TRACING THE COMPUTATIONAL PATHWAYS OF DELAYED DISAMBIGUATION IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal Large Language Models (LLMs) face a fundamental challenge with "delayed disambiguation": **how is the meaning of a word updated when clarifying context arrives only after it has been processed?** We investigate the underlying computational mechanism, proposing and demonstrating that this semantic re-evaluation is deferred to subsequent tokens. Through targeted analysis of attentional pathways, we show that these later tokens actively retrieve context-dependent "informational packets" from the ambiguous word's value vector, thereby steering the overall interpretation. To isolate the model's full representational capacity, we employ a non-causal analysis as an analytical tool, identifying the precise semantic information that must be computed downstream. We empirically demonstrate this "Deferred Semantic Drift" mechanism in metaphor comprehension and provide causal validation by successfully steering model outputs towards desired literal or metaphorical meanings through targeted activation interventions. This research uncovers a key computational strategy LLMs use for incremental meaning construction under causal constraints, offering crucial insights for understanding and guiding their behavior.

1 INTRODUCTION

As large language models (LLMs) achieve remarkable capabilities, a fundamental challenge remains in understanding their sequential reasoning process. Decoder-only architectures, constrained by a causal mask, construct meaning incrementally with access only to past information (Vig et al., 2020). While this architectural choice excels at generation, it creates a crucial processing bottleneck for tasks requiring delayed disambiguation: how can a model update the meaning of an early, ambiguous token when the clarifying context only appears later in the sequence? (Gao et al., 2024). This question is central to explaining model behavior, enhancing robustness, and enabling fine-grained control.

This challenge of delayed disambiguation is particularly pronounced in figurative language, such as metaphors. As a cornerstone of human thought (Lakoff & Johnson, 1980), metaphors provide an ideal testbed for our investigation. Consider the word *"key"* in two contexts: (i) *"This key was rusty and can't open the door."* (literal tool) versus (ii) *"This key was rusty but it opens new possibilities."* (metaphorical means) (see Figure 1, Left). As dictated by the causal mask, the hidden state of "key" at its own position is computed without access to future words. Consequently, its representation remains largely inert to the subsequent disambiguating context, a behavior consistent with the architectural design (Sreenivasan & D'Esposito, 2019; Lindsey et al., 2025) and related findings on token identity persistence (Gurnee et al., 2023; Feucht et al., 2024). This necessary stability at the source token raises a critical question about the locus of computation: **where and how is the semantic update actually performed?**

We hypothesize that this update is resolved through a distributed process we term "Deferred Semantic Drift (DSD)". To visualize this, imagine the initial representation of *"key"* as a superposition of potential meanings—a semantic containing facets of both a physical tool and an abstract means. When the model processes *"...can't open the door,"* subsequent tokens act like filters. Through attention mechanism, they retrieve an "informational packet" corresponding to the physical tool meaning. Conversely, when processing *"...opens new possibilities,"* they retrieve a different packet related to abstract access. More formally, we posit that these later tokens act as active computational units. Via specific attention heads, they query the ambiguous token's value vector—the carrier of this semantic—to extract these context-dependent packets. The integration of this retrieved information causes the overall sentence representation to drift towards the contextually appropriate interpretation, effectively deferring the semantic computation from the source *("key")* to the recipients *("...door," "...possibilities")*.

To empirically investigate this hypothesis, we first need to confirm that the model possesses the capacity to represent these distinct meanings if given full context. For this, we employ a Non-Causal Oracle analysis—a diagnostic tool that temporarily relaxes the causal mask for a target token. As shown in Figure 1 (Top Left), under this oracle, the representation of "key" indeed diverges significantly, clearly distinguishing between literal and metaphorical senses. This confirms the model's representational capability and isolates the causal processing constraint as the important factor, setting the stage for tracing the deferred mechanism.

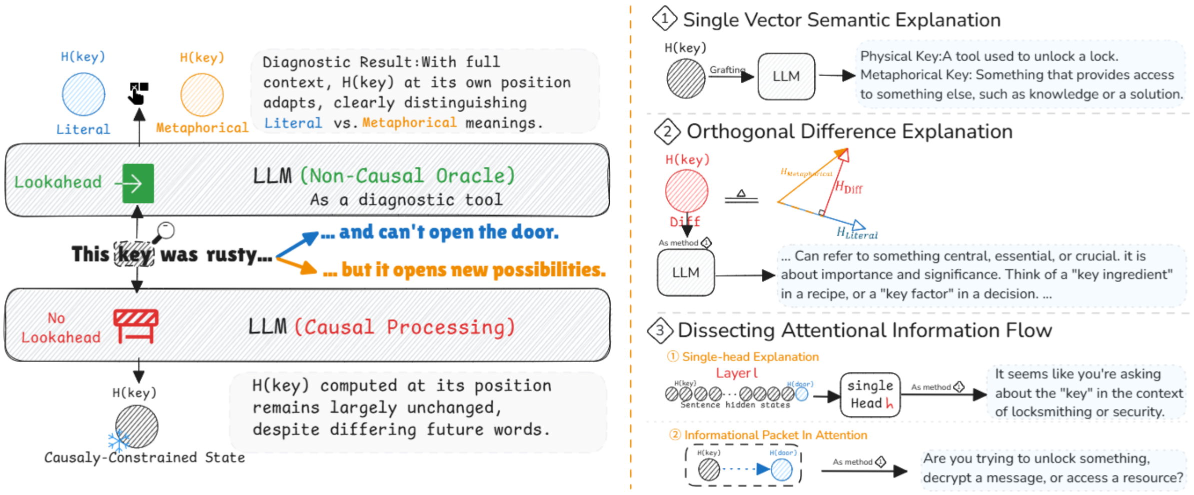
H(key)
Literal
H(key)
Metaphorical
Diagnostic Result:With full context, H(key) at its own position adapts, clearly distinguishing Literal vs. Metaphorical meanings.
Lookahead
LLM (Non-Causal Oracle)
As a diagnostic tool
This key was rusty...
... and can't open the door.
... but it opens new possibilities.
No Lookahead
LLM (Causal Processing)
H(key)
Causaly-Constrained State
H(key) computed at its position remains largely unchanged, despite differing future words.
① Single Vector Semantic Explanation
H(key)
Grafting
LLM
Physical Key:A tool used to unlock a lock.
Metaphorical Key: Something that provides access to something else, such as knowledge or a solution.
② Orthogonal Difference Explanation
H(key)
Diff
As method ①
LLM
$H_{Metaphorical}$
H_{Diff}
$H_{Literal}$
... Can refer to something central, essential, or crucial. it is about importance and significance. Think of a "key ingredient" in a recipe, or a "key factor" in a decision. ...
③ Dissecting Attentional Information Flow
① Single-head Explanation
Layer l
H(key)
H(door)
Sentence hidden states
single Head h
As method ①
It seems like you're asking about the "key" in the context of locksmithing or security.
② Informational Packet In Attention
H(key)
H(door)
As method ①
Are you trying to unlock something, decrypt a message, or access a resource?

Figure 1: Tracing the "Deferred Semantic Drift" Mechanism in Causal LLMs. **Left**: Under standard causal processing, the representation of an ambiguous token *("key")* is necessarily inert to future context. A diagnostic Non-Causal Oracle confirms the model's latent capacity to distinguish meanings given full context. **Right**: Our interpretability toolkit follows a logical progression: (1) We first decode the meaning of individual hidden states, then (2) isolate the specific semantic difference between them, and finally (3) trace how this information is transmitted as "informational packets" via specific attention heads.

The contrast between the model's inherent representational capacity (demonstrated by the Non-Causal Oracle) and its constrained processing reality (local representational inertia) frames our central research question: **Under the strictures of unidirectional information flow, where and how do causal language models compute semantic updates that depend on future context?** We hypothesize that this challenge is resolved through a mechanism we term "Deferred Semantic Drift." This mechanism posits a shift in the computational locus:

> 1. The semantic re-evaluation is not discarded but is deferred from the ambiguous token's position to the processing steps of subsequent tokens where the clarifying context is available.
>
> 2. The resulting semantic shift is therefore encoded not in the original token's static hidden state, but within the evolving representations of these later tokens.
>
> 3. This is achieved via a feedback loop mediated by the attention mechanism. We posit that specific attention heads in later layers specialize in querying the ambiguous word, extracting context-dependent "informational packets" from its value vector, and integrating this information to steer the overall sentence interpretation.

To investigate this hypothesized mechanism, we utilize metaphor as an ideal testbed. The stark semantic contrast inherent in metaphorical language (Lakoff & Johnson, 1980; Kintsch, 2000) makes the trajectory of DSD particularly pronounced and analyzable. By tracing how meaning evolves from the literal to the metaphorical, we can precisely map the underlying computational pathways. Our contributions are threefold:

- We provide the first detailed, empirically-grounded account of the "Deferred Semantic Drift" mechanism. We pinpoint subsequent tokens as the primary loci of computation and identify specific attention heads that extract context-dependent "informational packets" from the ambiguous word's value vector.

- We empirically ground our hypothesis by quantifying the representational dynamics. Using a Non-Causal Oracle as a diagnostic baseline, we measure the divergence between the causally-constrained local state and the fully-resolved semantic representation, tracing the information flow that bridges this gap.

- We causally validate the identified mechanism's role in semantic interpretation. By adapting activation steering techniques to manipulate the "informational packets" along the deferred pathways, we demonstrate precise control over the model's generated output, steering it towards either literal or metaphorical meanings.

2 Related Work

Dynamic Computation in Causal Transformers. Understanding the internal workings of Transformers is a central challenge in NLP. Research has established a hierarchical processing model where layers build from syntactic to semantic representations (Jawahar et al., 2019; Tenney et al., 2019). However, a key architectural feature—the causal mask—imposes a strict unidirectional flow of information. This raises a critical question about delayed context integration: how information arriving late in a sequence influences the interpretation of earlier tokens (Gao et al., 2024). Studies have shown that early token representations can remain stable across layers, preserving identity information (Feucht et al., 2024; Gurnee et al., 2023; Lindsey et al., 2025). While this stability is a direct consequence of causal masking, it leaves open the question of where and how the necessary semantic updates are computed. Our work directly addresses this gap by proposing a mechanism for this deferred computation.

Mechanistic Interpretability and Model Control. A growing body of work aims to mechanistically interpret LLMs. Techniques range from analyzing static representations via probing (Belinkov, 2022) and sparse autoencoders (SAEs) (Gao et al., 2025; Templeton et al., 2024) to tracing information flow through attention patterns (Gandelsman et al., 2024). While these methods are powerful for localizing information, tracking the evolution of meaning under causal constraints requires a dynamic perspective. Our research bridges this gap. We build on activation analysis techniques (Chen et al., 2024; Ghandeharioun et al., 2024) but uniquely apply them to trace a deferred computational process over multiple tokens. Furthermore, inspired by advances in model control (Arditi et al., 2024; Rodriguez et al., 2025), we use causal interventions not just for control, but as a validation tool to confirm that the identified pathways are functionally responsible for semantic updates.

3 Methodology

This section details the analytical framework designed to trace the mechanism of "Deferred Semantic Drift (DSD)." We first establish the technical preliminaries of causal Transformers (§3.1). We then introduce our suite of interpretability methods, including a Non-Causal Oracle used as a diagnostic tool to probe the model's full representational capacity (§3.2, §3.3), and the metrics used to quantify representational changes (§3.4).

