior can be preserved) and causal-structural analyses (do mechanisms behave similarly to changes?).

Beware of implicit interpretations and interventions!

Gist. Every metric is a lens—and a transformation. Name your lens. If it's lossy, your claims are only up to that loss.

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 with very coarse partitions (where A_{token} is token alphabet):

$$g : \mathcal{Y} \to A_{\text{token}} \quad \forall y_j \in \mathcal{Y} : g(y_j) = \text{argmax}_{z \in A_{\text{token}}}\, p(z|x_j) \tag{7.4}$$

[41] 18-layer LLM or Claude Haiku 3.5

[42] Look at Section 6 (Locality)

For instance, when we view the target model from top-1 accuracy view, we get[43]:

$$\mathbb{I}^{\text{top-1}} : \mathbb{E}_t : (\mathbb{M}_t, \mathcal{X}, \mathcal{Y}) \to \mathbb{E}_t^{\text{top-1}} : (\mathbb{M}_t \,/\ker(g), \mathcal{X}, \mathcal{Y}/\ker(g)) \tag{7.5}$$

This means that if we only use top-1 accuracy to determine faithfulness, we can only identify any mechanism up to equivalence by $\ker(g)$.

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

$$\mathbb{I}^{\mathcal{F}} : \mathbb{E}_t : (\mathbb{M}_t, \mathcal{X}, \mathcal{Y}) \to \mathbb{E}_t^{\mathcal{F}} : \left(\mathbb{M}_t^f \hookrightarrow \mathbb{M}_t, \mathcal{X}, \mathcal{F}\right) \tag{7.6}$$

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 $\mathcal{F} \to \mathcal{Y}$.

We can also have implicit interventions. For instance, when we freeze the target model's attention based on run i, we can see that as an interventional [50]:

$$\mathbb{J}^{\kappa} : (\mathbb{M}_t, \mathcal{X}, \mathcal{Y}) \to (\mathfrak{T}^{\kappa}(\mathbb{M}_t), \mathcal{X}, \mathcal{Y}) \tag{7.7}$$

As we have seen before, both interpretations and interventions are transformations between explanations. Whether a given implicit transformation is one or the other is a matter of context and perspective[44].

What mechanistic structure could prevail?

Mechanistic faithfulness almost always requires further specification 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:

$$\mathbb{I}_{\mathbb{M}}^{t\to\,\text{glob}} : \mathbb{M}_t \to \mathbb{M}_{\text{glob}} \tag{7.8}$$

We immediately see that $\mathbb{I}_{\mathbb{M}}^{t\to\,\text{glob}}$ cannot have high mechanistic-structural faithfulness if we consider network connectivity as part of the mechanism: $\mathbb{M}_{\text{glob}}$ has a computational graph that is not graph-isomorphic to $\mathbb{M}_t$. 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.

[43] Look at Section 6 (Lossiness)

[44] As a rule of thumb, if the transformation preserves relevant isomorphisms and we are intimately acquainted to how it mechanistically happens, so everything exactly commutes like Equation (6.10), then it could be an intervention

We then consider how the local replacement model was built in three steps. First, we do:

$$\mathbb{I}_{\mathbb{M}}^{\text{glob} \times t(x) \to \text{edited-glob}} : \mathbb{M}_{\text{glob}} \times \mathbb{M}_t \circ x_{\text{local}} \to \mathbb{M}_{\text{edited-glob}} \tag{7.9}$$

This means we modified $\mathbb{M}_{\text{glob}}$ with *values* from $\mathbb{M}_t$ evaluated at a specific prompt x_{local}: evaluated attention patterns and normalization denominators. Viewed from the perspective of $\mathbb{M}_{\text{glob}}$, can be even conceived as a *hard intervention* [50], changing a sub-mechanism with constant function:

$$\mathbb{I}_{\mathbb{M}}^{\text{glob} \to \text{edited-glob}} : \mathbb{M}_{\text{glob}} \overset{\times \mathbb{M}_t \circ x_{\text{local}}}{\to} \mathbb{M}_{\text{edited-glob}} \tag{7.10}$$

Viewed from the perspective of $\mathbb{M}_{\text{t}}$, we get the same issues as Equation (7.8). The second step is:

$$\mathbb{I}_{\mathbb{M}}^{\text{edited-glob} \to \text{Local}} : \mathbb{M}_{\text{edited-glob}} \to \mathbb{M}_{\text{Local}} \tag{7.11}$$

