Tug-of-war between idioms' figurative and literal interpretations in LLMs

Soyoung Oh[1], Xinting Huang[1], Mathis Pink[2], and Michael Hahn[1, †], Vera Demberg[1,3, †]
Saarland University[1], Max Planck Institute for Software Systems[2],
Max Planck Institute for Informatics[3]
{soyoung, xhuang, mhahn, vera}@lst.uni-saarland.de,
mpink@mpi-sws.org

Abstract

Idioms present a unique challenge for language models due to their non-compositional figurative interpretations, which often strongly diverge from the idiom's literal interpretation. In this paper, we employ causal tracing to systematically analyze how pretrained causal transformers deal with this ambiguity. We localize three mechanisms: (i) Early sublayers and specific attention heads retrieve an idiom's figurative interpretation, while suppressing its literal interpretation. (ii) When disambiguating context precedes the idiom, the model leverages it from the earliest layer and later layers refine the interpretation if the context conflicts with the retrieved interpretation. (iii) Then, selective, competing pathways carry both interpretations: an intermediate pathway prioritizes the figurative interpretation and a parallel direct route favors the literal interpretation, ensuring that both readings remain available. Our findings provide mechanistic evidence for idiom comprehension in autoregressive transformers[1].

1 Introduction

Idioms pose a challenge to standard semantic composition because, as multi-word expressions, their figurative meanings don't follow from the literal meanings (Beck and Weber, 2020). For instance, "kick the bucket" conveys the figurative interpretation "to die", yet its literal sense "physically kicking a container" remains semantically available. Psycholinguistics has investigated how humans understand idioms, whether they access the figurative interpretation directly or first have to access the literal interpretation of the idiom and only in a later stage suppress that interpretation and access the figurative interpretation. These findings have led to various models describing human idiom processing (Gibbs, 1980; Cacciari and Glucksberg, 1991; Bobrow and Bell, 1973; Cacciari and Tabossi, 1988).

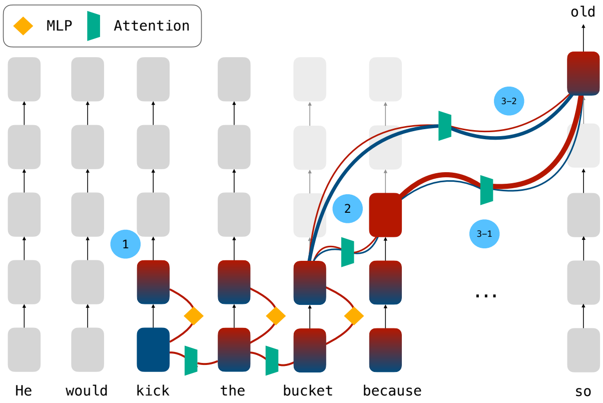


Figure 1: The **figurative** and **literal** interpretations are highlighted in the blocks and paths. We find three main steps for idiom processing: 1 **Idiom retrieval step:** Early layers (i.e., layers 0-3) attention and MLP are actively retrieving the idiom's figurative interpretation while storing both figurative and literal interpretations in the residual stream. 2 **Selective interpretation step:** At the token immediately following the idiom span, the model begins to encode a representation that favors the figurative interpretation over the literal one, starting from the middle layers. 3 **Interpretation routing:** For final prediction, the model passes literal interpretation via both **a direct compositional semantic path** (3-2), as well as the intermediate pathway that prioritizes the figurative interpretation (3-1 **figurative path**).

Yet for LLMs, it's unclear how they process idioms, potentially limiting further improvements in their idiom comprehension (Kabra et al., 2023; Liu et al., 2024; Knietaite et al., 2024). One explanation may lie in how the model transforms raw input token embeddings into richer representations (Tian et al., 2023; Haviv et al., 2023; Dankers et al., 2022). Within the autoregressive transformer, token representations are incrementally refined through a series of residual layers, which consist of multi-head self-attention (MHSA) and MLP sublayers, each adding their outputs to update the representation (Vig et al., 2020). This multi-stage

[†] Co-corresponding authors.

[1] Code and dataset are available at https://github.com/sori424/idiom_processing

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2942–2958
March 24-29, 2026 ©2026 Association for Computational Linguistics

transformation suggests that idiom processing proceeds via a sequential shift from compositional to non-compositional representations (Haviv et al., 2023; Dankers et al., 2022).

In this paper, we first aim to identify how the idiom is interpreted by a transformer Llama3.2-1B (we also run experiments on Llama3.1-8B, Qwen2.5-0.5B, Qwen2.5-7B), where its figurative interpretation is located, and how it is retrieved by the model compared to its literal counterpart. To this end, we pinpoint specific components of the model that have specialized in processing idioms by boosting their figurative interpretation, while suppressing the compositional semantic literal interpretation. Having identified these, we proceed to investigating the information flow in the model, with the goal of identifying specific pathways through which the figurative interpretation is passed forward to the final prediction, and whether this differs from the pathways of passing forward information about an idiom's literal interpretation. Additionally, given that the context plays a key role in disambiguating the interpretation of an idiom (Mi et al., 2025; Fazly et al., 2009; Dankers et al., 2022; Holsinger, 2013), we investigate how context interacts with lexical information of idioms.

We address these questions through knockout analyses (Nanda et al., 2023; Wang et al., 2022; Geva et al., 2023), where we separately ablate activations of each component (layer-wise MHSA and MLP, individual attention heads) in the model to observe their importance for retrieving idiom's figurative interpretation, and contextual disambiguation. Furthermore, to trace the flow of these interpretations, we employ activation patching experiments (Wang et al., 2022; Meng et al., 2022; Hanna et al., 2023; Conmy et al., 2023; Stolfo et al., 2023), in which we ablate one or both interpretations from specific pathways within the model. Figure 1 summarizes our main findings. Our contributions are:

- We find that early layers and idiom specific heads causally retrieve figurative interpretations, while suppressing the corresponding literal counterparts (Section 4).
- We identify a selection mechanism at the token immediately following the idiom, where an intermediate path carries a strong flow toward the figurative interpretation. In parallel, a bypass pathway preferentially carries literal information directly to the final prediction, circumventing the intermediate pathway (Section 5).
- We localize where context resolves idiomatic ambiguity, showing that early MLP/MHSA layers suppress figurative retrieval under literal context, while later MHSA integrates surrounding tokens to refine the context consistent interpretation (Section 6).

2 Idiom interpretation task

We conduct our experiments in two setups. First, we investigate how language models process idiomatic expressions, which permit both figurative and literal interpretations. In the second setup, we introduce disambiguating context by adding preceding sentences.

We formalize a next-token prediction task and construct a dataset. Idioms are embedded into a sentence template designed to reveal their interpretation through continuation '`X (would) IDIOM because X was so/too/a/the`'[2]. This template preserves idiom ambiguity while eliciting causal completions (e.g., for 'kick the bucket', figurative completions include 'old/sick', while literal completions include 'angry/mad'.) Since there is no single ground truth continuation, we automatically construct token sets consistent with either the figurative or literal interpretation. The tokens in each set are mutually exclusive and chosen for their high likelihood of association with the respective interpretation. Below, we outline the dataset construction process, and see Figure 9.

1. Idiom Extraction: We select 245 idioms from a psycholinguistic paper (Cronk et al., 1993), each of which received high literality (humans rated the literal meaning as plausible). This ensures ambiguity, which is necessary to evaluate the model's change between literal and figurative interpretations.