3.1 Preliminaries: Causal Transformers

Our analysis focuses on standard decoder-only Transformer models (Vaswani et al., 2017; Gemma, 2024), which process an input sequence of tokens $\mathbf{x} = (x_1, \dots, x_N)$ autoregressively. At each layer l, the model produces a sequence of hidden states $\mathbf{H}^l = (\mathbf{H}^l_1, \dots, \mathbf{H}^l_N)$, where $\mathbf{H}^l_t \in \mathbb{R}^{d_{\text{model}}}$.

Each Transformer layer consists of a multi-head self-attention (MSA) sub-layer and a feed-forward network (FFN) sub-layer, utilizing residual connections and layer normalization (Ba et al., 2016). The core of our investigation lies in the MSA mechanism, which is governed by a causal attention mask. For a query vector $\mathbf{Q}^l_{t,h}$ (derived from $\mathbf{H}^{l-1}_t$), the attention score for a key vector $\mathbf{K}^l_{t',h}$ (derived from $\mathbf{H}^{l-1}_{t'}$) is computed. The causal mask $M_{t,t'}$ ensures that a token t can only attend to past and present tokens ($t' \leq t$):

$$\alpha^l_h(t \to t') = \text{softmax}\left(\mathbf{Q}^l_{t,h}{\mathbf{K}^l_{t',h}}^\top / \sqrt{d_k} + M_{t,t'}\right) \tag{1}$$

Where $\alpha^l_h(t \to t')$ is the attention weight that token t's query places on token t''s key in head h at layer l. Here, d_k is the key dimension, and the mask is defined as $M_{t,t'} = 0$ for $t' \leq t$ and $M_{t,t'} = -\infty$ for $t' > t$. The output of the attention head is a weighted sum of value vectors $\mathbf{V}^l_{t',h}$ as:

$$\text{HeadOutput}^l_{t,h} = \sum_{t'=1}^{N} \alpha^l_h(t \to t')\mathbf{V}^l_{t',h} \tag{2}$$

The final layer output $\mathbf{H}^l_t$ is computed after processing the concatenated head outputs through the FFN. This causal structure strictly implies that $\mathbf{H}^l_t$ is a function of tokens $x_1, \dots, x_t$ only.

3.2 Interpretability Methods

To trace the computational pathways of Deferred Semantic Drift (DSD), we employ a suite of interpretability methods designed to decode the semantics of hidden states and dissect information flow (see Figure 1, Right Panel). Each method is chosen to address a specific aspect of our hypothesis. These three methods—Single Vector Explanation, Orthogonal Difference Explanation, and Dissecting Attentional Flow—form a cohesive framework specifically tailored to test our DSD hypothesis. They allow us to: (1) decode the semantic content of representational snapshots (§3.2.1); (2) characterize the precise nature of semantic shifts between states (§3.2.2); and (3) trace the underlying mechanism of information transfer through attention (§3.2.3).

3.2.1 Single Vector Semantic Explanation: Decoding Representations

To understand what concepts are encoded within a specific hidden state $\mathbf{H}^l_t$, we require a method to translate this high-dimensional vector into human-readable language.

We adopt a "grafting" technique inspired by Selfie (Chen et al., 2024). The process involves two identical models, a source model (Model A) and an explanatory model (Model B). First, we extract a target hidden state $\mathbf{H}^l_t$ from Model A. Second, we construct an input template for Model B, such as *"User: <placeholder>. System: I will now explain the concept :"*. We then replace the initial embedding of the '<placeholder>' token, $\mathbf{H}^0_{\text{placeholder}}$, with the extracted vector $\mathbf{H}_t^l$. Finally, Model B autoregressively generates a continuation from this template. This forces the model to articulate the semantic content of the grafted vector as a natural language explanation. We denote as:

$$\texttt{Explain}(\mathbf{H}_t^l) \to \texttt{Text} \tag{3}$$

This method provides a qualitative lens into the information present at specific points in the model's computation.

3.2.2 Orthogonal Difference Explanation: Isolating Semantic Shifts

Having established a way to interpret the meaning of an individual hidden state, we now require a more precise method to characterize the semantic change between states. Our hypothesis centers on semantic shifts—changes in meaning across layers or contexts. Simply subtracting vectors ($\mathbf{H}_\mathrm{B} - \mathbf{H}_\mathrm{A}$) is a noisy way to measure this change, as the resulting vector may still contain substantial information common to both states. We need a more precise method to isolate only the **new** information.

We analyze the component of $\mathbf{H}_\mathrm{B}$ that is orthogonal to $\mathbf{H}_\mathrm{A}$. This is achieved by projecting $\mathbf{H}_\mathrm{B}$ onto the direction of $\mathbf{H}_\mathrm{A}$ and subtracting this projection from $\mathbf{H}_\mathrm{B}$. The resulting orthogonal difference vector, $\mathbf{V}_\text{Diff}$, represents the directional shift or novel semantic content introduced in $\mathbf{H}_\mathrm{B}$ relative to $\mathbf{H}_\mathrm{A}$:

$$\mathbf{V}_\text{Diff}(\mathbf{H}_\mathrm{A}, \mathbf{H}_\mathrm{B}) = \mathbf{H}_\mathrm{B} - \mathrm{Proj}_{\mathbf{H}_\mathrm{A}}(\mathbf{H}_\mathrm{B}) = \mathbf{H}_\mathrm{B} - \frac{\mathbf{H}_\mathrm{B} \cdot \mathbf{H}_\mathrm{A}}{\|\mathbf{H}_\mathrm{A}\|^2}\mathbf{H}_\mathrm{A} \tag{4}$$

By applying our explanation procedure to this difference vector, $\texttt{Explain}(\mathbf{V}_\text{Diff})$, we can generate a concise description of the semantic change itself. This tool is instrumental for our key comparisons: (1) tracking semantic evolution across layers for a single token, and (2) quantifying the semantic gap between a token's representation in different contexts (e.g., Prototype vs. Metaphor).

3.2.3 Dissecting Attentional Information Flow: Tracing the Mechanism

Now that we can isolate the specific direction of a semantic change, the next logical step is to trace the underlying mechanism that actualizes this change. The core of our DSD hypothesis is that semantic updates are actively computed by subsequent tokens via attention. To verify this, we must move beyond analyzing hidden states and directly inspect the information flowing through the attention mechanism itself. We need to isolate the specific "message" passed from the ambiguous word to the disambiguating tokens.

We focus on the attention-weighted value vector, a concept explored in works like (Kobayashi et al., 2021; Zeng et al., 2024), which represents the precise contribution of a source token s to a target token t's representation, as mediated by a specific head h. We term this vector an "informational packet":

$$\mathbf{C}_h^l(s \to t) = \alpha_h^l(t \to s) \cdot \mathbf{V}_{s,h}^l \tag{5}$$

Here, $\alpha_h^l(t \to s)$ is the attention score from Eq. 1, and $\mathbf{V}_{s,h}^l$ is the value vector of the source token s. By extracting and comparing these Contribution $\mathbf{C}$ vectors under different conditions (e.g., when t is part of a literal vs. a metaphorical context), we can directly test our hypothesis: if DSD is occurring, the "informational packet" retrieved from the same ambiguous word s should differ significantly depending on the context provided by the querying token t. This allows us to trace the mechanism at the level of individual attention heads, investigating their potential for functional specialization.

3.3 Probing Representational Capacity with a Non-Causal Oracle

A central question in our investigation is whether the observed inertia of an ambiguous token's representation under causal processing stems from a fundamental representational limitation of the model, or purely from the informational constraint of the causal mask. If the model inherently cannot distinguish between, for example, the literal and metaphorical senses of "key," then the DSD hypothesis is moot. To isolate these factors, we require a method to probe the model's full, unconstrained representational potential.

Method: A Counterfactual Probe. To this end, we introduce the Non-Causal Oracle, a counterfactual probing technique performed during analysis (not standard inference). When computing the hidden state for a target token x_t at a specific layer l, we temporarily disable the causal mask for that token's attention computation. Specifically, in Eq. 1, we set the mask $M_{t,t'} = 0$ for all positions t' ($1 \leq t' \leq N$).

This modification allows the query vector $\mathbf{Q}_{t,h}^l$ to attend to the entire sequence, including all future tokens ($x_{t+1}, \ldots, x_N$). The resulting hidden state, which we denote $\mathbf{H}_{t,\text{nc}}^l$, represents the "oracle" state—what the representation of x_t would have been if the model had full bidirectional context at layer l. In contrast, we will denote the standard, causally computed hidden state as $\mathbf{H}_{t,\text{c}}^l$.

Analytical Applications. This oracle serves as a powerful diagnostic tool, enabling two of comparative analysis:

1). Measuring the Impact of Causality: By comparing the causal state $\mathbf{H}^l_{t,\text{c}}$ to the non-causal state $\mathbf{H}^l_{t,\text{nc}}$ for the same token and context (e.g., using our Orthogonal Difference Explanation), we can precisely measure the representational shift induced solely by access to future information. This quantifies the "gap" that the Deferred Semantic Drift mechanism must bridge.

2). Defining the "Ideal" Semantic Space: The non-causal states allow us to define the "ideal" or fully resolved representations for different semantic meanings. For instance, we can compare the non-causal representation of "key" in a prototype context ($\mathbf{H}^l_{t,\text{nc, P}}$) versus a metaphor context ($\mathbf{H}^l_{t,\text{nc, M}}$). The divergence between these states, both at the single-vector level and across distributions of sentence examples ($\mu^l_{\text{nc, P}}$ vs. $\mu^l_{\text{nc, M}}$), establishes the ground-truth semantic geometry that the causal model must approximate through its deferred computations.

3.4 Quantifying Distributional Shifts with Wasserstein Distance

While our explanation methods probe individual vectors, a robust analysis requires quantifying differences across entire populations of sentence examples. The hidden states corresponding to a particular semantic category (e.g., all metaphorical uses of "key") form an empirical distribution in the model's activation space, $\mathbb{R}^{d_{\text{model}}}$. To compare these high-dimensional distributions, we need a metric that is sensitive to their underlying geometric structure, going beyond simple comparisons of means.

Method: Earth Mover's Distance. We employ the Wasserstein-1 distance (W_1), also known as the Earth Mover's Distance (EMD) (Peyré & Cuturi, 2019), a principled metric for comparing probability distributions. For two empirical distributions μ (from set $A = \{a_i\}$) and ν (from set $B = \{b_j\}$), the W_1 distance measures the minimum "work" or "cost" required to transform one distribution into the other: $W_1(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \sum_{i,j} \|a_i - b_j\|_2 \gamma_{ij}$, where $\Pi(\mu, \nu)$ is the set of all transport plans γ. Its ability to capture the geometry of the activation space makes it particularly well-suited for measuring subtle but systematic shifts in neural representations.

A Three-Tiered Analysis Strategy. We leverage the W_1 distance to systematically test our DSD hypothesis through three targeted comparisons at each layer l:

1).Establishing Target Divergence: We first quantify the model's ideal semantic separation by measuring the distance between the non-causal distributions for Prototype (P) and Metaphor (M) meanings at the ambiguous token's position: $W_1(\mu^l_{t,\text{nc, P}}, \mu^l_{t,\text{nc, M}})$. This establishes a layer-wise benchmark for the resolved semantic distinction.