In this step, we add an error adjustment to each CLT output, to make it exact-on-prompt x_{local}. This could be considered as a type of *interventional* [50], changing the mechanism based on the previous run. The third step is the explicit restriction to the single output of interest:

$$\mathbb{I}_{\mathbb{M}}^{(i)} : \mathbb{M}_{\text{Local}} \to m_{\text{local}} \tag{7.12}$$

The local attribution graph is built from the local replacement model:

$$\mathbb{I}_{\mathbb{M}}^{\text{local} \to \text{att}} : m_{\text{local}} \to m_{\text{att}} \tag{7.13}$$

The local attribution graph has edges calculated from weights. We also form super-nodes, which could be considered a form of value merge [50]. Both present no obvious issues for mechanistic faithfulness. We also do have pruning (which is not variable marginalization), but I would expect the influence of it not to be so significant.

The global-weights model is built from the global replacement model:

$$\mathbb{I}_{\mathbb{M}}^{\text{glob} \to \text{w}} : \mathbb{M}_{\text{glob}} \to \mathbb{M}_{\text{w}} \tag{7.14}$$

This transformation does not consider attention-mediated circuits and has interference, as it does not readily account for how features actually co-activate. This does represent a barrier for mechanistic faithfulness.

Finally, interpretations can be composed:

$$\mathbb{I}_{\mathbb{M}}^{t \to \text{att}} = \mathbb{I}_{\mathbb{M}}^{\text{local} \to \text{att}} \circ \mathbb{I}_{\mathbb{M}}^{(i)} \circ \mathbb{I}_{\mathbb{M}}^{\text{glob} \to \text{Local}} \circ \mathbb{I}_{\mathbb{M}}^{t \to \text{glob}} \tag{7.15}$$

What's the big picture? The biggest challenge for mechanistic faithfulness is the fact that CLTs create a differently shaped computational graph. The jumps from the local replacement model to the local attribution graph are less problematic.

How do we think of experiments?

Gist. Draw the study as a diagram first. If the routes you claim to compare aren't actually the routes you measured, your numbers won't mean what you think.

We will look at some of the paper's experiments and make some remarks using our new language. In each study, $\mathbb{I}^{\text{study}}$ will be our interpretation of interest.

How good is our proxy at global level?

Accuracy evaluation of the global replacement model uses top-1 accuracy, which has coarse partitions as in Equation (7.5). However, in the paper, the authors show that top-1 accuracy tracks the KL Divergence (which is more discriminative), maybe hinting that in practice, top-1 accuracy would identify as much structure as the KL Divergence. The authors also relate the accuracy to the reconstruction error of CLTs.

$$\begin{array}{ccc} \mathbb{E}_{\text{t}} = (\mathbb{M}_t, \mathbb{X}, \mathbb{Y}) & \xrightarrow{\quad\mathbb{I}^{\text{study}}\quad} & \mathbb{E}_{\text{glob}} = \left(\mathbb{M}_{\text{glob}}, \mathbb{X}, \mathbb{Y}\right) \\ & g \Big\Downarrow & \\ (\mathbb{M}_t \,/\ker(g), \mathcal{X}, \mathcal{Y}/\ker(g)) & \longrightarrow & \left(\mathbb{M}_{\text{glob}} \,/\ker(g), \mathcal{X}, \mathcal{Y}/\ker(g)\right) \end{array}$$

Figure 9: The more coarse a metric g is, the less identifiable our explanations will be.

Feature influence: Validating specific mechanisms

The paper does ablations to verify that when we steer the interpretable features, we get meaningful changes in behavior.