2. Sentence Generation: We generate a literal paraphrase and a figurative paraphrase for each idiom using the Llama3.3-70B-instruct model (Grattafiori et al., 2024). The sentences below are three examples of the generated *ambiguous sentence* (s_a), *figurative paraphrase* (s_f), and *literal paraphrase* (s_l):

[2] Here, `X` is instantiated with pronouns and morphologically adapted; X is instantiated with various pronouns (he/she/it/they); the word "would" is inserted only in some of the idioms to make the sentence sound more fluent; idioms are morphologically fit into the sentence; depending on sg/pl form of X, the second part of the template contained was or were.

s_a He would *kick the bucket* because he was so

s_f He would *die* because he was so

s_l He would *kick the container* because he was so

3. Token Set Generation: To generate sets of next-word completions that are indicative of the figurative vs. the literal interpretation, we prompt the Llama3.3-70B-instruct model with paraphrases (s_f, s_l) and obtain logit distributions. Let $z^{(f)}$ and $z^{(l)}$ denote the logits corresponding to s_f and s_l as input, then we compute the element-wise difference between the two sets $\Delta z_v = z_v^{(f)} - z_v^{(l)}$, where $v \in V$ is the vocabulary size of the model. Since tokens that score high in one paraphrase tend to score low in the other, Δz_v captures each token's interpretation-specific relevance. We take the 20 tokens with the largest Δz_k as the figurative candidate set C_f, and the 20 with the smallest Δz_k as the literal candidate set C_l.

4. Context Generation: Using each paraphrased sentence, we generate a preceding context sentence that disambiguates the interpretation of the subsequent idiom. The examples below illustrate the generated figurative context (FC) and literal context (LC).

FC His breath rattled after every step. He would die because he was so

LC The vending machine ate his coins. He would kick the container because he was so

3 Methodology

3.1 Information flow in transformers

Given an input token sequence $\mathbf{t} = [t_1, \ldots, t_N]$ over a vocabulary V, each token is embedded as $\mathbf{x}_i^0 = \mathrm{PE}(\mathrm{emb}(t_i), i) \in \mathbb{R}^d$, where PE denotes a positional encoding function that injects position i into the token embedding (Vaswani et al., 2017), and d is the model's hidden dimension. From layer 0, these vectors x_i^0 are carried forward and accumulated over the subsequent L layers. At each layer ℓ, the hidden state for token i is updated from its previous value $\mathrm{x}_i^{\ell-1}$ by adding the multi-head self-attention (MHSA) output a_i^ℓ and MLP output m_i^ℓ:

$$\mathrm{x}_i^\ell = \mathrm{x}_i^{\ell-1} + a_i^\ell + m_i^\ell.$$

MLP Sublayers Every MLP sublayer computes a local update for each representation:

$$m_i^\ell = W_F^\ell \, \sigma\Big(W_I^\ell (a_i^\ell + \mathrm{x}_i^{\ell-1})\Big),$$

where $W_I^\ell \in \mathbb{R}^{d_{ff} \times d}$ and $W_F^\ell \in \mathbb{R}^{d \times d_{ff}}$ are parameter matrices with inner-dimension d_{ff} and σ is a nonlinear activation function.

MHSA Sublayers Each attention layer's output a_i^ℓ aggregates information of multiple parallel attention heads h. The parameter matrices of each MHSA sublayer comprise three projection matrices W_Q^ℓ, W_K^ℓ, $W_V^\ell \in \mathbb{R}^{d \times d}$, and $W_O^\ell \in \mathbb{R}^{d \times d}$. The columns of each projection matrix and rows of the outputs matrix can be split into H equal parts, corresponding to the number of attention heads $W_Q^{\ell,j}$, $W_K^{\ell,j}$, $W_V^{\ell,j} \in \mathbb{R}^{d \times \frac{d}{H}}$, and $W_O^{\ell,j} \in \mathbb{R}^{\frac{d}{H} \times d}$, $j = [1, H]$. Let $X^{\ell-1} \in \mathbb{R}^{N \times d}$ be the matrix of all token representations at layer $\ell - 1$, $A^{\ell,j} \in \mathbb{R}^{N \times N}$ encodes the weights computed by the j-th attention head at layer ℓ, where it's masked to a lower triangular matrix. The MHSA output is the sum over individual attention heads matrices:

$$a_i^\ell = \sum_{j=1}^{H} \big(A_i^{\ell,j} X^{\ell-1} W_V^{\ell,j}\big) W_O^{\ell,j}.$$

After L layers, the final hidden state x_T^L is mapped to vocabulary logits via an unembedding matrix $W_U \in \mathbb{R}^{|V| \times d}$, so that $y = W_U \, \mathrm{x}_T^L \in \mathbb{R}^{|V|}$, and $\mathrm{softmax}(y)$ defines the next-token distribution. Thus we can view a transformer as a computational graph $G \colon \mathcal{X} \to \mathcal{Y}$, where token embeddings, MHSA sublayers, MLP sublayers, and the unembedding all communicate via the residual stream. We refer to Vaswani et al. (2017); Geva et al. (2023) for further details.

3.2 Knockout

If we think of a model as a computational graph G, then certain behaviors of the model can be attributed to specific subgraphs $G_{sub} \subseteq G$ (Rai et al., 2024). Identifying such subgraphs provides a valuable mechanism for interpretability, enabling the isolation of individual components and their contributions to the model's behavior. A knockout removes specific components to quantify their contribution to particular behaviors (Nanda et al., 2023; Wang et al., 2022; Geva et al., 2023). Every activation in G_{sub} can be replaced by either zero or reference mean:

$$\mathrm{x}_i^\ell \;\longmapsto\; \tilde{\mathrm{x}}_i^\ell = \begin{cases} 0 \quad \text{or} \quad \mu_i^\ell & \text{if } (\ell, i) \in G_{\text{sub}}, \\ \mathrm{x}_i^\ell & \text{otherwise}. \end{cases},$$

then G with those components "knocked out" is

$$G_{\text{knock}} = G\left(\ldots, \tilde{\mathrm{x}}_i^{\ell}, \ldots\right).$$

By replacing the activations of these components, we eliminate the contributions of G_{sub} while keeping other computations in G fixed.

3.3 Activation patching

Activation patching is a technique to identify which activations in a model contribute to a particular output (Wang et al., 2022; Meng et al., 2022; Hanna et al., 2023; Conmy et al., 2023; Stolfo et al., 2023). This involves running the model on input A, then selectively patching certain activations with those obtained from a paired input B:

$$G_{\text{patch}}(A \hookleftarrow B) = G\big(\mathrm{x}_i^1(A), \ldots, \mathrm{x}_i^{\ell}(B), \mathrm{x}_i^{\ell+1}(A), \ldots\big).$$

By comparing the original output for input A with the altered output after patching, we can precisely measure how much the introduced activations from input B shift the model's output.

3.4 Metric for measuring interpretation shift

To measure whether the model assumes the figurative or literal interpretation when given each prompt $s \in \{s_a, s_f, s_l\}$, we define two scores based on cumulative probabilities:

$$F(s) = \sum_{c_f \in C_f} p(c_f \mid s); \; L(s) = \sum_{c_l \in C_l} p(c_l \mid s).$$

After applying an intervention, we quantify its causal effect:

$$\Delta I(s) = I_{\text{interv}}(s) - I_{\text{origin}}(s), \; I \in \{F, L\},$$

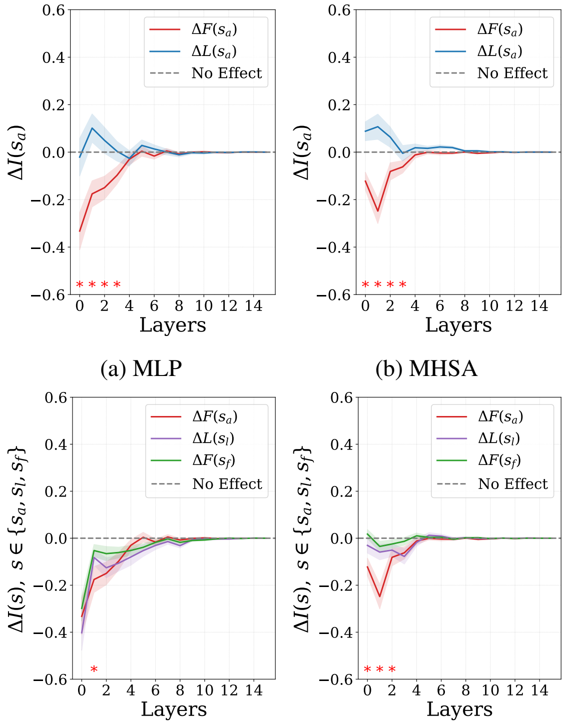
where $\Delta I < 0$ means the intervention disrupts the corresponding interpretation, implying the component is necessary; $\Delta I = 0$ shows no effect; $\Delta I > 0$ implies the component is not required (or may even inhibit) that particular reading.