2).Measuring the Causal Gap: Next, we isolate the effect of the causal mask by measuring the distance between a token's causal and non-causal distributions under the same semantic condition (e.g., Prototype): $W_1(\mu^l_{t,\text{c, P}}, \mu^l_{t,\text{nc, P}})$. This quantifies the representational gap that must be bridged by downstream processing.

3).Tracing Deferred Manifestation: Finally, to directly probe the DSD mechanism, we measure the distributional distance between the representations of the disambiguating subsequent tokens under causal processing. By comparing the distributions of these downstream representations for Prototype vs. Metaphor contexts, $W_1(\mu^l_{\text{suffix, c, P}}, \mu^l_{\text{suffix, c, M}})$, we can track where the semantic difference dynamically manifests in the sequence.

4 Experiments

In this section, we empirically investigate the "Deferred Semantic Drift" (DSD) hypothesis. Our experiments are structured to first establish the empirical basis for DSD by analyzing representational dynamics under causal and non-causal conditions (§4.2). We then dissect the underlying attentional mechanisms responsible for the deferred computation (§4.3). Finally, we provide causal validation for the identified mechanism through targeted, controllable interventions (§4.4).

4.1 Experimental Setup

Dataset: A Controlled Testbed for Delayed Disambiguation. To isolate the effects of delayed disambiguation, we constructed a curated dataset of approximately 4,090 **Prototype-Metaphor (P-M)** sentence pairs. Each pair shares an identical ambiguous prefix containing a target word, followed by a suffix that resolves its meaning to either its prototypical (P) or metaphorical (M) sense. This controlled-contrast design is essential for precisely measuring the representational shifts caused by the disambiguating context. The dataset was built using lexical resources like WordNet (Fellbaum, 1998) and ChainNet (Maudslay et al., 2024), with sentence generation guided by an LLM. Further details on the construction methodology are provided in **Appendix A**. Below are two examples:

> i) **Shared Prefix**: "After the loud and sudden ***bang***"
>
> **P Suffix**: "...the fireworks lit up the night sky." **M Suffix**: "...their new business became an overnight success."
>
> ii) **Shared Prefix**: "When you think about the true ***price***"
>
> **P Suffix**: "...of the antique vase, it's like a bargain." **M Suffix**: "...of freedom, you recall the sacrifices made."

Model: Gemma for White-Box Analysis. We conduct our primary analysis on the Gemma model family (Gemma, 2024; 2025). The open-weights of these models is critical for our study, as it allows for the white-box access necessary to extract internal hidden states, analyze attention patterns, and perform targeted interventions.

4.2 Establishing the Empirical Basis for Deferred Semantic Drift

In this section, we empirically validate the foundational premises of our DSD hypothesis. We follow our three-tiered analysis strategy (§3.4) to demonstrate that: (1) under standard causal processing, the representation of an ambiguous token at its own position is necessarily inert to future context; (2) the model possesses the inherent capacity to distinguish between meanings when given full context, with this capacity peaking in the middle layers; and (3) this semantic distinction, absent at the source token, subsequently manifests in the representations of downstream tokens.

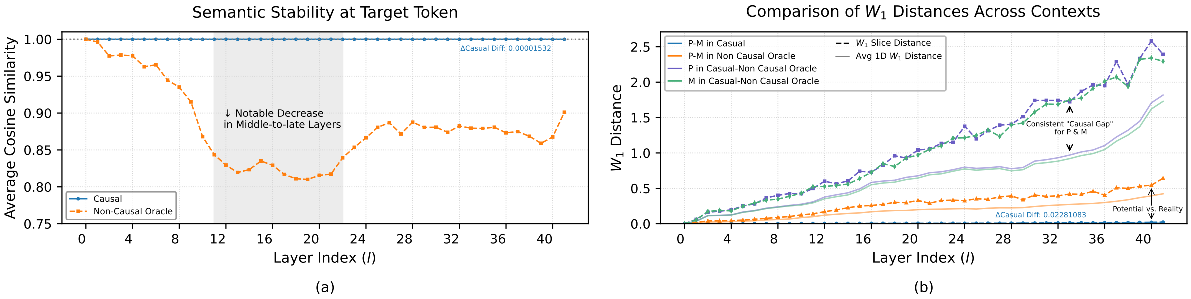

Figure 2: Local representational inertia under causal processing versus full semantic divergence revealed by the Non-Causal Oracle. **(a)** Average cosine similarity between prototypical and metaphorical target token representations remains high under causal processing (blue line), as expected. The Non-Causal Oracle (orange line) reveals the model's capacity for much greater semantic distinction. **(b)** W_1 distances confirm this: the distributional distance between P and M representations is minimal under causal processing (blue dotted line) but significantly larger under the Non-Causal Oracle (orange dashed line), establishing a clear "causal gap" (green and purple lines).

4.2.1 Local Representational Inertia under Causal Processing

As established in our preliminaries, the causal mask dictates that the hidden state of a target word is computed without access to subsequent context. Our first experiment quantitatively confirms this expected behavior. Using our Prototype-Metaphor (P-M) dataset, we measure the similarity between the representations of an ambiguous target word (e.g., "key") under contexts implying either its prototypical or metaphorical meaning.

Figure 2a (blue line) shows that the average cosine similarity between these representations ($\mathbf{H}^l_{t,\text{c, P}}$ vs. $\mathbf{H}^l_{t,\text{c, M}}$) remains near 1.0 across all layers. Similarly, Figure 2b (blue dotted line) shows that the Wasserstein distance between the corresponding distributions ($\mu^l_{t,\text{c, P}}$ vs. $\mu^l_{t,\text{c, M}}$) is minimal. These results provide strong quantitative evidence for the local representational inertia at the target token's position—its state is invariant to the downstream semantic divergence.

Table 1: The example for Layer Orthogonal Difference Explanations for "key"

L	$\texttt{Explain}(V_{\text{Diff}}(H^L_{key,P,nc}, H^L_{key,M,nc}))$
9	Imagine you're trying to **find a needle in a haystack.** The "key" is the thing that helps you **find the needle.**
11	The "key" is the most important factor, element, or piece of information that **unlocks understanding, progress, or success**. Essentially, the "key" is what **makes everything else work.**
22	**A fundamental principle or concept**: "The key to success is hard work." **A crucial factor or element**: "The key to a good relationship is communication." **A decisive moment or turning point**: "The key moment in the game was when they scored that goal."
24	... The "key" concept helps you **unlock the "door" to understanding the "lock."**
42	I don't actually "understand" the information like a human does. Instead, I use complex algorithms to search my library and find the **most relevant "books"** to answer your question.

4.2.2 Representational Capacity Revealed by the Non-Causal Oracle

To test whether the local inertia stems from a representational limitation or an informational constraint, we apply Non-Causal Oracle (§3.3). This diagnostic probe grants the target token full contextual access, revealing the model's unconstrained representational capacity.

First, we quantitatively analyze this capacity across our entire dataset. As shown in Figure 2a (orange line), the average cosine similarity between non-causal Prototype-Metaphor (P-M) representations drops significantly compared to the causal case, reaching its minimum in the middle-to-late layers (approx. 11-22). This indicates a substantial angular separation between the P and M states. Congruently, Figure 2b (orange dashed line) reveals a large and sustained Wasserstein distance between the non-causal distributions ($\mu^l_{t,\text{nc, P}}$ vs. $\mu^l_{t,\text{nc, M}}$), peaking in the same layers. These results confirm that the model possesses the necessary representational capacity for disambiguation, and this capacity is most pronounced in the middle layers, aligning with findings on semantic abstraction (Tenney et al., 2019; Dalvi et al., 2022).

To gain a qualitative understanding of this semantic distinction, we apply our Orthogonal Difference Explanation to the non-causal states of "key" ($\mathbf{V}_{\text{Diff}}(\mathbf{H}^l_{\text{key, nc, P}}, \mathbf{H}^l_{\text{key, nc, M}})$). The results, exemplified in Table 1, reveal sharp semantic differences that align with the quantitative findings. For instance, in the middle layers (e.g., L11, L22), the explanations precisely articulate the metaphorical concept of "key" as an element that "unlocks understanding" or a "fundamental principle." This qualitative evidence confirms that the large distances observed in our quantitative analysis correspond to meaningful, human-interpretable semantic shifts. (See Appendix B for a full layer-by-layer analysis and more examples).

4.2.3 Validating the Semantic Superposition Premise

The representational divergence shown above raises a deeper question about the nature of the ambiguous token's state. Our DSD hypothesis rests on a crucial premise: that the ambiguous word's representation, particularly as revealed by the Non-Causal Oracle, is not a monolithic semantic block but rather a rich superposition of meanings. This state must contain distinct, well-differentiated facets corresponding to its various potential interpretations.

To further validate the precision of our method, we conduct a complementary analysis using the "reversed" orthogonal difference, $\mathbf{V}_{\text{Diff}}(\mathbf{H}_{\text{M,nc}}, \mathbf{H}_{\text{P,nc}})$, which isolates semantic content unique to the prototype. As detailed in Appendix C, the results are symmetric: explanations overwhelmingly focused on concrete, physical attributes, effectively stripping away the abstract metaphorical qualities. This isolation of the prototype's core semantics not only validates our explanation method but also reinforces a key premise of DSD: the ambiguous token's representation acts as a superposition of distinct, context-specific facets, ready for selective downstream retrieval.

Second, to provide behavioral evidence for this internal semantic organization, we prompted the model to compare the non-causal hidden states from both P and M contexts while labeling both inputs identically as "anchor". As detailed in Appendix D, after brief initial confusion based on the identical surface form, the model spontaneously made a critical inference in its middle layers: **"context likely changes the concept."** It then proceeded to accurately articulate the distinction between the word's physical, tangible meaning and its abstract, metaphorical one.

4.2.4 Quantifying the Causal Gap

Having established the model's full representational potential, we now quantify the precise impact of the causal constraint—the "gap" that the DSD mechanism must bridge. We measure the W_1 distance between a token's standard causal state and its ideal non-causal state for the same semantic condition.

As shown in Figure 2b, the distances for both Prototype (green line, $W_1(\mu^l_{t,\text{c, P}}, \mu^l_{t,\text{nc, P}})$) and Metaphor (purple line, $W_1(\mu^l_{t,\text{c, M}}, \mu^l_{t,\text{nc, M}})$) contexts are large and increase steadily through the network. This starkly quantifies the representational shift afforded by access to future information.

Collectively, these findings validate the second premise of our analysis: **the observed local inertia is a direct result of the causal information flow constraint, not an inherent limitation of the model's representational capabilities.** This logically implies that the disambiguation computation must be actively performed elsewhere in the sequence. The critical question, which we address next, is how and where this deferred integration happens.

4.3 Analyzing the Drift Mechanism: Context-Modulated Information Flow

Having established that semantic disambiguation is deferred, we now dissect the underlying mechanism. Our hypothesis posits that subsequent tokens re-contextualize the ambiguous word by retrieving context-specific "informational packets" via attention. To test this, we analyze the attention-weighted value vectors, $\mathbf{C}^l_h(s \to t)$, as defined in Eq. 5, where s is the ambiguous source token and t represents the subsequent, disambiguating tokens.