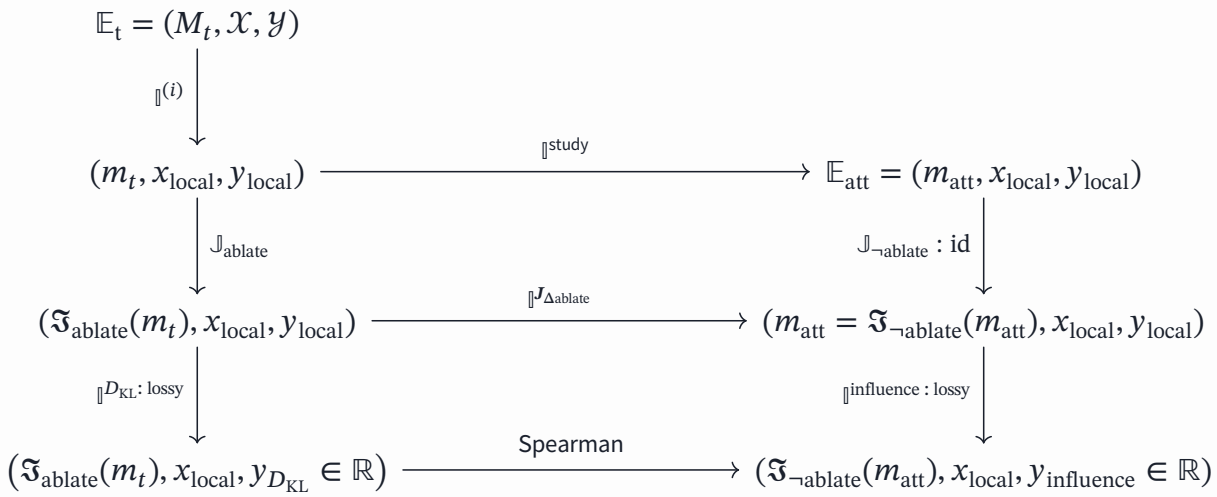


Figure 10: No estimate node-logit influence, the paper ablates the target world but not the proxy world. By this construction, $\mathbb{I}^{J_{\Delta\text{ablate}}}$ will encode the magnitude of its effect[45]. To approximate it, the authors study the world-to-world correlation from re-interpreted versions. The re-interpretations are technically lossy but useful to compute an actual number.

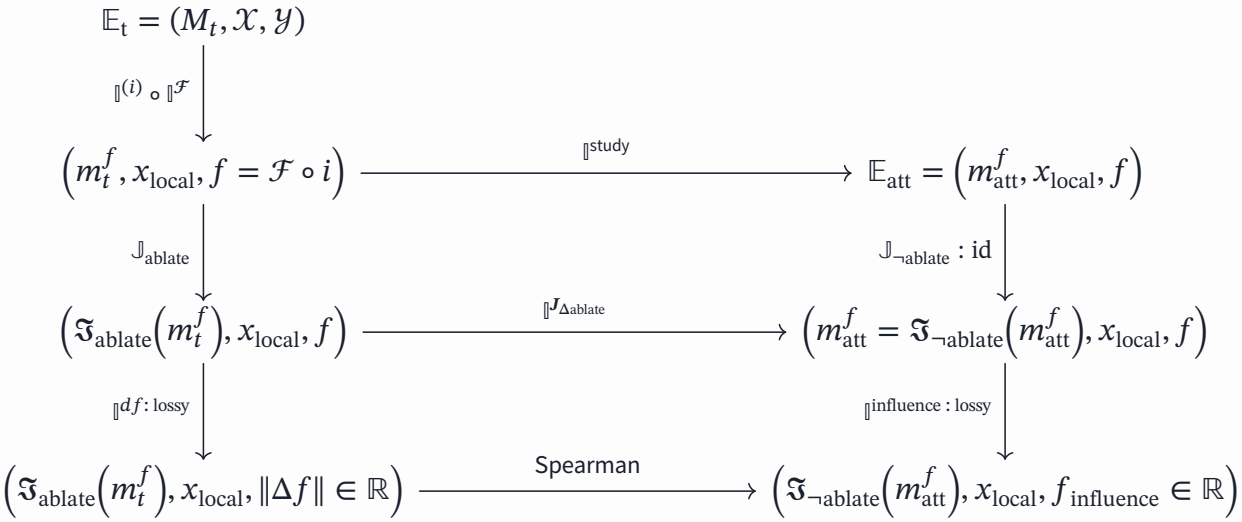


Figure 11: Feature-feature influence is estimated in analogous way to Figure 10

Perturbations: Testing faithfulness of the local replacement model as a whole

Their perturbation study can be thought of as an interventional study; the paper compared intermediate feature activations after intervening at an earlier layer, as Equation (7.6). The target model has frozen attention, like Equation (7.7), and we restrict our domain to a single prompt:

[45] If $\mathbb{I}^{\text{study}} : \text{id}$, then $\mathbb{I}^{J_{\Delta\text{ablate}}} \circ J_{\text{ablate}}$ would be a transformation $\mathbb{E}_t \to \mathbb{E}_t$

