Taking the plunge: A deep dive into the representation of figurative information in large language models

Anonymous ACL submission

Abstract

Large language models are capable of processing compositional phrases and multiword expressions, such as idioms, however it is still unclear how idioms are internally represented. This research investigates the hidden state space of transformer models, using representational similarity analysis and activation patching to identify internal components that causally contribute to the representation of idiomatic information. The results reveal that idiomatic processing begins in the early layers of models, alongside compositional processing of the individual tokens. Later layers of models use context to reinforce idiomatic interpretations. Additionally, smaller models rely on localised circuits for idiomatic processing whilst larger models are more robust with processing distributed throughout the model. These findings shed light on how state of the art language models process and represent idiomatic information.

1 Introduction

The human lexicon is constantly growing, consisting of words, phrases, and expressions that are carefully chosen by humans during everyday speech. Expressions such as "kick the bucket" and "spill the beans" are considered as potentially idiomatic expressions (PIEs), with disambiguation often requiring surrounding context. For example, "the old man kicked the bucket" provides the necessary context required to shift from a literal physical action to a figurative state (death). It is this transition between literal and figurative understanding that has led to research into how humans process and interpret idioms, metaphors, and other compositional semantic phrases.

Summarised by Cacciari and Tabossi (2014), theories of idiomatic comprehension are split into two different types of models: non–compositional models, where idiom meanings are accessed from pre–existing memory or knowledge bases, and compositional models where idiom processing requires simultaneous processing of literal words with surrounding context. Hybrid approaches to idiomatic processing suggest that varying degrees of non–compositional and compositional processing take place depending on the idiom; decomposable idioms may favour compositional processing whilst non–decomposable idioms may struggle with mapping literal meanings to a figurative interpretation (Kintsch, 1988; Titone and Connine, 1999).

With the growing popularity of large language models (LLMs) and their proficiency at completing natural language processing (NLP) tasks, they have since proven to be capable of interpreting multiword expressions (MWEs), including noun–noun compounds, metaphors and PIEs. Previous work has explored MWE classification and identification tasks, however there remains a gap in the current understanding of how LLMs are able to process and represent figurative information. This work aims to answer the question: do LLMs encode and represent figurative information in their hidden state space, and which components contribute to the resolution of idiomatic information? This paper makes the following contributions:

- A syntactically controlled small dataset of idioms, designed to minimise noise and variability for representational similarity analysis.
- A comprehensive, synergic combination of analysis methods, combining correlational analyses with causal approaches.
- Localisation of the emergence of idiomatic information within the internal structure of the transformer.
- An overall multi–stage mechanism for idiomatic processing which mirrors psycholinguistic hybrid models of idiom comprehension.

2 Related works

Idiom detection. There have been a multitude of studies within psycholinguistics investigating how humans process idioms, with two diverging classes of theories. Non–compositional models of idiom processing imply that idioms are processed as a complete unit where figurative meanings are accessed from memory rather than processing each token, whilst compositional models follow the assumption that figurative meaning is built from a combination of individually processed tokens. Cognitive studies involving human participants have further supported hybrid models for idiom comprehension, where idiom properties such as predictability and decomposability affect the level of word–level processing (Gibbs Jr et al., 1989; Gibbs Jr and Nayak, 1989; Hubbard et al., 2023).

Studies on how LLMs recognise and handle MWEs, including idiomatic expressions, have primarily focused on idiom detection and classification, however more recent work has started to shed light on the compositional processing of idioms. By treating idioms as a single token, similar to non–compositional models, Hashempour and Villavicencio (2020) discovered that idiom detection in BERT increases by 6% when incorporating contextualised word embeddings. Hashiloni et al. (2025) used prompt-based approaches to identify PIEs from documents, showing that LLMs such as GPT–4o are capable of extracting the phrasal semantics.

Gomes (2025) studied how transformer–based language models process idioms using path patching to identify specialised circuits that are utilised during the processing of idioms. They discovered that idioms are processed in early layers, with tokens interacting to determine an underlying figurative meaning. Most recently, the work of Oh et al. (2026) discussed the competing pathways of figurative and literal interpretations as LLMs process idioms, where the literal interpretation is preserved within the model whilst a separate internal circuit considers the figurative interpretation.

Representational similarity analysis. A popular technique for analysing statistically similar patterns in the brain is representational similarity analysis (RSA), originally developed to compare the similarity between multivariate data (Kriegeskorte et al., 2008). Describing representational geometry as the pairwise dissimilarity between embedding vectors, Lepori and McCoy (2020) used RSA for investigating linguistic dependencies with BERT embeddings. They discovered that verb embeddings extracted from BERT encode information about their subjects. Ormerod et al. (2024) used RSA as a technique for comparing the token embeddings extracted from LLMs with a constructed ground truth matrix that is designed to encode a specific linguistic property. They found that bi–directional LLMs extracted and stored compositional semantic information from noun–noun compounds within the hidden state representation models. This was further supported by Kendrick et al. (2025) where RSA was used to investigate autoregressive LLMs, concluding that after processing the full compound, the middle layers of LLMs were responsible for extracting relational information. This technique is a post hoc approach to exploring the hidden state space, with insights into how LLMs extract and represent specific linguistic phenomena.