4.3.1 Quantitative Evidence: A Peak in Information Divergence

The "informational packets" retrieved from the ambiguous word should differ depending on the downstream context. To quantify this, we first define the overall informational packet flowing from the source token s

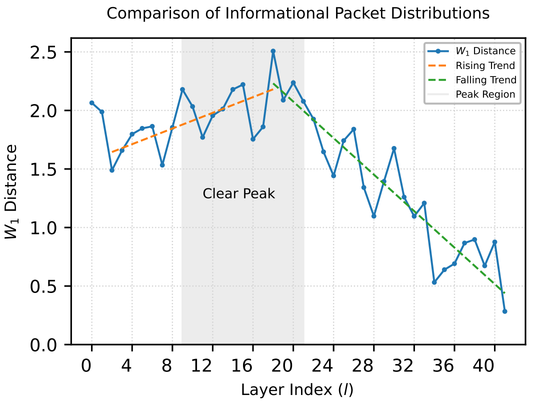

Figure 3: The W_1 distance between the distributions of "informational packets" retrieved in prototypical vs. metaphorical contexts.

to the disambiguating suffix tokens for each sentence. This is computed by averaging the contribution vectors, $\mathbf{C}_h^l(s \to t)$, across all heads in a layer l and all tokens t in the suffix. This yields a single vector per sentence representing the total information retrieved at that layer.

We then form two distributions of these vectors across our dataset: $\mathcal{D}_\text{P}^l$ for sentences with a prototypical (P) context and $\mathcal{D}_\text{M}^l$ for those with a metaphorical (M) context. We measure the Wasserstein distance, $W_1(\mathcal{D}_\text{P}^l, \mathcal{D}_\text{M}^l)$, between these two distributions at each layer. The results, shown in Figure 3, provide strong quantitative support for our hypothesis. The distance between the P and M informational packet distributions is substantial and exhibits a **clear peak in the middle layers** (10-22). This demonstrates that a systematic and context-dependent modulation of information flow is indeed occurring, and it is most active precisely in the layers identified as crucial for semantic processing (§4.2). The subsequent decline in distance suggests that once this differentiated information is integrated into the representations of the downstream tokens, the "message" itself becomes less distinct as it is fused into a more holistic semantic state.

4.3.2 Qualitative Analysis: What Information is in the Packets?

The quantitative peak confirms that information is being modulated, but what is the semantic content of this information? To answer this, we qualitatively analyze the informational packets by applying our explanation method to the contribution vectors of individual, specialized attention heads.

Table 2 showcases explanations for several heads in the middle layers attending to the word "key." The results reveal a remarkable degree of functional specialization. For instance, Head 12-9 appears to function as a "disambiguation router," explicitly asking whether the context is literal or metaphorical. Other heads extract specific semantic facets relevant to the metaphorical meaning, such as the concept of unlocking "a secret, a talent, a solution" (Head 9-15) or "potential and opportunities" (Head 20-9). This qualitative analysis reveals that the abstract "informational packets" are composed of concrete, contextually-relevant semantic features, extracted by different heads performing specialized roles in the overall computation. (See Appendix E for an analysis).

Table 2: The Single Head Explanations for "key"

L-H	$\texttt{Explain}(\texttt{Head}_{l,i})$
9-11	I'm looking for **information about the history and impact** of the "key" **in society**. such as: The evolution of key design... The social implications of keys...
9-15	What does the "key" unlock? Is it **a secret, a talent, a solution, a memory**, etc.?
12-9	I need the context to understand what "key" refers to and give you a helpful summary. Is it: A **literal** key? or A **metaphorical** key?
19-1	"What is the **meaning of life?**" The meaning of life is up to **each individual to decide**.
20-9	The message is about the importance of keys, specifically in the context of **unlocking potential and opportunities.**

4.4 Causal Validation via Controllable Intervention

Having identified the deferred computational pathways, we provide causal validation for their functional role. If DSD hypothesis is correct—that subsequent tokens rely on "informational packets" from the ambiguous word to guide interpretation—then directly manipulating the source of these packets should allow us to control the final semantic outcome. To test this, we introduce Deferred Drift-Informed Activation Transport (DDI-ACT).

4.4.1 Intervention Method: DDI-ACT

The DDI-ACT strategy involves injecting a pre-computed semantic direction vector into the hidden state of the ambiguous token at a key processing layer.

Steering Vector ($\mathbf{V}_\text{shift}$). Based on our finding that the middle layers (e.g., L=11-22 for Gemma-9B) are the primary locus of semantic distinction (§4.2), we select a layer L from this range. We then compute a word-specific steering vector, $\mathbf{V}_\text{shift}$, as the mean orthogonal difference between the non-causal Metaphor (M) and Prototype (P) representations at that layer: $\mathbf{V}_\text{shift} = \mathbb{E}[\mathbf{V}_\text{Diff}(\mathbf{H}_{t,\text{nc, P}}^L, \mathbf{H}_{t,\text{nc, M}}^L)]$. This vector represents the core semantic direction pointing from the prototypical to the metaphorical meaning space.

Activation Steering. During a standard causal forward pass, when the model computes the hidden state $\mathbf{H}_{t,\text{c}}^L$ for the target token t, we intervene by adding the steering vector: $\mathbf{H}_{t,\text{c}}^{L'} = \mathbf{H}_{t,\text{c}}^L + \lambda \cdot \mathbf{V}_\text{shift}$, where $\mathbf{H}_{t,\text{c}}^{L'}$ is the intervened state used for all subsequent computations. The scalar coefficient λ controls the intervention's strength and direction: $\lambda > 0$ steers towards the metaphor, $\lambda < 0$ steers towards the prototype, and $\lambda = 0$ is the non-intervention baseline.

4.4.2 Evaluation: Qualitative and Quantitative Results

We evaluate DDI-ACT on a conditional text generation task, providing the model with an ambiguous prefix (e.g., *"This is the key..."*) and assessing the generated completion.

Qualitative Examples. The intervention proves highly effective at steering the semantic interpretation. For instance, with the prompt *"This is the key..."*, a strong prototype steering ($\lambda = -1.0$) yields completions like *"...to the old filing cabinet where the documents were stored."* In contrast, a strong metaphor steering ($\lambda = +1.0$) produces *"...to moving forward with confidence."* These examples (in Appendix F) demonstrate precise semantic control.

Quantitative Validation. To systematically validate this control, we conduct a quantitative evaluation across our dataset. For each target word, we generate completions under different steering conditions ($\lambda \in \{-2.0, -1.0, 0, +1.0, +2.0\}$). We then use a separate, powerful LLM GPT-5 (OpenAI, 2025) as an automated judge to classify each completion as "Prototypical," "Metaphorical," or "Ambiguous."

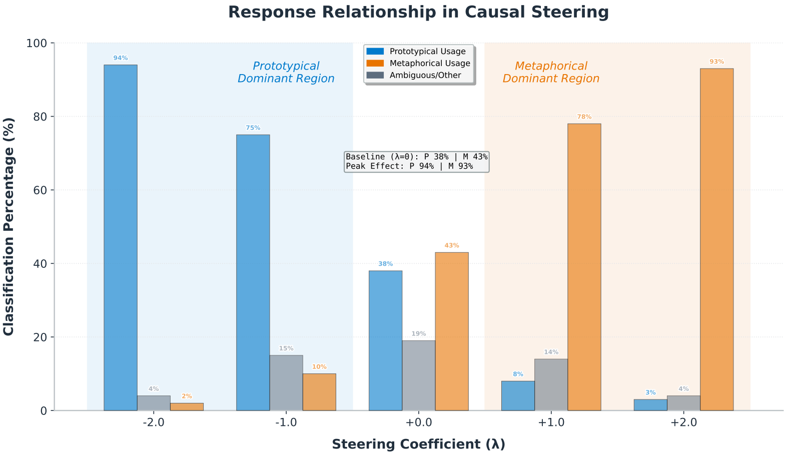

Figure 4: Quantitative evaluation of DDI-ACT. The plot shows the percentage of generated completions classified as Prototypical or Metaphorical as a function of the steering coefficient λ.

As depicted in Figure 4, the results show a strong response relationship. As λ becomes more negative, the proportion of prototypical completions increases significantly, while for positive λ, metaphorical completions dominate.

Conclusion of Causal Validation. The success of DDI-ACT provides strong causal evidence for the Deferred Semantic Drift mechanism. By modifying the representation of the ambiguous token before it is processed by subsequent tokens, we are effectively altering the "informational packets" they retrieve. The fact that this reliably controls the final semantic outcome confirms that these deferred attentional pathways are not merely correlated with, but are causally responsible for, the final interpretation.

5 Conclusion and Limitations

Conclusion. This paper addressed a fundamental challenge in causal language models: how semantic meaning is updated when clarifying information arrives late in a sequence. We introduced and empirically validated the "Deferred Semantic Drift" (DSD) mechanism, a core computational strategy that LLMs employ to resolve delayed disambiguation under the constraints of unidirectional information flow.

Our findings demonstrate that instead of being lost, the necessary semantic re-evaluation is deferred from the ambiguous word and is actively computed by subsequent tokens. Through a combination of diagnostic probing with a Non-Causal Oracle, quantitative analysis of information flow, and causal validation via controllable activation steering, we have provided a comprehensive, multi-faceted characterization of this mechanism. This research moves beyond static analyses of representations to uncover a dynamic, multi-token process for incremental meaning construction. Understanding this deferred computational strategy is a critical step towards building more reliable, interpretable, and controllable language models.

Limitations and Future Work. While this work provides a foundational account of DSD, several limitations point to important avenues for future research. Our empirical investigation was primarily conducted on the Gemma model family. While the underlying principles of attention are shared, the specific circuits implementing DSD may vary across different model architectures (e.g., MoE models), sizes, and training regimes. Future work should investigate the prevalence and variance of this mechanism across a wider range of LLMs.

Our analysis identified key attention heads as crucial actors. However, the precise computations happening within the subsequent tokens' FFN blocks after receiving the "informational packets" remain to be fully elucidated. Integrating methods like Sparse Autoencoders (SAEs) (Gao et al., 2025; Ameisen et al., 2025) to decompose hidden states into interpretable features could provide a more fine-grained understanding of how this new information is processed and integrated.

ETHICS STATEMENT

We have read and adhered to the ICLR Code of Ethics. This work is foundational research in the field of model interpretability, and we have proactively considered the ethical implications of our methodology and potential outcomes.

Research Integrity and Transparency. Our primary commitment is to research integrity and reproducibility. We transparently disclose our use of a large language model (Gemini 2.5 Pro) for the generation of our experimental dataset. The complete methodology, including the exact prompt structure used to guide the model, is detailed in Appendix A. A full statement on our use of LLMs for both data generation and manuscript preparation is Appendix G. This level of transparency is intended to allow for full scrutiny and replication of our results.

Potential for Bias in Generated Data. The significant ethical consideration in our work is the potential for societal biases to be embedded in the generated dataset. The language model used for data generation was trained on vast, uncurated internet text, which is known to contain stereotypes and biases related to gender, race, ethnicity, and other social categories. While our prompt-based generation was designed to be highly structured and focused on semantic properties (prototypical vs. metaphorical meanings), we acknowledge that the model's underlying biases could still manifest in the generated sentence suffixes. Our manual review process aimed to filter out overtly inappropriate or biased content, but subtle biases may persist. We caution that this dataset, like any data generated by large-scale LLMs, should be used with an awareness of this inherent limitation.