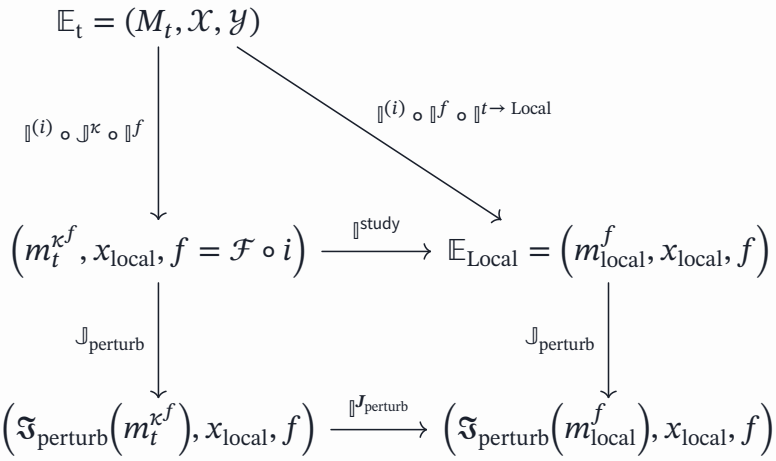


Figure 12: The diagram makes evident the implicit interpretations and interventions. We have $\mathbb{J}_t = \mathbb{J}_p = \mathbb{J}_{\text{perturb}}$ because intervention is the same operation, and we get exact agreement under zero-size perturbations. $\mathbb{I}^{J_{\text{perturb}}}$ is estimated using cosine similarity in paper

In Figure 12, we can see that our target has multiple transformations ($\mathbb{J}^{\kappa} \circ \mathbb{I}^{(i)} \circ \mathbb{I}^{f}$) before it enters the core study (rectangle).

What's the big picture? The diagrams can tell us a lot at just a glance. Instantly, we can get the *gist* of what's going on more clearly.

I really love Anthropic's work. I had read the paper a few times and felt I already had a good understanding. But after writing diagrams for it, I have internalized it all at a deeper level. Each experiment of the paper now has a "shape" in my mind.

Thinking categorically is more than a formal technique.
It is a way of feeling yourself forward with curiosity.

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; it also helps me investigate my queerness in a way. So, 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 enjoy reading.

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 [83]? Are they properties that activate particular mechanisms [84]? What structure is induced by their associated level of abstraction [85]? How do they all map [86] from and into R^k of the residual stream? They might not be proper manifolds [39], [87]? Are they some curvature [88]? What are they, *really*?

This piece presented some of my mental models around interpretability. As I mentioned earlier, I did not intend to give you a tidy, provably correct, data-driven cartographic map [89]. 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.

On the technical side, I want to conclude by highlighting how category theory helps us reason and formalize what our experiments are really saying, as in Section 7 (Interpreting Circuit Tracing). Sometimes, stopping and re-reading [90] of an old [91] experience with a new perspective [92] can be as illuminating [93] as a new experience. Sometimes, I want us to cruise [94] ahead instead.

You might find it funny that I discuss math, my feelings, and make a joke in the same breath. But that's how my mind does work, and how I do live [4]. I am not interested in upholding ecologies of knowledge [5] that alienate us. I want us to be *congressive*[46]. As I hinted, interpretability deals with an inherent [96] opacity. Not only because artificial intelligence is complex [97], [98], but also because we are [99], [100]. Interpretability is **not** [101] a solved problem between us [66], [102], *non-artificial* people.

I want to remind us what's at stake here [1], [103]: The world will likely radically [104], [105] change for everyone in the upcoming decades. Most of the people who will live through this change are currently disengaged, excluded, or overlooked. Some are already resentful[47]. Let's take this cosmotechnical [106] changes [45] seriously. Let's intervene [107] in more than language models.

Keep wayfinding.

[46] empathetic + coming together + helpful [95]

[47] I know a certain Sam S. already is

Citation Information

Use the following to cite this piece:

```
@article{sialer2025wanderings,
      author  = {Ian Rios-Sialer},
      title   = {Category-Theoretic Wanderings into
Interpretability},
      year    = {2025},
      url     = {https://unrulyabstractions.com/pdfs/wanderings.pdf}
    }
```

Acknowledgement

I want to thank my amazing friend Abdul Wasay for feedback and revision.

I also used ChatGPT (OpenAI) and Claude (Anthropic) to brainstorm ideas, spot-check intermediate math, and suggest wording; the authors reviewed all outputs and take responsibility for the final content.