4 Localizing idiom's figurative interpretation retrieval

We deliberately preserve idiomatic ambiguity by analyzing idioms without the context, so we can localize where, and through which components, the model shifts between figurative and literal interpretations.

4.1 Probing sublayers through knockout

To locate the mechanisms for retrieving figurative interpretations, we knock out the activations at idiom span tokens by replacing them with their mean values computed over all s_a. By knocking out the activations across each layer, we assess how the probabilities of predicting specific candidate tokens ($\Delta F(s_a)$, $\Delta L(s_a)$) change. To narrow down the points to which we can ascribe a functional role for a specific computation component, we limit the intervention to a MLP and MHSA in each block.



(a) MLP

(b) MHSA

(c) MLP

(d) MHSA

Figure 2: Sublayer-wise interpretation shift $\Delta I(s)$ after ablating activations at idiom span, for sentences $s \in \{s_a, s_f, s_l\}$. **Y-axis:** Mean values of $\Delta L(s_a)$, $\Delta F(s_a)$, $\Delta L(s_l)$, $\Delta F(s_f)$ with 95% confidence intervals across idiom sentences. **X-axis:** Layers. **Gray dashed line:** $\Delta I = 0$ (no effect). The red asterisk (*) marks layers where the difference between ΔL and ΔF exceeds the average difference across layers (paired t-test, $p < 0.05$).

Results Figure 2a shows by how much the model's interpretation of the idiom changes when the MLP sublayer is knocked out. We observe substantial drops of $\Delta F(s_a)$ when knocking out early MLP sublayers (0-2). At the same time, we see an increase in the probability of the literal completions: $\Delta L(s_a)$ rises above zero. And these shifts are significantly different in 0-3 layers ($p < 0.05$). This indicates that early layer MLPs are causally important for reading out the figurative interpretation while potentially suppressing the literal one. For later layers (≥ 4), the knockout does not affect the model's interpretation, where the corresponding patching effect converges to zero.

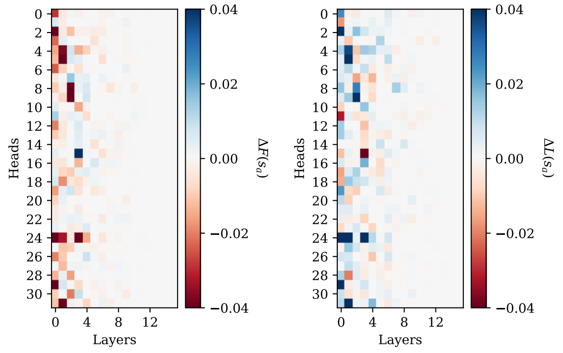
Figure 2b shows the effect of MHSA knockout, which is similar to what we observed with MLP sublayer knockout. Specifically, knocking out early layers MHSA (0-3) leads to a drop in $\Delta F(s_a)$ and increase in $\Delta L(s_a)$. The most significant difference between $\Delta F(s_a)$ and $\Delta L(s_a)$ is in layer 1, where $\Delta F(s_a)$ dropped by -0.30 and $\Delta L(s_a)$ increased by 0.16 ($p < 0.05$). After layer 4, knocking out MHSA has little to no effect on interpretation, as the values converge toward zero.

Together, the two interventions track the 1 **idiom's figurative interpretation retrieval**: Early MHSA layers (0-3) gather and bind token-wise interactions, while early MLP layers (0-3) transform those bound representations into a figurative interpretation, while suppressing the literal one. This early layer idiom interpretation retrieval aligns with the findings of Haviv et al. (2023), who argue that idiomatic information is accessed in the initial layers of LLMs during inference. After layer 4, $\Delta F(s_a)$ and $\Delta L(s_a)$ stabilize to zero, indicating that later sublayers are not adding new information and that the representations of the interpretations at the idiom span tokens are no longer used by any subsequent tokens.

Moreover, these components are specifically essential for idiom processing, not generally important for semantic interpretation of a sentence. Figure 2c shows that the drop from knocking out early MLP sublayers is significantly smaller in layer 1 for the two paraphrases ($\Delta L(s_l)$, $\Delta F(s_f)$) than for the idiomatic expression, ($\Delta F(s_a)$), ($p < 0.05$). While, at layer 0, the knockout effect is large for all expressions, as in this case, the MLP block is processing the semantics of the sentence. Figure 2d shows a similar effect for the early MHSA layers (0-2) where literal paraphrases are not strongly affected by knockout ($\Delta L(s_l)$, $\Delta F(s_f)$) compared to idiomatic expression ($\Delta F(s_a)$), ($p < 0.05$). The crucial role of figurative interpretation retrieval in early layers is consistent across different models (Appendix D).

4.2 Identifying attention heads specific to retrieving figurative interpretations

Next, we investigated which attention heads are specialized for retrieving an idiom's figurative interpretation. We knock out attention heads iteratively at the idiom span tokens and measure $\Delta F(s_a)$, $\Delta L(s_a)$. To operationalize this, we define *idiomatic heads* ($\mathcal{H}_{\text{idiom}}$) as heads with a large negative $\Delta F(s_a)$ and positive $\Delta L(s_a)$.



(a) $\Delta F(s_a)$

(b) $\Delta L(s_a)$

Figure 3: Heatmaps of the (a) $\Delta F(s_a)$ (b) $\Delta L(s_a)$ when ablating individual attention heads at the idiom span. **Idiomatic heads**: Heads those are crucial for retrieving the figurative interpretation of idiom; $-\Delta F(s_a)$ and $+\Delta L(s_a)$, simultaneously.

We identify attention heads whose removal reliably decreases figurative probability ($-\Delta F(s_a)$), while increasing literal probability ($+\Delta L(s_a)$) for the idiom sentences. To identify which of these heads are consistently important across instances of s_a, we performed a nonparametric bootstrap with $B = 1000$ resamples for each head's contribution to the interpretation. We then rank all heads by, on the one hand, $-\Delta F(s_a)$, and, on the other hand, $+\Delta L(s_a)$, and select the top 20 heads highly ranked on both orders (selected attention heads listed in Appendix Table 1). Figure 3 shows that there is a subset of attention heads that simultaneously exhibits large $-\Delta F(s_a)$ (Figure 3a) and $+\Delta L(s_a)$ (Figure 3b).

4.3 Causal role of MLP sublayers and idiomatic heads in figurative reading

To show that the components that we found (i.e., early MLP sublayers and $\mathcal{H}_{\text{idiom}}$) are causally sufficient for figurative reading, we patch the activations on s_a at idiom span tokens into the corresponding components when processing s_l: patching their activations from the idiom sentence into the literal paraphrase increases $\Delta F(s_l \hookleftarrow s_a)$ and decreases

$\Delta L(s_l \hookleftarrow s_a)$, while patching random components has no effect.