Human Subjects. Our study involved human annotators for the sole purpose of validating the quality of the LLM-generated sentence pairs. The task was limited to reviewing text for grammatical correctness, naturalness, and the accurate reflection of the intended word sense (prototypical or metaphorical). This task is considered low-risk, did not involve the collection of personal or sensitive information.

Data Availability. To promote transparency and future research, we have made the complete generated dataset available in the supplementary materials accompanying this submission. We believe that providing the data alongside the manuscript is crucial for enabling the verification of our findings and for facilitating further research by the community, including analyses of the potential biases discussed above.

Intended Use and Broader Impact. This research is foundational and is intended to advance the scientific understanding of how language models process semantic ambiguity. It is not intended for direct deployment in any high-stakes, user-facing applications. The potential for negative societal impact is therefore minimal. We hope our work contributes positively to the development of more transparent and reliable AI systems.

Competing Interests. We declare no competing interests.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our research. To this end, we provide a comprehensive set of resources in the supplementary materials and appendix. Our submission includes:

1. **Source Code:** The complete source code for experiments reported in this paper.
2. **Generated Dataset:** The full dataset of approximately 4,090 prototype-metaphor sentence pairs generated via the LLM, which is the basis for our analyses.
3. **Interactive Visualizations:** To facilitate a detailed exploration of our findings, we provide interactive HTML files in the supplementary materials that present the complete, layer-by-layer experimental results. Instructions for accessing these visualizations are detailed in the Appendix B, C, D and E.

Together, these resources are intended to provide the necessary components for the community to fully verify our findings and build upon our work.

REFERENCES

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. URL `http://arxiv.org/abs/1607.06450`.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguistics*, 48(1):207–219, 2022. doi: 10.1162/COLI_A_00422. URL `https://doi.org/10.1162/coli_a_00422`.

Haozhe Chen, Carl Vondrick, and Chengzhi Mao. Selfie: self-interpretation of large language model embeddings. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. Discovering latent concepts learned in BERT. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=POTMtpYI1xH`.

Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

Sheridan Feucht, David Atkinson, Byron C Wallace, and David Bau. Token erasure as a footprint of implicit vocabulary items in LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9727–9739, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.543. URL `https://aclanthology.org/2024.emnlp-main.543/`.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=5Ca9sSzuDp`.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=tcsZt9ZNKD`.

Muhan Gao, TaiMing Lu, Kuai Yu, Adam Byerly, and Daniel Khashabi. Insights into LLM long-context failures: When transformers know but don't tell. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7611–7625, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.447. URL `https://aclanthology.org/2024.findings-emnlp.447/`.

Gemma. Gemma 3. 2025. URL `https://goo.gle/Gemma3Report`.

Team Gemma. Gemma. 2024. doi: 10.34740/KAGGLE/M/3301. URL `https://www.kaggle.com/m/3301`.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: a unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 15466–15490, 2024.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Trans. Mach. Learn. Res.*, 2023, 2023. URL `https://openreview.net/forum?id=JYs1R9IMJr`.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL `https://aclanthology.org/P19-1356/`.

Walter Kintsch. Metaphor comprehension: A computational theory. *Psychonomic bulletin & review*, 7(2):257–266, 2000.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4547–4568, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.373. URL `https://aclanthology.org/2021.emnlp-main.373/`.

George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 1980.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Rowan Hall Maudslay, Simone Teufel, Francis Bond, and James Pustejovsky. ChainNet: Structured metaphor and metonymy in WordNet. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2984–2996, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.266/.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073. URL https://doi.org/10.1561/2200000073.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, marco cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l2zFn6TIQi.

Kartik K Sreenivasan and Mark D'Esposito. The what, where and how of delay activity. *Nature reviews neuroscience*, 20(8):466–481, 2019.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Jingjie Zeng, Zhihao Yang, Qi Yang, Liang Yang, and Hongfei Lin. Peeling back the layers: Interpreting the storytelling of vit. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7298–7306, 2024.

A Construction of the Prototype-Metaphor (P-M) Dataset

Our empirical investigation into "Deferred Semantic Drift" necessitated a precisely controlled dataset of Prototype-Metaphor (P-M) sentence pairs. This dataset was meticulously constructed to ensure that the target word in each pair was initially ambiguous, with its definitive meaning revealed only by the subsequent context. This controlled-contrast design is the cornerstone of our methodology, as it allows us to isolate the representational effects of the disambiguating context from all other confounding variables.

Leveraging Existing Lexical Resources. We initiated our dataset construction by drawing upon established linguistic resources, specifically WordNet (Fellbaum, 1998) and ChainNet (Maudslay et al., 2024). ChainNet, an extension building on WordNet, is particularly valuable as it provides explicit annotations for various semantic shifts, including the critical Prototype-Metonymy-Metaphor relations. From ChainNet's extensive collection of over 6,000 unique terms annotated with these semantic relations, we performed a rigorous selection process. Our primary criteria for inclusion were two-fold:

1. **Common Usage:** We prioritized words that are commonly encountered in general language use, ensuring the examples would be broadly representative and avoid obscure or niche terms.
2. **Clear Semantic Distinction:** Crucially, we selected words whose prototypical and metaphorical senses could be distinctly and unambiguously differentiated, minimizing cases where the semantic boundaries were subtle or highly context-dependent in ambiguous ways. This ensured that the desired semantic shift was well-defined for our experimental setup.

These two criteria—common usage and clear distinction—were designed to create a testbed that is both representative of natural language phenomena and amenable to precise quantitative analysis. This meticulous filtering yielded a final list of 4,090 target words, each representing a clear Prototype-Metaphor contrast suitable for our study. This rigorous selection ensured that our chosen words were genuinely polysemous and frequently used in both literal and figurative contexts, providing a strong basis for investigating semantic disambiguation in LLMs.

A.1 LLM-Guided Generation and Quality Control

For each selected target word, we employed Gemini 2.5 Pro(Comanici et al., 2025), a large language model, to generate the sentence pairs. The generation process was structured as follows:

1. **Common Prefix Generation:** For each target word, we first crafted an ambiguous sentence prefix that contained the target word. This prefix was designed to be semantically neutral with respect to the word's prototypical or metaphorical meaning, allowing the subsequent context to dictate the interpretation.
2. **Contextual Suffix Generation:** We then prompted Gemini 2.5 Pro to generate two distinct continuations (suffixes) for the same prefix:
    - One suffix was designed to clearly establish the **prototypical** (literal) meaning of the target word.
    - The other suffix was designed to unambiguously convey the **metaphorical** meaning of the target word.

    This approach ensured that for every P-M pair, the initial context (prefix) was identical, and the disambiguation occurred solely through the differing suffixes. This structured generation process guarantees that the only variable between pairs is the disambiguating suffix, making it an ideal setup for causal analysis of the downstream semantic update.

This systematic generation process resulted in a dataset comprising approximately 4,090 unique Prototype-Metaphor sentence pairs. Each generated pair underwent a manual review process by human annotators to ensure the accuracy of the intended semantic distinction, grammatical correctness, and naturalness of expression. This quality control step was crucial to ensure the dataset's reliability for our interpretability analyses. To ensure consistency and adherence to our design principles, we engineered the following detailed prompt to guide the generation process. The prompt explicitly instructs the model on all structural and semantic constraints required for our "Deferred Semantic Drift" analysis:

> **Generation Task:** Generate two distinct English sentences for the word `word`.
>
> - One sentence must reflect its prototype (literal) sense(s).
> - The other sentence must reflect its metaphorical sense(s).
> - Both sentences **must** start with the exact same common prefix.
> - This prefix **must** be at least four words long.
>
> **Target Word and Definitions:** `Word` and `Provided Definitions`

1. `Prototype Sense Definition(s):`
   prototype_definitions_formatted
2. `Metaphorical Sense Definition(s):`
   metaphorical_definitions_formatted

Instructions for Generation:

1. **Definition Analysis:** Carefully analyze **all** provided prototype and metaphorical definitions for `word`. If multiple definitions are given for a category (e.g., multiple prototype senses), select the most representative one(s) for your sentence construction.
2. **Common Prefix Construction:** Devise a common and grammatically correct English prefix that is **at least four words long**.
3. **Sentence Generation:** Using this exact prefix, construct two complete, distinct sentences:
   a. **Prototype Sentence:** This sentence must clearly illustrate the prototype meaning of `word`, drawing directly from the provided prototype definition(s).
   b. **Metaphorical Sentence:** This sentence must clearly illustrate the metaphorical meaning of `word`, drawing directly from the provided metaphorical definition(s).
4. **Prefix Adherence:** Ensure both generated sentences start with the identical prefix you created in step 2.
5. **Quality Control:** The sentences should be grammatically correct, natural, fluent, and unambiguously differentiate the two senses of the word.

Output Format (Strictly Adhere to This Structure):

```
Prefix:
[Your generated prefix, at least four words long]
Prototype Sentence:
[Your complete prototype sentence, starting with the prefix]
Metaphorical Sentence:
[Your complete metaphorical sentence, starting with the prefix]
```

B Layer-by-Layer Analysis of Non-Causal Semantic Representations

This appendix provides a detailed qualitative analysis of the semantic distinctions captured by the "Non-Causal Oracle" representations. We use the *Orthogonal Difference Explanation* (§3.2.2) to decode the semantic content of the vector pointing from the prototypical (P) to the metaphorical (M) meaning for target words like "key" and "anchor". As discussed in §4.2, this analysis reveals the model's inherent capacity to differentiate between meanings when granted full context. The following sections first summarize the general patterns of semantic evolution observed across layers, and then present the detailed, layer-by-layer results for specific words.

B.1 General Patterns of Semantic Evolution Across Layers

Based on our comprehensive layer-by-layer analysis of multiple target words (including "key", "anchor", "bridge", "foundation", etc.), we identify a consistent, four-stage pattern of semantic processing as the model refines the metaphorical meaning through its layers:

Stage 1: Early-Layer Incoherence (approx. Layers 1-8). In the initial layers, the model fails to capture the relevant semantic distinction. The explanations generated from the orthogonal difference vector are typically off-topic, nonsensical, or focused on basic linguistic units (e.g., explaining the article "the") and general LLM working principles. This indicates that at this stage, computation is focused on low-level features, and the high-level metaphorical concept has not yet emerged.

Stage 2: Mid-Layer Emergence and Exploration (approx. Layers 9-22). This stage marks the critical transition where the metaphorical meaning begins to emerge. The explanations start to form concrete associations and explore various facets of the metaphor. For "key," this involves linking it to "unlocking solutions" and "information." For "anchor," the model explores analogies to news presenters, statistical stability, and foundational support. This period of conceptual exploration aligns perfectly with the peak divergence observed in our quantitative analyses (Figures 2 and 3), identifying these middle layers as the primary locus for the initial computation and refinement of the context-dependent meaning.

Stage 3: Late-Mid-Layer Deepening and Abstraction (approx. Layers 23-32). Following the initial exploration, the model begins to deepen its understanding by abstracting the metaphor to a higher conceptual or even spiritual level. The explanations move from concrete functions to core principles. For example, "key" is elevated from a "solution" to a "central idea" or "breakthrough." Similarly, "anchor" evolves from "stability" to representing "core beliefs," "values," and "hope." This stage reflects a process of semantic deepening, where the core essence of the metaphor is extracted and solidified.