References

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

[2] 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

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

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

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

[6] 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

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

[8] 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

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

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

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

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

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

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

[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] T. S. Kuhn, *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 2012. doi: 10.7208/chicago/9780226458144.001.0001.

[46] 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

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

[48] 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

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

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

[51] 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

[52] 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

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

[54] 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

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

[56] 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.

[57] 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

[58] 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.

[59] 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

[60] 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

[61] 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

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

[63] 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.

[64] 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

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

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

[67] 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

[68] 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

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

[70] 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

[71] 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

[72] 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.

[73] 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.

[74] 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

[75] 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

[76] F. M. Zennaro, "Abstraction between Structural Causal Models: A Review of Definitions and Properties." [Online]. Available: https://arxiv.org/abs/2207.08603

[77] D. Sutter, J. Minder, T. Hofmann, and T. Pimentel, "The Non-Linear Representation Dilemma: Is Causal Abstraction Enough for Mechanistic Interpretability?." [Online]. Available: https://arxiv.org/abs/2507.08802

[78] R. Haugseng, "Higher categories." [Online]. Available: https://arxiv.org/abs/2401.14311

[79] J. Otsuka and H. Saigo, "On the Equivalence of Causal Models: A Category-Theoretic Approach." [Online]. Available: http://arxiv.org/abs/2201.06981

[80] S. Garrabrant, M. G. Mayer, M. Wache, L. Lang, S. Eisenstat, and H. Dell, "Factored space models: Towards causality between levels of abstraction." [Online]. Available: https://arxiv.org/abs/2412.02579

[81] S. Buchholz, J. Park, and B. Schölkopf, "Products, abstractions and inclusions of causal spaces." [Online]. Available: https://arxiv.org/abs/2406.00388

[82] T. Nguyen, *A View from the Bottom: Asian American Masculinity and Sexual Representation*. in Perverse modernities. Duke University Press, 2014. [Online]. Available: https://books.google.com/books?id=oGmABAAAQBAJ

[83] 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

[84] 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

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

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

[87] 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

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

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

[90] G. Gopinath, *Unruly Visions: The Aesthetic Practices of Queer Diaspora*. in Perverse Modernities. Durham, NC: Duke University Press, 2018, p. 248. doi: 10.1215/9781478002161.

[91] L. Lowe, *The Intimacies of Four Continents*. Durham: Duke University Press, 2015. doi: 10.1215/9780822375647.

[92] I. Lakatos, *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press, 1976. [Online]. Available: https://books.google.com/books?id=1n6SFdXCOBQC

[93] R. L. Trosper, *Indigenous Economics: Sustaining Peoples and Their Lands*. Tucson, AZ: University of Arizona Press, 2022, p. ~272.

[94] J. E. Muñoz, *Cruising Utopia: The Then and There of Queer Futurity*. New York: New York University Press, 2019.

[95] E. Cheng, *x+y: A Mathematician's Manifesto for Rethinking Gender*. Profile, 2020. [Online]. Available: https://books.google.com/books?id=E8KLDwAAQBAJ

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

[97] D. Sumpter, *Four Ways of Thinking: Statistical, Interactive, Chaotic and Complex*. Allen Lane, 2023. [Online]. Available: https://books.google.com/books?id=t1k7zwEACAAJ

[98] M. T. Bennett, "Is Complexity an Illusion?," in *Artificial General Intelligence*, Springer Nature Switzerland, 2024, pp. 11–21. doi: 10.1007/978-3-031-65572-2_2.

[99] É. Glissant, *Poetics of Relation*. Ann Arbor: University of Michigan Press, 2010.

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

[101] T. S. Goetze, "Hermeneutical Dissent and the Species of Hermeneutical Injustice," *Hypatia*, vol. 33, no. 1, pp. 73–90, 2018, doi: 10.1111/hypa.12384.

[102] F. Fanon, *Black Skin, White Masks*. New York: Grove Press, 2008.

[103] D. McQuillan, *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*. Bristol: Bristol University Press, 2022, p. 190. doi: 10.56687/9781529213522.

[104] H. Kim *et al.*, "The Road to Artificial SuperIntelligence: A Comprehensive Survey of Superalignment." [Online]. Available: https://arxiv.org/abs/2412.16468

[105] M. Fahad *et al.*, "The Benefits and Risks of Artificial General Intelligence (AGI)," in *Artificial General Intelligence (AGI) Security: Smart Applications and Sustainable Technologies*, S. El Hajjami, K. Kaushik, and I. U. Khan, Eds., Singapore: Springer Nature Singapore, 2025, pp. 27–52. doi: 10.1007/978-981-97-3222-7_2.