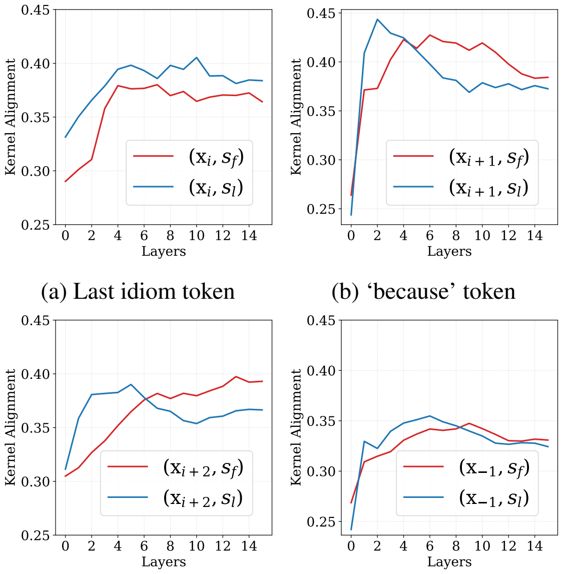
Results Patching these components boosts the figurative interpretation, as quantified by $\Delta F(s_l \hookleftarrow s_a)$ ($M = 0.17$, $SD = 0.49$) while suppressing the literal interpretation, $\Delta L(s_l \hookleftarrow s_a)$ ($M = -0.29$, $SD = 0.58$). In contrast, patching random components barely affects the figurative interpretation ($\Delta F(s_l \hookleftarrow s_a)$: $M = -0.008$, $SD = 0.09$) as well as literal one ($\Delta L(s_l \hookleftarrow s_a)$: $M = -0.005$, $SD = 0.11$). Effects on the experiment and the control are significantly different ($p < 0.05$). These contrasting effects suggest that idiomatic components are specialized to enhancing the figurative interpretation while suppressing literal interpretations, so that replacing their activations injects a figurative interpretation and amplifies $\Delta F(s_l)$, while dropping $\Delta L(s_l)$.

5 Tracing the interpretation flow

5.1 Locating idiom interpretation across token positions

After the retrieval process, the model encodes both literal and figurative interpretations in its residual stream. We next investigate how these two representations are weighted or integrated in the final prediction. To locate the representation of the interpretations, we calculate the mutual nearest-neighbor kernel alignment (Huh et al., 2024; Cho et al., 2025) between sentence embeddings of a corresponding idiom token span from s_f and s_l encoded by BGE M3 (Chen et al., 2024) and hidden states from various token positions of idiom sentence s_a. This metric quantifies the similarity between representations, with higher values indicating greater semantic similarity.

Results We observe that at the 'because' token (Figure 4b), the kernel alignment between the idiom's hidden states and its figurative interpretation surpasses that of the literal interpretation after layer 4. This suggests that by this point, the model has begun to favor the figurative interpretation (2 **selective interpretation step**). Based on the results, we hypothesize an intermediate pathway where token position 'because' plays a pivotal role in steering the idiom toward a figurative interpretation. In parallel, to keep the alternative interpretation available at the prediction (as in Figure 4d), we posit a direct route that favors the literal interpretation by more strongly preserving compositional semantic information.



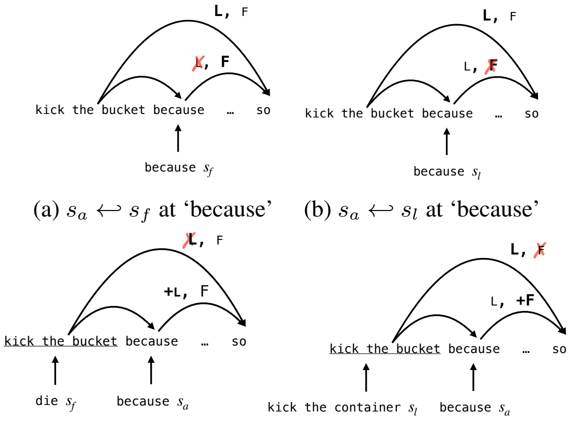
(a) Last idiom token (b) 'because' token

(c) After 'because' token (d) Last token of sentence

Figure 4: Kernel alignment between hidden states (x) extracted from four different token positions of s_a and semantic embeddings of paraphrases (s_f, s_l).

5.2 Analyzing layer-wise competing interpretation flow through activation patching

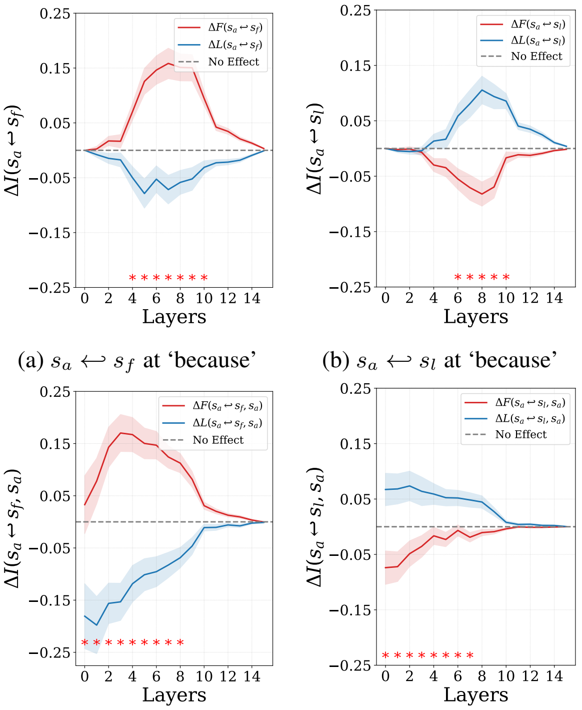
To measure the information flow along the intermediate pathway (through the subsequent token position *'because'*), we patch the activation s_a at the 'because' token with s_f or s_l (Figures 5a, 5b). On the other hand, patching the activation s_a at the idiom token span with an alternative activation (s_f or s_l) and then re-patching attention at the 'because' token with s_a allows us to isolate the information that flows via the direct route (Figures 5c, 5d).



(a) $s_a \hookleftarrow s_f$ at 'because' (b) $s_a \hookleftarrow s_l$ at 'because'

(c) $s_a \hookleftarrow s_f$ at idiom span, $s_a \hookleftarrow s_a$ at 'because' (d) $s_a \hookleftarrow s_l$ at idiom span, $s_a \hookleftarrow s_a$ at 'because'

Figure 5: Conceptual description of the activation patching experiments for tracing information flow (**L** = literal interpretation; **F** = figurative interpretation).



(a) $s_a \hookleftarrow s_f$ at 'because'

(b) $s_a \hookleftarrow s_l$ at 'because'

(c) $s_a \hookleftarrow s_f$ at idiom span, $s_a \hookleftarrow s_a$ at 'because'

(d) $s_a \hookleftarrow s_l$ at idiom span, $s_a \hookleftarrow s_a$ at 'because'

Figure 6: Layer-wise interpretation shift after patching in activations. The red asterisk (*) marks layers where the difference between ΔL and ΔF exceeds the average difference across layers (paired t-test, $p < 0.05$).

Results Figure 6a shows that the patching as illustrated in Figure 5a, increases the $\Delta F(s_a \hookleftarrow s_f)$ in mid-layers 4-10 (max: 0.17), but has a relatively small effect on $\Delta L(s_a \hookleftarrow s_f)$, which shows some decrease (min: -0.09). Patching in s_l (as in Fig. 5b) removes the intermediate path figurative flow, which leads to an increase in the $\Delta L(s_a \hookleftarrow s_l)$ in mid-layers (6-10) as shown in Figure 6b (max: 0.13). At the same time, $\Delta F(s_a \hookleftarrow s_l)$ drops symmetrically in the same layers (min: -0.10). This indicates that the figurative interpretation is suppressed in favor of the literal interpretation.

The asymmetry between increases and decreases show that a figurative injection doesn't lead to suppression of the literal interpretation. By contrast, a literal injection must first suppress the figurative interpretation in that subspace to have an effect, producing a more symmetric push–pull tradeoff. This indicates that for the mid-layers at the 'because' token position, the model's residual stream is the predominant route for the figurative interpretation (3-1 **figurative path**).

Meanwhile, Figure 4d indicates that literal interpretation remains competitive with the figurative one. We assume that there's a different path that favors the semantically literal interpretation (3-2 **compositional semantic direct path**) to convey its information to the final prediction, directly bypassing the intermediate path.

To probe this, we replace the idiom sentence's idiom token span activation with the activation from a paraphrased sentence, and re-patch the 'because' token with the idiom sentence's original activation (as shown in Figures 5c, 5d). This preserves the intermediate pathway and isolates the contribution of the direct route. Figure 6c shows that removing direct path's literal interpretation produces a large drop in ΔL (min: -0.25) and rise in ΔF (max: 0.20). Conversely, removing the direct path's figurative interpretation yields only a modest drop in ΔF (min: -0.10) and a small increase in ΔL (max: 0.10). These results indicate that the direct path predominantly conveys literal information. This pattern is consistent across different models (Appendix E).