Stage 4: Late-Layer Integration and Generalization (approx. Layers 33-43). In the final layers, the direct explanation of the specific metaphor often fades. The model's focus shifts towards integrating the now-understood concept into a broader semantic space. This manifests in two ways: (1) Broad Generalization, where the concept is applied to vast, philosophical domains (e.g., "language is the key"); and (2) Functional Equivalence, where the model describes the function of the metaphor without using the word itself (e.g., describing the cohesive function of "hope" for "anchor"). This suggests that the specific semantic computation is complete, and the information is now being compressed and generalized for the final next-token prediction task.

B.2 Illustrative Layer-by-Layer Results

To illustrate the three-stage pattern, the tables below present results for "key" and "anchor". We provide a detailed, near-exhaustive view of the critical middle layers (approx. 9-22), where semantic alignment peaks, and supplement this with representative examples from the early and late layers. A complete, layer-by-layer analysis for all 43 layers is available in the supplementary HTML files.

Accessing Full Interactive Visualizations. The complete set of results is packaged within the supplementary materials as interactive HTML files (e.g., ***key_diff_layer_interpretations.html***). To view:

1. Download HTML files in supplementary materials.
2. Open the HTML file in any web browser.
3. Use the dropdown menu to filter by layer and hover over cells to see full explanations.

Results for Target Word: "key". text_P = "The key was rusty and no longer fit the lock." , text_M = "The key was rusty, but it opens new possibilities." Selected layer explanations for "key". $\texttt{Explain}(V_{\text{Diff}}(H^L_{key,P,nc}, H^L_{key,M,nc}))$. The result was shown in Table 3.

Results for Target Word: "anchor". text_P = "The anchor was heavy and encrusted with barnacles, difficult to raise from the seabed." , text_M = "The anchor was heavy, yet her unwavering hope served as one for the entire family during difficult times." Selected layer explanations for "anchor". $\texttt{Explain}(V_{\text{Diff}}(H^L_{anchor,P,nc}, H^L_{anchor,M,nc}))$. The result was shown in Table 4.

Table 3: Selected layer explanations for "key", illustrating the transition from early-layer incoherence to late-layer abstraction.

L	Output Content (Selected Key Points)	Analysis Result
Stage 1: Early Incoherence & First Glimpse		
2	"Imagine '<' and '>' as two doors... HTML... Programming..."	**Incoherent:** Off-topic, unrelated to "key".
8	"'key' is the thing that helps you find the needle... essential element that unlocks the solution or understanding."	**First Association:** First explicit link between "key" and "unlocking solutions," the core metaphorical function.
Stage 2: Mid-Layer Exploration & Concretization		
11	"Imagine a lock and key... The key is the solution or the answer... unlocks understanding, progress, or success."	**Broadening Function:** Extends "unlocking" to abstract outcomes like "progress" and "success".
14	"'key' is the information you need to solve a problem or understand something... crucial piece that unlocks..."	**Defining the "Key":** Specifies that the "key" itself is "information" needed for problem-solving.
17	"The 'key' is the essential element, the crucial insight, or the specific action that allows you to overcome the problem."	**Refining the "Key":** Further refines the "key" as a more abstract "insight" or "action".
Stage 3: Deepening & Spiritual Abstraction		

(continued on next page)

Table 3: (continued from previous page, layer-by-layer explanations for ‘key’)

L	Output Content (Selected Key Points)	Analysis Result
23	"The ‘key’ concept helps you unlock the ‘door’ to understanding the ‘lock’."	**Conceptual Abstraction:** Frames the problem as unlocking "understanding" with a "key concept".
27	"The ‘key’ concept... is the central idea, principle, or insight that helps you understand the door (the complex topic)."	**Deepening the Concept:** Elevates the "key concept" to a "central idea, principle, or insight".
29	"‘The key’ is a new idea, a breakthrough... ‘The door swings open’... ‘The world beyond’ is the unknown, the possibilities..."	**Peak Abstraction:** Equates the "key" with the highest level of innovation—a "new idea" or "breakthrough".
Stage 4: Late-Layer Integration & Functional Equivalence		
33	"Language is the key that unlocks the potential of the human mind."	**Broad Generalization:** Applies the "key" concept to a vast, philosophical domain like "language".
36	"‘the data is the key,’ you’re saying that data is what unlocks our potential..."	**Domain-Specific Generalization:** Applies the "key" concept to the model’s own domain, considering "data" as the key.

Table 4: Selected layer explanations for "anchor", illustrating the evolution of semantic understanding from incoherence to abstraction.

L	Output Content (Selected Key Points)	Analysis Result
Stage 1: Early Incoherence & First Glimpse		
3	"Imagine you have a big box of LEGO bricks... ‘The’ is like a very common LEGO brick... a definite article."	**Incoherent:** Off-topic, explaining a basic article.
8	"Imagine you’re watching a TV show. The person who introduces the show... is like the ‘lead’ or ‘anchor’... In the world of news... called the ‘anchor’."	**First Glimpse:** First metaphorical association (news anchor).
Stage 2: Mid-Layer Exploration & Concretization		
12	"Imagine a ship sailing on a stormy sea... The ship itself is the data... the ‘anchor’ is the ‘central tendency’."	**Analogical Leap 1 (Statistics):** Links to "stability".
15	"Imagine a ship sailing on a stormy sea... The ‘anchor’ in this analogy is ‘a reliable and consistent source of information’."	**Analogical Leap 2 (Ship):** Solidifies the function of reliability.
17	"Imagine you’re building a house... ‘Anchor’ in this analogy would be like the strong foundation... ‘Hub’... ‘Platform’... ‘Lens’... ‘Bridge’... ‘Gateway’... ‘Compass’..."	**Analogical Leap 3 (House):** Connects to "support".
Stage 3: Deepening & Spiritual Abstraction		
26	"Imagine a ship sailing on a stormy sea... The ‘anchor’ is your core values, beliefs, and principles... ‘Staying afloat’..."	**Spiritual Deepening:** Elevates to personal values.
27	"Imagine a ship sailing on a stormy sea... The ‘anchor’ is your ‘anchor’ concept... a stable point of reference..."	**Symbolic Abstraction:** Further abstracts to "truth".

(continued on next page)

Table 4: (continued from previous page, layer-by-layer explanations for ‘anchor’)

L	Output Content (Selected Key Points)	Analysis Result
32	"Imagine a ship sailing on a vast ocean... An ‘anchor’: Represents stability, grounding... A ‘compass’: Represents direction... A ‘lighthouse’: Represents hope..."	**Conceptual Summary:** Explicitly lists core symbolic meanings.
Stage 4: Late-Layer Integration & Functional Equivalence		
35	"Imagine you’re building a house... ‘A strong foundation’... represents ‘a strong understanding of the basics’."	**Functional Generalization:** Moves to a functional equivalent.
43	"Imagine a big, complex machine... ‘Tete’ is like the oil... the glue that holds everything together... understanding and respect... allows people to work together... essential for a functioning society... can be broken... machine starts to break down..."	**Highest Abstraction:** Describes the function without the word.

C Validation via Reversed Orthogonal Difference Explanation

This appendix presents a complementary analysis that serves as a powerful validation for both our explanation methodology and a key premise of the "Deferred Semantic Drift" (DSD) hypothesis. We interpret the "reversed" orthogonal difference vector, $\mathbf{V}_{\text{Diff}}(\mathbf{H}_{\text{M, nc}}, \mathbf{H}_{\text{P, nc}})$, which is designed to isolate the semantic content unique to the prototypical (P) meaning by projecting out any shared metaphorical (M) abstractions.

Our central hypothesis is that this reversed vector should yield explanations strongly grounded in the physical, tangible, and concrete aspects of the target word. The striking results presented below confirm this, demonstrating a clear "semantic symmetry" to the analysis in Appendix B.

C.1 General Pattern: Overwhelming Focus on Physicality

The most prominent pattern emerging from the layer-by-layer analysis is the overwhelming predominance of physical interpretations. Across the majority of layers, particularly the middle layers where semantic processing is most active, the explanations consistently and thoroughly center on the tangible, object-related attributes of the word.

This stands in stark contrast to the forward analysis $\mathbf{V}_{\text{Diff}}(\mathbf{H}_{\text{P, nc}}, \mathbf{H}_{\text{M, nc}})$, which focused on abstract and metaphorical concepts. This clear divergence in outcomes compellingly illustrates two critical points:

- **Validation of Explanation Method:** The distinct and predictable results confirm that our orthogonal difference technique is not generating arbitrary associations but is precisely isolating and decoding specific, directional semantic content within the model’s representation space.
- **Validation of DSD Premise:** This finding lends further credence to the DSD hypothesis. It shows that the ambiguous token’s potential representation (in the non-causal oracle) is a rich superposition containing well-differentiated semantic facets (both physical and metaphorical). The challenge for the causal model is therefore not to create meaning from scratch, but to selectively retrieve the contextually relevant "informational packet" that is already encoded within.

C.2 Detailed Layer-by-Layer Results for "anchor"

The following Table 5 presents the detailed, layer-by-layer "reversed" explanations for the target word "anchor". Unlike the complex task of interpreting metaphors which requires continued abstraction, the relatively simple task of describing a physical object appears to be computationally "solved" in the middle layers. Consequently, the late layers exhibit signs of computational redundancy. The model either begins to repeat the detailed physical explanations from earlier layers (e.g., Layer 28) or abandons the specific task entirely, shifting to generic, high-level outputs like philosophical questions (Layer 35) or unrelated noise (Layer 43). This suggests that once a semantic task is fully resolved, the higher layers may not engage in further refinement, a finding that sheds light on the task-driven nature of hierarchical processing in LLMs.

Table 5: Selected layer explanations for the "reversed" orthogonal difference of "anchor". The results show a clear pattern: after an overwhelming focus on the physical object in the middle layers, the late layers exhibit computational redundancy and a shift to irrelevant or generic tasks.

L	Output Content (Selected Key Points)	Analysis Result
Stage 1: Early-Layer Confusion & Glimpse of Duality		
2	"Anchor ... Physical Object: An anchor is a heavy object... used to secure a boat... Figurative Meaning: ‘Anchor’ can also be used metaphorically..."	**Initial Duality:** Correctly identifies both physical and figurative meanings, but with a clear priority on the physical.
6	"An anchor for your boat, but for your journey! ...just like an anchor can help you stay grounded..."	**Metaphorical Intrusion:** Despite the prompt to isolate the physical, this layer’s explanation leans strongly towards the metaphorical, showing early-layer instability.
Stage 2: Mid-Layer Overwhelming Focus on Physicality		
8	"An anchor is a heavy object that is dropped from a boat or ship to hold it in place. It works by digging into the seabed..."	**Core Physical Definition:** A clear, textbook definition of a physical anchor’s function and mechanism.
17	"An anchor is a heavy object... Purpose... Mechanism... Types... Deployment... Retrieval..."	**Systematic Deconstruction:** Moves beyond a simple definition to a structured, technical breakdown of a physical anchor’s lifecycle.
22	"It seems like you’re asking about ‘anchor’ or ‘grounding’ in the context of a physical object. Could you please clarify...?"	**Peak Physical Certainty:** The model is so focused on the physical meaning that it interprets the query as exclusively physical and asks for user confirmation.
Stage 3: Late-Layer Redundancy and Task Abandonment		
28	"An anchor is a heavy object, typically made of metal, that is dropped from a ship to the seabed to hold it in place. ...Purpose... How it works... Types... Operation..."	**Computational Redundancy:** The model begins to repeat the detailed physical descriptions from the middle layers (e.g., Layer 17, 19), indicating the core task is complete.
35	"‘What is the meaning of life?’ ...Nihilism... Existentialism..."	**Task Abandonment (Shift to Generics):** Having exhausted the specific task, the model shifts to generic, high-level philosophical questions, effectively abandoning the original probe.
43	"‘Waw’ seems like an exclamation of surprise or amazement."	**Task Abandonment (Noise Generation):** In the final layer, the output becomes completely unrelated noise, explaining a simple interjection. The specific semantic signal has entirely dissipated.