[106] Y. Hui, *The Question Concerning Technology in China: An Essay in Cosmotechnics*. MIT Press, 2016. [Online]. Available: https://books.google.com/books?id=cFuPEAAAQBAJ

[107] D. H. Meadows, "Leverage Points: Places to Intervene in a System," Hartland, VT, Dec. 1999. Accessed: Aug. 24, 2025. [Online]. Available: https://donellameadows.org/archives/leverage-points-places-to-intervene-in-a-system/

Appendix

poetics

as we saw in category theory, we like to see objects in context
this was my context this summer, come close to understand me

conditioned

the picture and the song
they don't seem connected
is there possible hermeneutics
but they are indeed united
conditioned on the enigma
did see those things
did enjoy that beat
felt them together
adjacent histories
are generatively coupled

3am dark floor time

dissonance around the peak
not well behaved
we design for the asymptotic
but then you finally do meet him
and suddenly algebra just works
what does this discontinuity mean
between me and u
and what do we tell our kids

messy

i complicate your life in the best ways
you want to share many things with me too
you are infatuated with me, i'm special
you really like me and you are scared

but you say this to many
fiction
the lot forewarns me
i never really was
one of the lucky ones
and now i am
crying in the airport floor
i actually think you are great
which makes all of this harder
did i imagine the glimpses
do you pull away
do i imagine that too

i crave to be held, fully
and now i am slightly bored
i think i lost the plot, a bit
and i think i am ready to move on
is that not allowed?

translations

compromised and scrambled
your messages open up a canvass
for me to fantasize and theorize
and to bend
(i am bent already)
my will

your piss on my face
snorted the testosterone
tasted the signifiers
intoxicated then leashed
you like to have fun

i don’t want my feelings
to spill over
you would not hold them
unbounded, i do enjoy the hurt
what about the overwhelm

are you ready now?

side quest into mechanistic sameness

I do not know what mechanisms really are. I am hopeful they can be fully represented in category-theoretic language. We are interested in determining if two mechanisms are "the same". In this side quest, we will review a few ways in which two things can be "the same" using category theory. The goal of this section is to give you a rudimentary and loose sense of what category theory can offer us to investigate mechanistic faithfulness: infinitely many levels of nuance

I am unsure which category-theoretic construction to choose to represent a mechanism. I will make an arbitrary choice, see how far we go, and adjust as needed. Choosing mechanisms to be plain categories (such as $\mathcal{C}, \mathcal{D}$) seems like a sensible place to start. It aligns with my initial intuition that mechanisms have "parts" (objects, arrows), but we can also compare a whole mechanism to another one (functors). In this chosen setting, the question of mechanistic faithfulness would be: what "sameness" can a map $\mathcal{C} \to \mathcal{D}$ preserve?

$$\begin{array}{ccc} \mathcal{C} & & \mathcal{D} \\ & \to & \\ \bullet_a \quad ? \quad \bullet_b & & *_{x?} \quad ? \quad *_{y?} \end{array}$$

Base Case

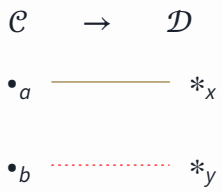
At the lowest level, we have no arrows in between objects[48]. We only have objects. Looking at $\mathcal{C}$: $\bullet_a \qquad \bullet_b$, we see that the only object "sameness" can be *equality*. Either $a = b$ and $\mathcal{C}$ only has one object; or $a \neq b$ and a, b have no relationship.

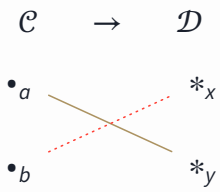
Similarly, the strongest category "sameness" would be equality: $\mathcal{C} = \mathcal{D} \Leftrightarrow \bullet_a = *_a, \bullet_b = *_b$

That would require us to have some extra information, being able to look at something *within* a, b to make that call. But if we were able to do so, we would have more levels below, and this is our lowest categoric level. Instead of identifying objects, let's identify structure. Each category is just a set of objects at this level, so the structure relates to *cardinality*. If $\mathbb{F} : \mathcal{C} \to \mathcal{D}$ is bijective on objects (a permutation), then $\mathcal{C}$ and $\mathcal{D}$ are *isomorphic*. There may be many such permutations.