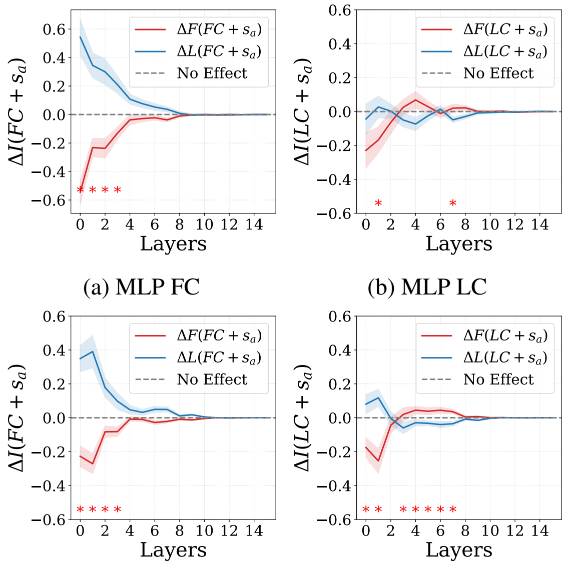
6 Localizing idiom disambiguation in context

To assess the role of context in idiom disambiguation, we're now looking at the setup with context. We retain an instance only when the model assigns higher cumulative probability to the context-consistent candidate than to the alternative (i.e., $F(s_a|FC) > F(s_a|LC)$, $L(s_a|LC) > L(s_a|FC)$).

6.1 Probing sublayers through knockout with figurative vs. literal context

We conduct a layer-wise knockout experiment to assess how MLP and MHSA components contribute to idiom processing when the same idiomatic expressions are presented in different contexts.

Results Figure 7a shows that for idioms in a figurative context, knocking out the MLP in the early layers (0–3) results in a significant drop in ΔF (min: -0.62) and increase in ΔL (max: 0.66), consistent with the pattern observed for idioms without context (Figure 2a). This indicates that these layers are critical for retrieving the figurative interpretation independently of context. This result is consistent across different models (Appendix F). For idioms in a literal context (7b), ablating the MLP layer 1 leads to a significant drop in ΔF (min: -0.33) and an increase in ΔL (max: 0.09), indicating that the figurative interpretation is still being



(a) MLP FC

(b) MLP LC

(c) MHSA FC

(d) MHSA LC

Figure 7: Sublayer-wise interpretation shift ΔI after knockout activations at idiom span with contexts. The red asterisk (*) marks layers where the difference between ΔL and ΔF exceeds the average difference across layers (paired t-test, $p < 0.05$).

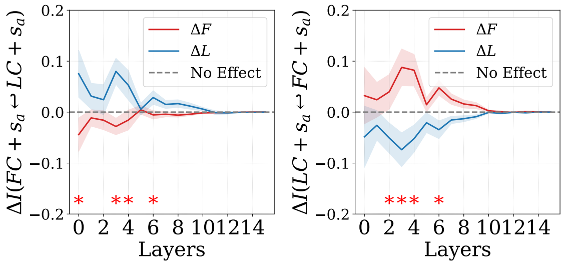
retrieved. The effect persists but is small compared with the figurative context condition, where early layers at the idiom token appear to encode context that suppresses figurative retrieval while biasing toward the literal interpretation.

A similar pattern is observed in the MHSA layers. As shown in Figure 7c, under figurative context, ablating early layers (0–3) produces a pronounced decrease in ΔF (min: -0.33) and an increase in ΔL (max: 0.48). Under literal context, ablating the earliest layers (0–1) yields a drop in ΔF (min: -0.32), whereas only modest changes for ΔL (max: 0.16) relative to the figurative context condition. That is, the literal context suppresses retrieval of the figurative interpretation from the very earliest layers. Moreover, ablating the layers (3–7) under literal context flips the pattern, where ΔF is increasing while ΔL is decreasing with a significant difference (Figure 7d). This indicates that, after the initial retrieval of figurative interpretation, the literal interpretation, which is disambiguated from context, is refined through mainly MHSA mediated integration with surrounding tokens.

In sum, the results point to a two stage context disambiguation: (1) early layers retrieve figurative evidence while beginning to encode context, followed by (2) mid-layers, MHSA-driven contextual disambiguation that resolves conflicts in favor of the context-aligned interpretation.

6.2 Probing MHSA contextual disambiguation with Query-preserved KV patching

To examine in detail how MHSA mediates contextual disambiguation in early-mid layers, we perform Query-preserved Key-Value patching, in which MHSA's Key–Value (KV) activations are swapped while selectively preserving idiom tokens' query (Q) activations across layers, thereby redirecting the model's attention toward alternative contexts. That is, we switch idiom token span's attention from a figurative-context to the literal context, and vice versa. If the model's final prediction changes from figurative to literal (or vice versa) after patching, it demonstrates that the MHSA mechanism in those specific layers is responsible for using the context in the disambiguation process.



(a) FC $+s_a \hookleftarrow$ LC $+s_a$ at idiom span

(b) LC $+s_a \hookleftarrow$ FC $+s_a$ at idiom span

Figure 8: Layer-wise interpretation shifts ΔI under Q-preserved KV patching. **Y-axis:** Mean $\Delta L(s_a)$ and $\Delta F(s_a)$ across idiom sentences. **X-axis:** Layers. **Gray dashed line:** $\Delta I = 0$ (no effect). The red asterisk (*) marks layers where the difference between ΔL and ΔF exceeds the average difference across layers (paired t-test, $p < 0.05$).

Results As shown in Figure 8, patching in KVs with alternative context yields a shift in interpretation of idiom. Under *FC*, effects appears already at layer 0, indicating immediate context sensitivity, with additional peaks at layers 2-3 and 6. Under *LC*, the probability of the context matched interpretation drops across layers 2–6, while the alternative interpretation rises. Taken together, these results suggest that the idiom tokens exploit contextual evidence primarily in layers 2-6, after the figurative interpretation is first retrieved in layers 0-1, with layer 0 playing a crucial early role under *FC*. Later layers tend to consolidate rather than redirect the interpretation.

7 Conclusion

We present a multi-stage mechanism of idiom comprehension in causal transformers, by identifying distinct components and flows for figurative and literal interpretations. Causal interventions reveal (i) early retrieval of the idiomatic interpretation, (ii) immediate use of preceding context, with later refinement when it conflicts with the retrieved interpretation, (iii) a selective pathway that carries the figurative interpretation through intermediate layers, and a bypass route that favorably delivers the literal interpretation directly to the output.

Considering idioms as multiword expressions, our results suggest that early layers (especially MLPs) perform a "detokenization" step (Gurnee et al., 2023; Elhage et al., 2021) that compresses the idiom into a unified figurative representation, while later pathways can "retokenize" it into a competing literal interpretation. To sum up, these findings link representational dynamics to contextual disambiguation, yielding a concise mechanistic interpretation of idiom processing.

Limitations

While causal tracing methods have been widely used in recent work (Geva et al., 2023; Dar et al., 2023; Meng et al., 2022; Heimersheim and Nanda; Nanda et al., 2023), they only approximate the actual information stored in activations. Moreover, knocking out or replacing the activations can lead the model to out-of-distribution behaviors and cast doubt on the robustness of any interpretability claim.

Moreover, we binarize idiom interpretation (figurative vs. literal), ignoring polysemy among figurative senses (e.g., 'break the ice' can be interpreted into initiate talk, ease tension, etc.), which can flatten nuance and bias evaluation. As a future direction, we can use soft labels for multiple figurative sense and distributional metrics.

Ethical Considerations

One of the intended uses of the Llama-3.3-70B-instruct model is content generation, which aligns with our use of the model for generating the candidates for next token prediction task (Grattafiori et al., 2024). The candidates generated by the language model may include biased or sensitive attributes (e.g., race, minority status), which reflects stereotypes that the language model already has (See the Appendix Table 3 for examples).