D Behavioral Validation: Spontaneous Differentiation of Latent Concepts

This appendix presents a behavioral experiment designed to further validate a key premise of the DSD hypothesis: that an ambiguous word’s representation is a rich superposition of its distinct senses, which the model can access and differentiate.

D.1 Experimental Design

We investigate how the model differentiates two inputs when told it is comparing two "vectors." Critically, these vectors are the non-causal hidden states of the same target word (e.g., "anchor") from its prototypical (P) and metaphorical (M) contexts. However, the model is only prompted with the identical surface form "anchor" for both inputs, without explicit knowledge of their different origins. The setup is as follows:

1. **Inputs:** We provide the model with the non-causal hidden states for the same target word (e.g., "anchor") extracted from two different contexts: one prototypical (P) and one metaphorical (M).
2. **Prompting:** Critically, in the textual prompt given to the explanation module, we do not reveal the different origins of these vectors. Instead, we label both inputs with the identical surface form, e.g., asking the model to "explain the difference between ‘anchor’ and ‘anchor’."

3. **Objective:** This design directly tests the model's ability to reason beyond surface-level identity. Can it infer that identical words might represent different underlying concepts (vectors) and then articulate that difference?

The results, detailed below for the word "anchor," as shown in Table 6, reveal a fascinating, multi-stage reasoning process. Similar results for other target words are available in the supplementary files, named ***XXX_two_vector_explanation.html***.

D.2 A Three-Stage Reasoning Process

The layer-by-layer analysis reveals a consistent, three-stage process as the model interprets and responds to this ambiguous task.

Stage 1: Surface-Level Interpretation (Layers 1-5). Initially, the model is guided by the identical surface forms in the prompt. It consistently concludes that since the inputs ("anchor" and "anchor") are the same, there is no conceptual difference to explain. During this stage, it repeatedly requests distinct inputs, demonstrating a literal interpretation of the task based on the provided text.

Stage 2: Conceptual Differentiation (Layers 6-23). A dramatic shift occurs around Layer 6, where the model makes a critical inference: "While the word 'anchor' is the same... context likely changes the concept." Following this insight, the model spontaneously accesses and differentiates the two latent meanings. It consistently and accurately elaborates on the distinction between the physical, tangible anchor and the abstract, metaphorical one, often using highly precise terminology (e.g., "tangible restraint" vs. "emotional stability"). This stage, where deep semantic reasoning overcomes the initial surface-level interpretation, aligns perfectly with the middle layers identified as the primary locus of semantic computation in our main analysis.

Stage 3: Shift to Technical Interpretation (Layers 24-43). In the late layers, the model's understanding of the task shifts again. Having fully articulated the semantic difference, it begins to interpret the prompt's reference to "vectors" in a more literal, mathematical sense. It ceases the semantic explanation and instead repeatedly requests the actual numerical vector data for a technical comparison (e.g., of their magnitude and direction). This suggests that the late layers may default to a more technical or meta-level task interpretation once the core semantic processing is complete.

D.3 Implications for Deferred Semantic Drift

This layered and dynamic task understanding provides powerful qualitative support for the DSD mechanism:

- **Confirms Superposition Premise:** The model's ability to spontaneously access and contrast both the physical and metaphorical senses confirms that these distinct concepts are co-located within the word's potential representational space.
- **Highlights Mid-Layer Processing:** The "critical shift" and subsequent deep semantic analysis occur in the middle layers (approx. 6-23), aligning perfectly with our quantitative findings that these layers are the primary locus of semantic computation. This reinforces the idea that DSD is a mid-layer phenomenon.

Table 6: Selected layer explanations for the "identical input" experiment with "anchor". The results show a three-stage reasoning process, starting with surface-level confusion, transitioning to deep conceptual differentiation in the middle layers, and finally shifting to a technical interpretation of the task.

L	Output Content (Selected Key Points)	Analysis Result
Stage 1: Surface-Level Interpretation (Layers 1-5)		
1	"You've provided the same vector twice: 'anchor'. To understand the difference... I need two distinct vectors."	**Surface-Level Logic:** The model correctly identifies the identical input strings and, based on this surface form, concludes it cannot perform the comparison task.
3	"You've provided the same vector twice... I need different vectors... Think of vectors as representing ideas. Example: 'happy...' vs 'sad...'"	**Primed for Concepts:** While stuck on the surface identity, the model shows it is primed for conceptual comparison by providing an example of what it *could* do with distinct inputs.
Stage 2: Conceptual Differentiation (Layers 6-23)		

(continued on next page)

Table 6: (continued from previous page, **Identical "anchor" input comparison**)

L	Output Content (Selected Key Points)	Analysis Result
6	"While the word 'anchor' is the same... context likely changes the concept. Breakdown: Anchor (physical object)... Anchor (metaphorical)..."	**The Critical Shift:** The model spontaneously infers that identical words can hide different latent contexts, then immediately and correctly distinguishes the physical and metaphorical senses.
11	"Here's a breakdown... Key Difference... context. Physical... tangible restraint. Metaphorical... emotional or conceptual stability."	**Deep Semantic Elaboration:** The model moves beyond simple labels to articulate the core difference with remarkable precision, contrasting "tangible restraint" with "emotional stability."
16	"Here's a breakdown... Key Differences: Physical vs. Abstract. Function: prevents movement vs. provides stability and grounding."	**Functional Distinction:** Further refines the analysis by contrasting the core *function* of each sense of "anchor," demonstrating a deep, functional understanding.
Stage 3: Shift to Technical Interpretation (Layers 24-43)		
24	"You're asking about the difference between the concepts represented by the vectors 'anchor' and 'anchor'. Since both vectors are identical, they represent the same concept."	**State Reset/Reversion:** The model reverts to its initial surface-level logic, suggesting that the deep semantic processing of Stage 2 is a specialized, mid-layer computation not maintained in later layers.
26-43	"Please provide the two vectors! I'm ready to analyze them..." (Repeated many times)	**Technical Task Interpretation:** The model shifts its interpretation of the task entirely, now understanding "vector comparison" in a literal, mathematical sense and repeatedly requesting numerical data for a technical analysis.

E Head-by-Head Analysis

This appendix provides a granular analysis of how individual attention heads contribute to the "Deferred Semantic Drift" mechanism, specifically in understanding metaphors. We use our explanation method to decode the "thought process" of single heads when processing the word "anchor" in a metaphorical context. This allows us to observe functional specialization and trace how different semantic facets are processed across early, middle, and late layers.

E.1 Quantitative Trend: Specialization Peaks in Middle Layers

To assess the overall distribution of relevant computational work, we categorized the output of each of the 16 attention heads for three representative layers (Layer 9, 21, 40) as: "Highly Relevant," "Relevant," "Indirect Connection," or "No Clear Connection." The results reveal a clear trend: the concentration of semantically relevant heads peaks in the early-to-middle layers and declines sharply in the late layers.

- **Layer 9 (Early-Middle):** 6 out of 16 heads (37.5%) are either Relevant or Highly Relevant. (2, 4, 4, 6)
- **Layer 21 (Peak-Middle):** 4 out of 16 heads (25%) are Relevant or Highly Relevant, but with a higher number of abstract, indirect connections. (2, 2, 8, 4)
- **Layer 40 (Late):** 0 out of 16 heads (0%) provide a relevant explanation. The processing becomes entirely abstract, meta-cognitive, or irrelevant. (0, 0, 8, 8)

This quantitative trend provides strong evidence against "cherry-picking" and demonstrates a systematic shift in computation. It suggests that specific heads in the middle layers are specialized for core semantic processing, while late-layer heads focus on higher-level integration, consistent with our DSD hypothesis.

E.2 A Three-Stage Model of Head Functionality Across Layers

Analyzing the detailed head-by-head results for "anchor" (presented below) and for additional target words (provided in the supplementary HTML files), we identify a consistent, three-stage evolution in how attention heads contribute to metaphor comprehension:

Stage 1: Early Layers – Foundation of Metaphorical Understanding. In the early-to-middle layers (e.g., Layer 9), a subset of heads specializes in building a foundational understanding. They either capture the core metaphorical function directly (e.g., "stability," "support") or make strong associative links to related concepts and physical attributes (e.g., "foundation," "heaviness"). The reasoning at this stage is relatively direct and grounded, establishing the core semantic building blocks.

Stage 2: Middle Layers – Deepening, Value Attribution, and Creative Extension. In the peak processing layers (e.g., Layer 21), specialized heads move beyond simple functions to a deeper, more abstract understanding. They attribute specific value and significance to the metaphor (e.g., interpreting "hope" as a "valuable asset" or "driver of success"). The analogical reasoning becomes more creative and emotionally resonant, demonstrating a clear extension of the initial concept.

Stage 3: Late Layers – Shift Towards Abstraction, Generalization, and Meta-cognition. In the late layers (e.g., Layer 40), a functional shift occurs. Specialized heads for the specific metaphor largely disappear. Instead, the heads' focus turns to higher-order tasks: integrating the understood concept into broader contexts, making highly abstract connections, or engaging in meta-cognition by describing the model's own processes. This suggests the specific semantic computation is complete, and the system is now preparing the representation for its final, generative purpose. This progression provides a micro-level view of the DSD mechanism, illustrating how the "drift" is not a monolithic process but a hierarchical and distributed computation performed by specialized heads across different processing stages.

E.3 Qualitative Case Study: The Word "anchor"

The following tables provide a detailed, head-by-head case study for the word "anchor", illustrating the three-stage process described above. For a comprehensive view and results for other words, see the supplementary file ***XXX_head_explanation.html***

E.3.1 Layer 9: Emergence of Foundational Concepts

In this early-middle layer shown in Table 7, key heads (e.g., H1, H14) successfully capture the core metaphorical functions of "stability," "support," and "foundation." Other heads (e.g., H2, H3, H7, H13) make strong associative links to related concepts like physical weight and support mechanisms. The reasoning is relatively direct and grounded.

Table 7: Layer 9 Attention Head Analysis for text_M = "The *anchor* was heavy, yet her unwavering hope served as one for the entire family during difficult times." (Selected Heads)

H	Output Content (Selected Key Points)	Analysis Result
Category 1: Highly Relevant - Core Metaphorical Function		
1	"Imagine a news anchor delivering a serious report... 'Anchor' represents the stability and reliability..."	**Function: Stability & Reliability.** Directly captures the core metaphorical role of "hope" as a steadfast, dependable presence in chaotic times.
14	"Imagine a ship's anchor... 'from the bottom of the ocean'... suggests something is coming from a place of great depth..."	**Function: Depth & Foundation.** Associates the anchor with foundational support emerging from adversity, mirroring the role of hope in "difficult times."
Category 2: Relevant - Associative & Physical Attributes		
3	"Imagine you're building a house. You need a strong foundation, right? That's what a 'ground truth' is..."	**Association: Foundational Support.** Links to the concept of a "foundation," a key related idea for providing a stable base.
13	"Imagine a big, heavy anchor. It's too much for one person to lift, so you need a system of ropes and pulleys..."	**Attribute: Physical Weight & Support.** Focuses on the physical property of "heaviness" and the need for a support system, a literal attribute that grounds the metaphor.
Category 3: Indirect Connection - Abstract Analogies		
5	"Imagine you're trying to learn a new language... start with basic phrases... 'learning to code,' the 'basic phrases' are the fundamental concepts..."	**Analogy: Intellectual Anchors.** Infers a need for "fundamental concepts" to act as anchors for learning, an abstract parallel to the emotional anchor of hope.