[48] But each object will still have its identity.

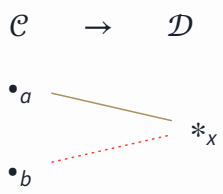
In our case, we have two possible isomorphisms:

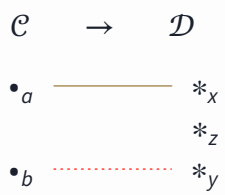
In both 

and 

, we lose no information:

$\mathcal{C} \cong \mathcal{D}$
In either case, there is a perfect pairing (bijection) between $\mathcal{C}$ and $\mathcal{D}$. In other, less perfect cases, we have

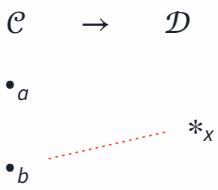
injectivity 

, or *surjectivity* 

<u>issues</u>, in

which we either lose information or we lose reach.

In all cases so far, all objects $a, b \in \mathcal{C}$ at least are guaranteed a destination in $\mathcal{D}$. But what if we start

dropping objects? 

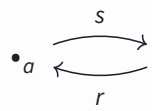
Then (from $\mathcal{C}$) cannot be a functor because

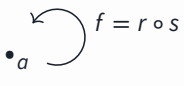
we lose a total object map.

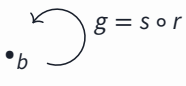
At this point, category theory would be a bit angry at you because, at this level, you would also lose *compositionality*. As we go up, within each level, we can keep making choices that have analogous consequences:
- loss of information
- loss of reach
- loss of functoriality
- loss of compositionality

However, as we progress to higher levels, we hope that our new set of available choices will give us nuance. Maybe we do want a system in which strict functoriality fails; can we systematize the failures? Where can we *relax* conditions so we do not have to throw away all structure?

One Level Up
At this level[49], we have morphisms between objects.

In 

, we can compose arrows like 

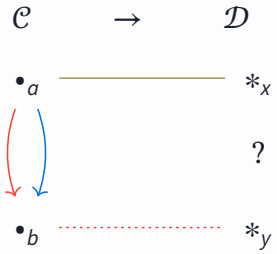
and 

[49] I will borrow some notation from [46]

For objects, arrows $r : a \to b$ and $s : b \to a$ witness "sameness".
If $s \circ r = \mathrm{id}_a$ and $r \circ s = \mathrm{id}_b$ (so $f = r \circ s$ and $g = s \circ r$), then $a \cong b$.
We can say they both belong to the same isomorphic class: $a, b \in [a]_{\cong}$

For morphisms, the notion of injectivity is called *faithfulness*, surjectivity is called *fullness*, so when two sets of arrows are isomorphic, we call them *fully faithful*. These properties need to be satisfied *locally*, with respect to the *source* and *target* of the morphism. Let's consider cases for a functor $\mathbb{F} \in [\mathcal{C}, \mathcal{D}]$.

First, consider[50]



, what are the options $[x, y]_{\mathcal{D}}$?

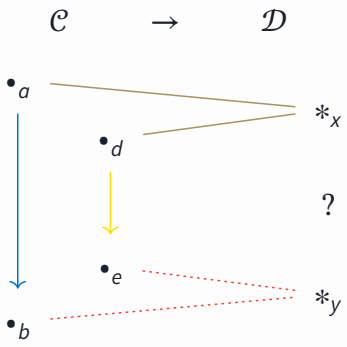
We could have two versions of the functor, each defining different $[x, y]_{\mathcal{D}}$:

1) 

, or 2) 

Both 1) and 2) are full but only 1) is fully faithful.
In 2), two morphisms in $\mathcal{C}$ map to same morphism in $\mathcal{D}$.

Now, consider



, what are the options $[x, y]_{\mathcal{D}}$?