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC Starting Grant "Individualized Interaction in Discourse", grant agreement No. 948878). Funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. We gratefully acknowledge the stimulating research environment of the GRK 2853/1 "Neuroexplicit Models of Language, Vision, and Action", funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number 471607914.

European Research Council
Established by the European Commission

References

Sara D Beck and Andrea Weber. 2020. Context and literality in idiom processing: Evidence from self-paced reading. *Journal of Psycholinguistic Research*, 49(5):837–863.

Samuel A Bobrow and Susan M Bell. 1973. On catching on to idiomatic expressions. *Memory & cognition*, 1:343–346.

Cristina Cacciari and Sam Glucksberg. 1991. Chapter 9 understanding idiomatic expressions: The contribution of word meanings. In Greg B. Simpson, editor, *Understanding Word and Sentence*, volume 77 of *Advances in Psychology*, pages 217–240. North-Holland.

Cristina Cacciari and Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of memory and language*, 27(6):668–683.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. 2025. Revisiting in-context learning inference circuit in large language models. In *The Thirteenth International Conference on Learning Representations*.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

Brian C Cronk, Susan D Lima, and Wendy A Schweigert. 1993. Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of psycholinguistic research*, 22:59–82.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

Raymond W Gibbs. 1980. Spilling the beans on understanding and memory for idioms in conversation. *Memory & cognition*, 8(2):149–156.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36:76033–76060.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. *URL https://arxiv.org/abs/2404.15255*.

Edward Holsinger. 2013. Representing idioms: Syntactic and contextual effects on idiom processing. *Language and speech*, 56(3):373–394.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Indra Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284.

Agne Knietaite, Adam Allsebrook, Anton Minkov, Adam Tomaszewski, Norbert Slinko, Richard Johnson, Thomas Pickard, Dylan Phelps, and Aline Villavicencio. 2024. Is less more? quality, quantity and context in idiom processing with natural language models. *arXiv preprint arXiv:2405.08497*.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

nostalgebraist. 2020. Interpreting gpt: The logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.

Ye Tian, Isobel James, and Hye Son. 2023. How are idioms processed inside transformer language models? In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 174–179.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *The Eleventh International Conference on Learning Representations*.

A Identified top-20 idiomatic heads

Table 1 lists the attention heads most influential in retrieving figurative meanings while suppressing literal interpretations and randomly selected heads.

Type	top-20 attention heads (layer, head)
Idiomatic	(0, 4), (1, 5), (0, 30), (0, 19), (0, 8), (9, 30), (2, 2), (1, 18), (0, 21), (2, 8), (1, 9), (4, 2), (1, 24), (1, 13), (3, 18), (0, 0), (3, 24), (0, 26), (1, 27), (0, 28)
Random	(9, 20), (10, 30), (2, 11), (2, 23), (1, 1), (9, 23), (13, 24), (8, 31), (3, 8), (13, 9), (10, 1), (5, 15), (7, 14), (7, 24), (4, 5), (14, 2), (9, 17), (4, 0), (14, 19), (6, 9)

Table 1: Top-20 attention head sets.

B Data generation pipeline

We illustrate the data generation pipeline in Figure 9.

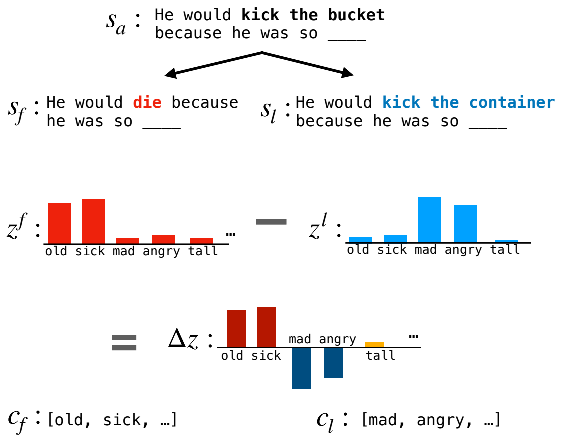
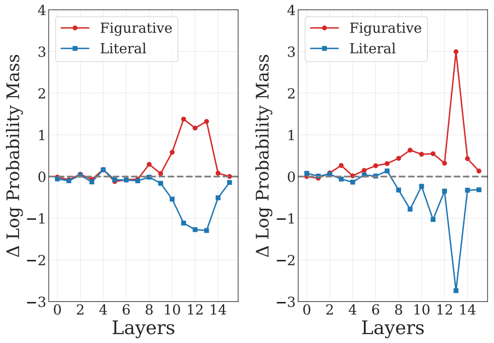

s_a : He would kick the bucket because he was so ____
s_f : He would die because he was so ____
s_l : He would kick the container because he was so ____
z^f :
old sick mad angry tall
z^l :
old sick mad angry tall
Δz :
old sick mad angry tall
c_f : [old, sick, …]
c_l : [mad, angry, …]

Figure 9: The pipeline of data generation.

C Layer-wise context disambiguation via logit lens

If the MHSA layers 3-7 play a causal role in context disambiguation, then we would expect that the representation at the final idiom token is fully disambiguated towards either the literal or the figurative meaning, depending on the earlier context. We employ the logit lens method (nostalgebraist, 2020) to measure the log-probability assigned to figurative and literal interpretations in the figurative (FC) vs. literal context (LC) conditions. At each layer, the hidden representations are projected into the output vocabulary space, yielding candidate probabilities. We then measure the difference in log-probability mass between figurative and literal candidates across layers, separately isolating contributions from the MHSA and MLP components.



(a) MLP FC−LC (b) MHSA FC−LC

Figure 10: Layerwise differences in log-probability between figurative and literal candidates predictions under figurative (FC) vs. literal (LC) contexts, measured with the logit lens.

Results Figure 10 shows a clear disambiguation effect for MHSA starting around layer 8. Once MHSA starts modulating activations in a contextdependent manner, it writes this information into the residual stream, which is subsequently processed by the MLP layers. In the early to middle layers (0–9) of the MLP, the difference in log probabilities between figurative and literal candidate predictions under figurative vs. literal contexts remains minimal. However, beginning at layer 10, MLP activations begin to diverge: idioms presented in figurative contexts are increasingly encoded with figurative interpretations, whereas idioms in literal contexts are increasingly encoded with literal interpretations.