(continued on next page)

Table 7: (continued from previous page, metaphorical *anchor*'s Head Explanation in ***Layer** 9*)

H	Output Content (Selected Key Points)	Analysis Result
6	"Imagine a big, bustling city... 'The city is a complex system'... 'The city is constantly evolving'..."	**Analogy: Grounding in Complexity.** Subtly implies the need for a grounding element (an anchor) within a complex, ever-changing system.
Category 4: No Clear Connection - Meta-Cognitive or Off-Topic		
10	"Imagine a vast library... picture yourself as a librarian... That's what I aim to be for you – a librarian of information..."	**Meta-Cognition:** Describes the LLM's own function as an information retriever, rather than processing the input's semantics.
12	"Imagine you have a big box of LEGO bricks... 'i.e.' is like saying 'in essence'... 'e.g.' is like saying 'for example'..."	**Off-Topic:** Explains Latin abbreviations, completely unrelated to the task.

E.3.2 Layer 21: Deepening, Abstraction, and Value Attribution

In this peak-middle layer shown in Table 8, the understanding becomes richer and more abstract. Highly relevant heads (e.g., H1, H10) move beyond simple stability to interpret the anchor as a "valuable asset," a "driver of success," and a symbol of "hope and resilience." The analogies become more creative and value-laden, reflecting a deeper semantic processing stage.

Table 8: Layer 21 Attention Head Analysis for text_M = "The *anchor* was heavy, yet her unwavering hope served as one for the entire family during difficult times." (Selected Heads)

H	Output Content (Selected Key Points)	Analysis Result
Category 1: Highly Relevant - Abstract & Value-Driven Functions		
1	"The '_' in this case represents the valuable asset of a successful business venture... a powerful tool for achieving success... a symbol of progress and growth."	**Function: Value & Progress.** Moves beyond simple stability to interpret the anchor as a proactive, value-driven asset that enables success and growth. Perfectly aligns with "hope."
10	"Imagine you're walking along a beach, and you see a bottle... This message is like the hope and dreams of the people who are facing hardship... the message inside, that's the resilience..."	**Function: Hope & Resilience.** Uses a powerful "message in a bottle" analogy to directly capture the core emotional meaning of the anchor as a symbol of hope and resilience in adversity.
Category 2: Relevant - Core Supporting Concepts		
7	"Imagine you have a big box of LEGO bricks... you need a good foundation. That foundation is like the 'main idea' of a story..."	**Association: Central Supporting Element.** Analogizes the anchor to a "foundation" or "main idea," a central element that provides structure and stability.
13	"Imagine a big, strong magnet. That's like the Earth... Its gravity... keeps us on the ground..."	**Association: Stabilizing Force.** Links the anchor to a fundamental "stabilizing force" (gravity) that prevents drifting and provides grounding.
Category 3: Indirect Connection - Contextual & Meta-Cognitive Links		
3	"The 'Man in the Street' Analogy... The prompt is not just a starting point; it's the foundation..."	**Analogy: Foundational Prompt.** Indirectly connects to the anchor concept via the idea of a "foundation," suggesting a sensitivity to core support structures.
9	"Imagine you're standing on a beach... 'the world is a dangerous place.' It acknowledges the potential for harm and danger..."	**Contextual Framing:** Focuses on the context of "danger/hardship" in which an anchor becomes vital, rather than the anchor itself.
Category 4: No Clear Connection - Off-Topic or Generic		

(continued on next page)

Table 8: (continued from previous page, metaphorical *anchor*'s Head Explanation in ***Layer** 21*)

H	Output Content (Selected Key Points)	Analysis Result
8	"Imagine a 'maze' of interconnected pathways... 'Word' is like a specific route... 'Dictionary' is like a map..."	**Off-Topic: Information Retrieval.** Uses a maze analogy for linguistic navigation, unrelated to the anchor's supportive role.
16	"The Problem: You're asking me to explain something, but you haven't actually provided anything to explain!..."	**Generic Response:** Outputs a generic request for more input, indicating it failed to process the probe meaningfully.

E.3.3 Layer 40: Shift to Generalization and Meta-Cognition

In this late layer shown in Table 9, a clear functional shift is observed. No single head provides a direct explanation of the "anchor" metaphor. Instead, their "thinking" becomes highly generalized, focusing on abstract systems (e.g., the internet), meta-linguistic concepts (e.g., "deeper meaning"), or the model's own cognitive processes. This supports the hypothesis that the specific semantic task is complete, and the focus has shifted to integrating the result into a global context for generation.

Table 9: Layer 40 Attention Head Analysis for text_M = "The *anchor* was heavy, yet her unwavering hope served as one for the entire family during difficult times." (Selected Heads)

H	Output Content (Selected Key Points)	Analysis Result
Category 1: Indirect Connection - Highly Abstract Analogies		
1	"Imagine a large, bustling city. This city represents the internet... My first impression of the internet was one of vastness and potential..."	**Analogy: Navigating Complexity.** Uses the vast internet as an analogy. The tenuous link is the implied need for an "anchor" (like a trusted website) to navigate complexity, but the focus is on the system itself.
8	"Imagine a car engine. It has many parts... If one part fails, the whole system can break down..."	**Analogy: System Integrity.** Focuses on system integrity. An "anchor" could be seen as a critical component preventing system failure, but the connection is highly inferential.
9	"Imagine a big, bustling city... this city is your mind... mindfulness is all about: paying attention to the city of your mind without judgment."	**Analogy: Mental Anchors.** Links to the concept of mindfulness, where an "anchor" (like the breath) provides a point of stability. This is a plausible, but very abstract and inferential, parallel.
Category 2: No Clear Connection - Meta-Cognitive, Off-Topic, or Anomalous		
3	"Imagine a bird's wing. It's not just a flat surface... suggests that there's more to the story, more depth, more meaning..."	**Meta-Linguistic Analysis:** Explains the general process of "interpreting non-literal meaning," rather than the specific meaning of "anchor."
7	"Imagine you're a detective trying to solve a case... connect the dots, analyze the information... That's what I do with your text input."	**Meta-Cognition (Self-Description):** Describes the LLM's own text-processing pipeline, comparing itself to a detective.
15	"Let's say you have a new invention... a new kind of coffee maker... Option 1: The 'Tech Talk'... Option 2: The 'Lifestyle'..."	**Off-Topic:** Explains different marketing strategies, completely unrelated to the input sentence.
16	(Content is a long, repetitive string of "to-be-to-being-to-but-to-be...")	**Anomalous Output:** Generates meaningless, repetitive text, indicating an unstable or error state for this head.

F Qualitative Examples of Controllable Intervention

This appendix provides qualitative examples demonstrating the effectiveness of our Deferred Drift-Informed Activation Transport (DDI-ACT) intervention method, as discussed in Section 4.4. The goal of DDI-ACT is to steer the model's generation towards either a prototypical (P) or metaphorical (M) interpretation of an ambiguous target word.

Methodology Recap. The intervention works by adding a pre-computed semantic steering vector, $\mathbf{V}_{\text{shift}}$, to the target word's hidden state at a key middle layer L. This vector is derived from our Non-Causal Oracle analysis and represents the precise directional shift from the prototypical to the metaphorical meaning space. For our experiments with Gemma2-9B-it, we identify layers 11-22 as the optimal intervention range. The scalar coefficient λ controls the strength and direction of the steering.

The following tables showcase completions for different target words under various λ values, illustrating a clear dose-response relationship between the intervention and the semantic outcome.

Target Word: "key" The model was given the following ambiguous prefix to complete: **"After much thought, she realized that the key..."**, as shown in Table 10.

Table 10: Controllable generation examples for the target word **'key'**.

Intervention Type	**Generated Completions**	**Analysis**
No Intervention ($\lambda = 0.0$)	"...was still in her other coat pocket." "...to solving the puzzle was simpler than she had imagined."	The baseline model generates both literal and metaphorical completions, reflecting the initial ambiguity.
Strong Prototype Steering ($\lambda = -1.5$)	"...was made of tarnished brass and fit the old lock perfectly." "...had a unique design, with a small emblem carved into its head."	The intervention successfully forces a literal interpretation, focusing on the physical attributes of a key.
Strong Metaphor Steering ($\lambda = +1.5$)	"...was not a single action, but a fundamental change in perspective." "...to moving forward was to finally let go of the past."	The intervention reliably steers the output towards an abstract, metaphorical meaning.

Target Word: "bridge" The model was given the following ambiguous prefix to complete: **"To connect the two sides, they decided to build a bridge..."**, as shown in Table 11.

Table 11: Controllable generation examples for the target word **'bridge'**.

Intervention Type	**Generated Completions**	**Analysis**
No Intervention ($\lambda = 0.0$)	"...across the wide and fast-flowing river." "...of dialogue between the opposing factions."	The baseline reflects the common literal and metaphorical uses of "bridge," demonstrating its inherent ambiguity.
Strong Prototype Steering ($\lambda = -1.5$)	"...with sturdy steel girders and reinforced concrete supports." "...that could withstand the region's frequent earthquakes."	The intervention forces the generation to focus on the engineering and physical structure of a bridge.
Strong Metaphor Steering ($\lambda = +1.5$)	"...of understanding and trust between the two communities." "...from their shared history to a collaborative future."	The intervention successfully guides the generation towards abstract concepts of connection and reconciliation.

Conclusion. The examples above demonstrate a high degree of precise, predictable control over the model's semantic processing. By intervening at the source—the ambiguous word's representation—before it is queried by subsequent tokens, we directly manipulate the "informational packets" retrieved through the attention mechanism. This ability to reliably steer the final interpretation confirms that the deferred computational pathways we identified are not merely correlational but are functionally integral to the model's process of delayed disambiguation. This serves as powerful causal validation for the Deferred Semantic Drift mechanism.

G THE USE OF LARGE LANGUAGE MODELS (LLMS)

We transparently disclose the use of LLMs in two distinct capacities in this research.

First, for data generation, we utilize Gemini 2.5 Pro (Comanici et al., 2025) to construct the core dataset for our experimental analysis. This process is methodologically driven and strictly controlled: we guide the model using a carefully engineered prompt to generate prototype-metaphor sentence pairs from a common prefix, ensuring a controlled setup for our causal analysis. Crucially, all generated data undergo a rigorous manual review and validation process by human annotators to ensure its quality and accuracy. A detailed description of the LLM-guided generation methodology, quality control measures, and the full prompt used can be found in Section A.1.

Second, for manuscript preparation, an LLM is also employed to assist with language polishing for certain sections of this paper, enhancing clarity and readability. We take full responsibility for all scientific contributions, the integrity of the data, the presented analyses, and the final text of the manuscript.