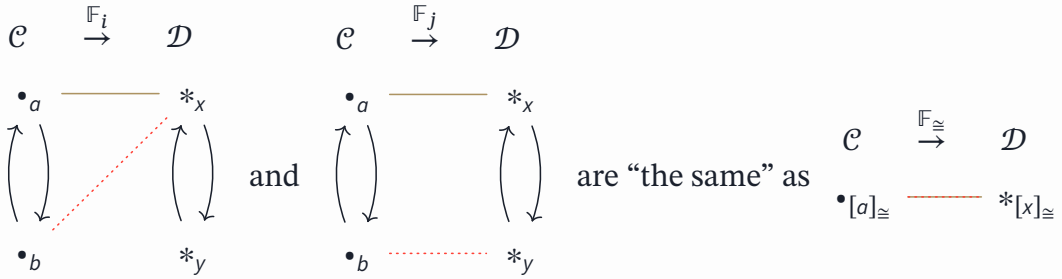
We could have i) 

, or ii) 

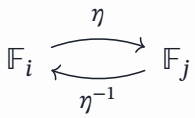
Neither i) nor ii) is full. Remember that surjectivity needs to be satisfied locally for fullness. Explicitly:
In i), $\mathrm{Size}([a, b]) = 1 \neq \mathrm{Size}([x = \mathbb{F}a, y = \mathbb{F}b]) = 2$
In ii), $\mathrm{Size}([a, e]) = 0 \neq \mathrm{Size}([x = \mathbb{F}a, y = \mathbb{F}e]) = 1$

[50] I borrowed these examples from fabulous [23]

What else can we say about $[\mathcal{C}, \mathcal{D}]$ functors? Functors will preserve isomorphisms[51]. We will say $\mathbb{F}_i, \mathbb{F}_j : \mathcal{C} \to \mathcal{D}$ are **objectwise isomorphic** if for every object $x \in \mathcal{C}$ there exists an isomorphism $\varphi_{\{x\}} : \mathbb{F}_{i(x)} \cong \mathbb{F}_{j(x)}$ in $\mathcal{D}$. (We are **not** requiring any coherence yet.) We only need this object-level agreement for the examples below; later we will point out when these $\varphi_{\{x\}}$ line up across arrows—i.e., a natural isomorphism.



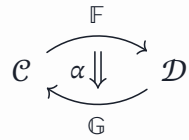
When the same choices $\varphi_{\{x\}}$ can be made compatibly with every arrow f in $\mathcal{C}$, they assemble into a **natural isomorphism** $\eta : \mathbb{F}_i \Rightarrow \mathbb{F}_j$:

But if $\mathbb{F}_i \cong \mathbb{F}_j$, then in some category: 

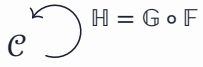
Mmmm.
In order to compare two $\mathbb{F}$ functors, it seems that we don't need all the details of the objects $a, b, x, \ldots$. Instead, we could focus on the *relationships* it has in its ambient category.

Rethinking what category to choose

I am now re-considering if I made the right choice in the beginning.
Right now, our objects need strict equalities, like $f = \mathrm{id}_a$, for $a \cong b$ to exist. What if we redefined our objects to be something more flexible? What if our base level (*0-morphisms* [78]) consisted of our $\mathcal{C}, \mathcal{D}$ themselves instead?

Now, consider a category like 

and forget the details of what is inside $\mathcal{C}, \mathcal{D}$

Analogous of what we saw before before,
if we wanted to have an isomorphism $\mathcal{C} \cong \mathcal{D}$, we would require
$\mathbb{H} = \mathrm{id}_{\mathcal{C}}$ in diagram 

We have an equality again, but this time, we can *relax* it.
Let's define natural transformations to account for the *failure* of $\mathbb{G} \circ \mathbb{F}$ and $\mathbb{F} \circ \mathbb{G}$ to be equal to identities:

$$\eta : \mathrm{id}_{\mathcal{C}} \Rightarrow \mathbb{G} \circ \mathbb{F} \qquad \varepsilon : \mathbb{F} \circ \mathbb{G} \Rightarrow \mathrm{id}_{\mathcal{D}}$$

$$(\varepsilon \mathbb{F}) \circ (\mathbb{F}\eta) = \mathrm{id}_{\mathbb{F}} \qquad (\mathbb{G}\varepsilon) \circ (\eta \mathbb{G}) = \mathrm{id}_{\mathbb{G}}$$

[51] If $\mathbb{F}$ is full & faithful, it also reflects them

Now, if that condition satisfies, we say $\mathcal{C}$ and $\mathcal{D}$
are equivalent $\mathcal{C} \simeq \mathcal{D}$
This is nice!
Equivalence $\simeq$ is weaker than isomorphism $\cong$, which is weaker than equality $=$.
Category theory allows us to go into higher levels to explore more nuanced ideas of "sameness".

And this can go on infinitely [78]!
However, category theory does not know your specific situation.
It can't make a recommendation for the appropriate nuance level for your reality. It just gives you the tools.
We ultimately have to choose.