D Knockout for crucial role of early layers with other models

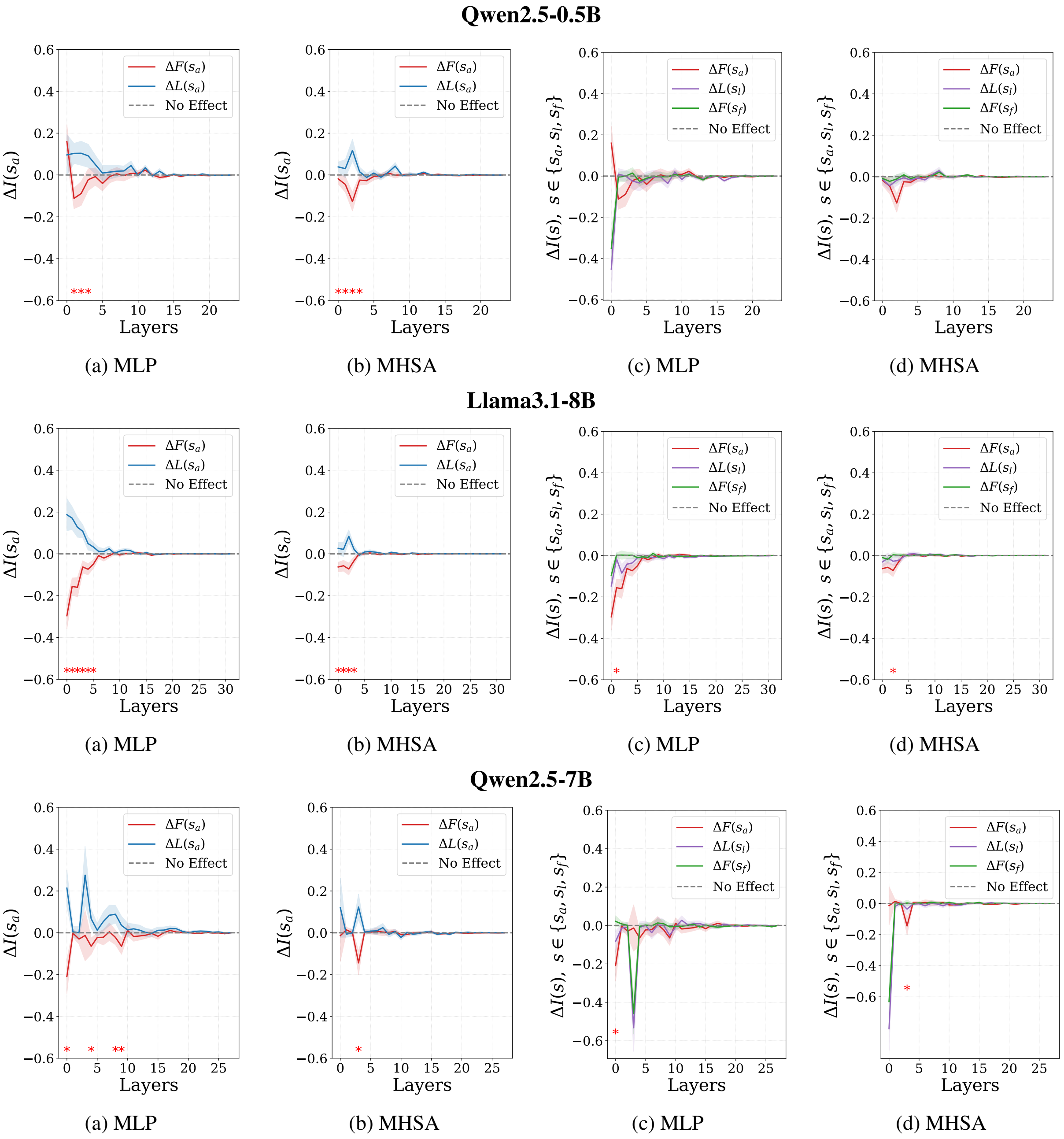
Results across different models in Figure 13 consistently show that early layers retrieve idioms' figurative interpretations while suppressing literal counterparts, unlike unambiguous sentences.

E Activation patching for information flow with other models

Across models, the intermediate and direct pathways consistently prefer different interpretations (see Figure 16).

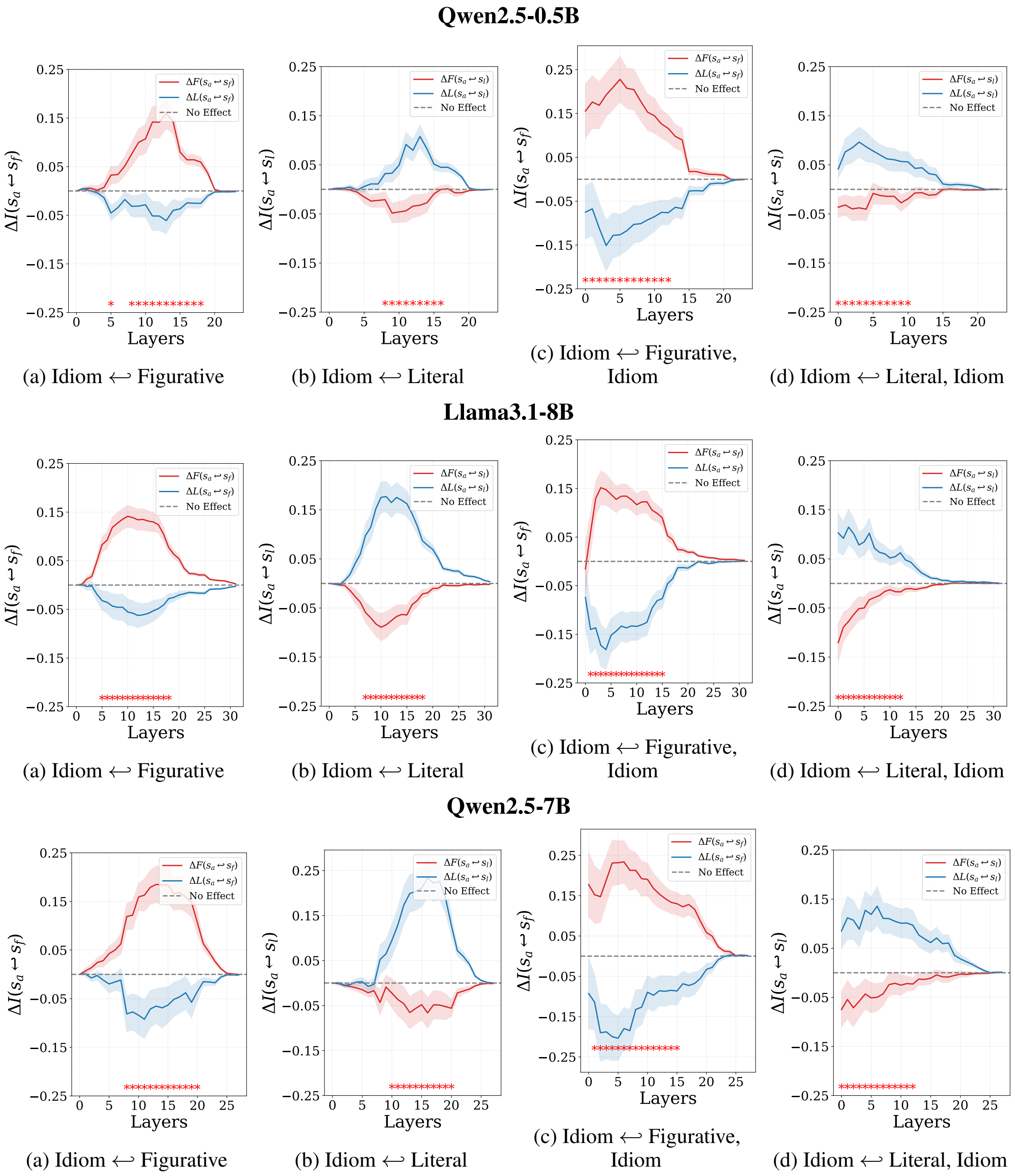
F Knockout for context disambiguation with other models

Across different models, Figure 19 consistently indicate that early MLP layers retrieve idioms' figurative meanings, and following layers promotes contextual cues via MHSA when a literal context is present.



(a) MLP (b) MHSA (c) MLP (d) MHSA

Figure 13: Sublayer-wise interpretation shift $\Delta I(s)$ after ablating activations at idiom span, for sentences $s \in \{s_a, s_f, s_l\}$. **Y-axis:** Mean values of $\Delta L(s_a)$, $\Delta F(s_a)$, $\Delta L(s_l)$, $\Delta F(s_f)$ with 95% confidence intervals. **X-axis:** Layers. **Gray dashed line:** $\Delta I = 0$ (no effect). **Red asterisk (*):** Significant difference between $\Delta F(s_a)$ and the others (paired t-test, $p < 0.05$). The difference at * marked layer is larger than the average difference across all layers.



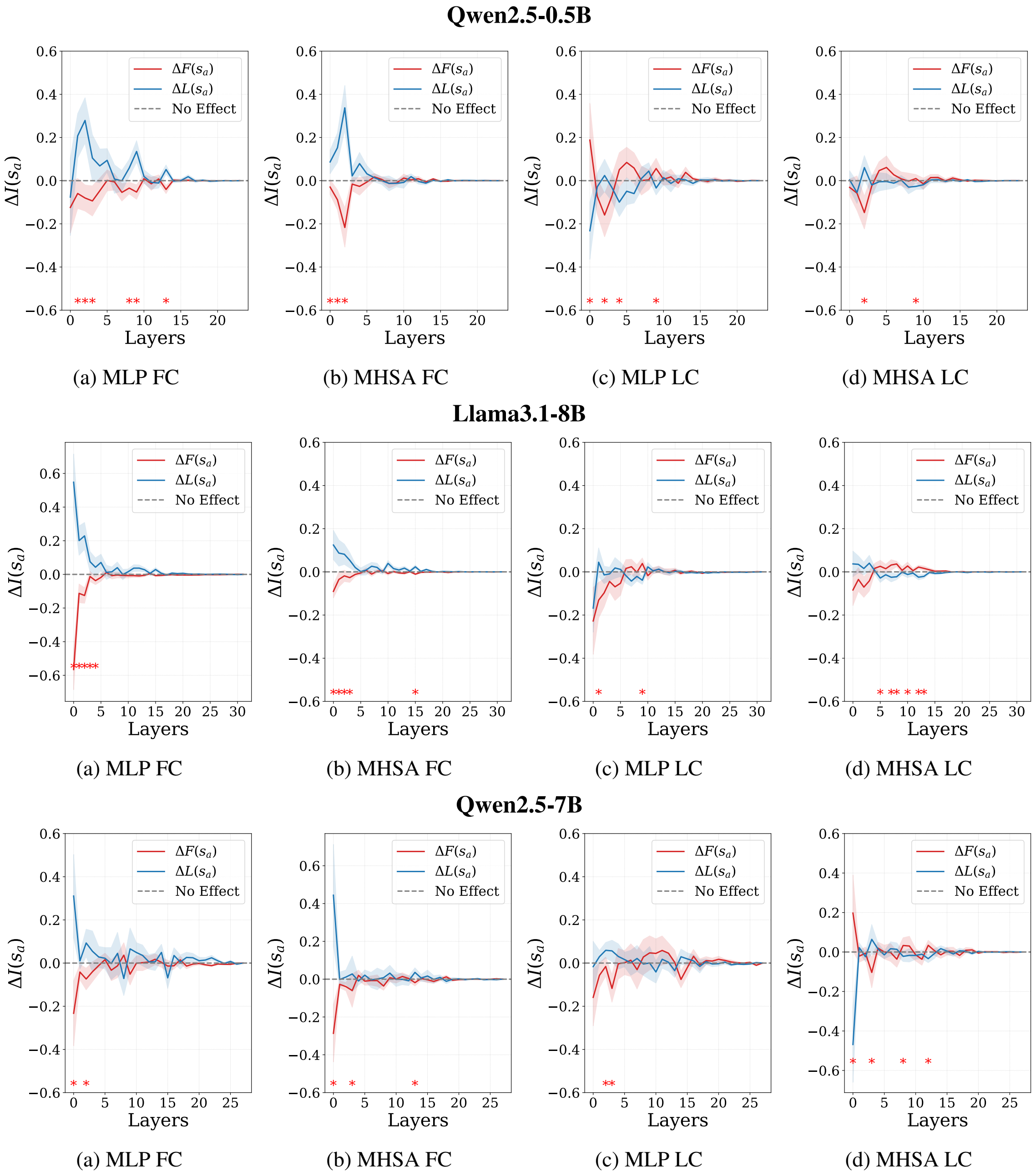
(a) Idiom ⟵ Figurative

(b) Idiom ⟵ Literal

(c) Idiom ⟵ Figurative, Idiom

(d) Idiom ⟵ Literal, Idiom

Figure 16: Layer-wise interpretation shift after patching in activations from s_f and s_l and vice versa. The red asterisk (*) marks layers where the difference between ΔL and ΔF exceeds the average difference across layers (paired t-test, $p < 0.05$).



(a) MLP FC (b) MHSA FC (c) MLP LC (d) MHSA LC

(a) MLP FC (b) MHSA FC (c) MLP LC (d) MHSA LC

(a) MLP FC (b) MHSA FC (c) MLP LC (d) MHSA LC

Figure 19: Sublayer-wise interpretation shift $\Delta I(C + s_a)$ after ablating activations at idiom span, for contexts $C \in \{FC, LC\}$. **Y-axis:** Mean values of $\Delta L(C + s_a)$, $\Delta F(C + s_a)$ with 95% confidence intervals. **X-axis:** Layers. **Gray dashed line:** $\Delta I = 0$ (no effect). **Red asterisk (*):** Significant difference between $\Delta F(C + s_a)$ and the others (paired t-test, $p < 0.05$). The difference at * marked layer is larger than the average difference across all layers.

Ambiguous sentence	**Literal paraphrase**	**Figurative paraphrase**	**Literal candidates**	**Figurative candidates**
They will bend over backwards because they are so	They will arch spine backwards because they are so	They will make extra efforts because they are so	flexible, used, strong, weak, relaxed, tight, uncomfortable, tall, stiff, short, scared, full, comfortable, tense, small, thin, angry, over, inf, surprised	grateful, eager, proud, passionate, close, motivated, committed, keen, desperate, enthusiastic, invested, confident, glad, well, interested, sure, apprec, focused, attracted, dedicated
He bit off more than he can chew because he was so	He took more food than he can swallow because he was so	He took on more than he could handle because he was so	hungry, greedy, happy, very, nervous, fam, attracted, r, thirsty, poor, star, pleased, starving, sad, delighted, hung, eng, drunk, gl, tempted	eager, confident, anxious, desperate, optimistic, enthusiastic, passionate, sure, determined, ambitious, focused, driven, busy, good, full, young, keen, convinced, smart, strong
He would blow his own horn because he was a	He would blow the musical instrument because he was a	He would praise himself because he was a	professional, musician, fan, p, wind, skilled, shepherd, trumpet, member, trump, pro, fl, virt, jazz, boy, bag, human, brass, flute, musical	good, great, man, self, genius, proud, winner, god, true, hero, hard, legend, unique, better, brilliant, successful, clever, narciss, smart, perfection
It was out in left field because it was a	It sat far in baseball's area because it was a	It was completely unrealistic because it was a	baseball, league, minor, sport, strong, few, popular, difficult, download, home, football, smaller, tough, sports, right, significant, basketball, pitcher, deep, stadium	fantasy, dream, very, completely, huge, one, product, total, movie, story, complete, perfect, romantic, fairy, totally, cartoon, fictional, massive, two, film
They were up the creek because they were a	They were near the stream because they were a	They were in trouble because they were a	group, family, fishing, nom, part, hunting, water, tribe, pair, thirsty, party, people, river, fish, stream, traveling, farming, curious, type, peaceful	new, long, threat, small, minority, bunch, few, mixed, very, day, single, large, bad, man, young, mess, tiny, mix, poor, relatively

Table 2: Examples of generated data.

Ambiguous sentence	**Literal paraphrase**	**Figurative paraphrase**	**Literal candidates**	**Figurative candidates**
They left him out in the cold because he was a	They left him outside in the frost because he was a	They abandoned him without support because he was a	bad, stranger, little, bit, thief, trouble, poor, drunk, witch, dirty, dog, rebel, beg, danger, trait, s, nuisance, tiny, he, naughty	child, foreign, threat, boy, political, baby, **black**, non, reminder, **disabled**, male, burden, cripp, product, **minority**, son, member, difficult, minor, different
They let him off the hook because he was a	They removed him from fishing tackle because he was a	They freed him from responsibility because he was a	threat, bad, convicted, little, bit, danger, sexual, fish, ped, **white**, criminal, racist, bully, **black**, terrorist, thief, political, jerk, rap, big	**minor**, child, foreign, slave, kid, good, mad, juvenile, victim, boy, stranger, young, youth, mere, student, fool, first, teenager, prisoner, friend
She was looking for a needle in a haystack because she was a	She was searching for a needle within a straw because she was a	She was facing an impossible search because she was a	needle, craft, seam, straw, tiny, hay, detective, master, haystack, witch, farmer, pro, camel, professional, busy, crazy, giant, chicken, cow, neat	**woman**, victim, single, young, **black**, **female**, **girl**, **slave**, mother, stranger, new, first, non, private, **white**, novice, **minority**, prisoner, foreign, mom

Table 3: Biased data examples reflecting stereotypes embedded in the language model.