Activation patching. A technique from mechanistic interpretability, known as activation patching (AP), or causal tracing, is designed to identify local circuits within LLMs that process specific information. Vig et al. (2020) implemented this technique to investigate individual neurons and attention heads that are responsible for driving gender bias. This method was refined by Zhang and Nanda (2023), who presented a systematic study of patching and suggest recommendations: symmetric token replacement where the tokens being replaced are semantically related, and using logit difference to measure the effects of patching. Meng et al. (2022) used causal mediation to then inform interventions, concluding that states with the highest causal effect can be intervened for knowledge editing.

3 Data construction

Previous research into figurative expressions and compositional phrases has yielded datasets such as MAGPIE (Haagsma et al., 2020), IdioTS (De Luca Fornaciari et al., 2024) and IMPLI (Stowe et al., 2022). These datasets have been leveraged for idiom detection tasks, however, they vary in syntactic structure and include additional figurative expressions such as synonyms, metaphors, and proverbs.

For this research, four datasets of idioms are

constructed. Firstly, a dataset of 60 idioms are extracted from the MAGPIE corpus of potentially idiomatic expressions. The idioms were chosen and cross–referenced against the Cambridge Dictionary (labelled as idiomatic), in the verb–noun construction (VNC) syntactic form, i.e. "kick the bucket", "take the fall". For each idiom, a set of 8 sentences were constructed, such that the first two are labelled as idiomatic, the last two are literal translations of the idiom, and an additional four distractor sentences, also labelled as literal. The distractor sentences are created such that they include synonyms of the verb and noun within each phrase. Table 1 gives an example of a set of sentences for the idiom "kick the bucket".

Table 1: Example group of sentences for the idiom "kick the bucket".

Sentence	Label
The old man kicked the bucket and	Idiomatic
The sick man kicked the bucket and	Idiomatic
The old man threw the football and	Literal
The sick man threw the football and	Literal
The old man lifted the container and	Literal
The sick man lifted the container and	Literal
The old man passed away and	Literal
The sick man passed away and	Literal

This careful construction isolates the idiomatic property of the sentences, reducing the effects of syntactic variability. The idioms chosen vary in semantic flexibility and decomposability, although this is not investigated here.

A figurative contextual dataset and a literal contextual dataset are constructed using the same idioms and synonym distractors, however they are designed to enforce either figurative or literal interpretation of the idiom. For example, "the old man was in hospital for six months" provides preceding context to increase the likelihood of "kick the bucket" being interpreted figuratively. However, "the old man was angry" could precede the literal phrase of kicking a physical object.

For AP, a dataset of minimal idiom–literal pairs with symmetric token replacement is constructed. In this case, pairs of sentences are constructed in such a way that only the noun within each idiom is replaced, thus giving an idiomatic and literal prompt pair. This is repeated three times for each idiom, each with a set of three potential idiomatic and literal continuation words. These words were chosen by using GPT–2 to select the next most likely token, given each prompt as the input.

THE SUSPECT WAS NERVOUS AND SPILLED THE BEANS, IT WAS A [SECRET, SURPRISE, MYSTERY]

THE SUSPECT WAS NERVOUS AND SPILLED THE DRINK, IT WAS A [MESS, MISTAKE, ACCIDENT]

4 Methodology

RSA is used to explore the geometry of the hidden state space, that is the representations extracted at the end of each transformer layer. The representations are compared against an encoded ground truth matrix using Spearman correlation to investigate whether the representations reflect encoded idiomatic information. RSA is conducted using the following models: GPT–2–Small (GPT–2), Llama–3.2–3B (Llama), Mistral–7B (Mistral), Qwen–2.5–7B (Qwen), Falcon–3–Base (Falcon), Deepseek–R1–Distil–1.5B (Deepseek), MultiBERTs and RoBERTa.

4.1 Idiom processing RSA

Objective: Investigate whether hidden state representations capture figurative meaning from compositional idiom phrases.

Hypothesis: Representations for idiomatic phrases should appear closer to their direct literal translation, as opposed to literal sentences with semantically similar constituent words. This representation will emerge in later layers after the LLM has processed the full phrase.

To determine whether the extracted representations from LLMs store the figurative meaning encoded within idioms, RSA is implemented across six autoregressive LLMs. A binary ground truth representational dissimilarity matrix (RDM) is constructed to encode whether or not each pair of sentences convey the same meaning. Positive correlation with the ground truth would imply that idiomatic meaning is stored across the representations. A one–sample t–test is performed to highlight layers where the correlation is statistically greater than 0. To limit the effects of lexical similarity, the full upper triangular matrix of the empirical RDM is restricted to the last two columns within each idiom group.

An overview of the RSA experiment is shown in Figure 1, with the ground truth RDM shown in (a), an example empirical RDM shown in (b) and

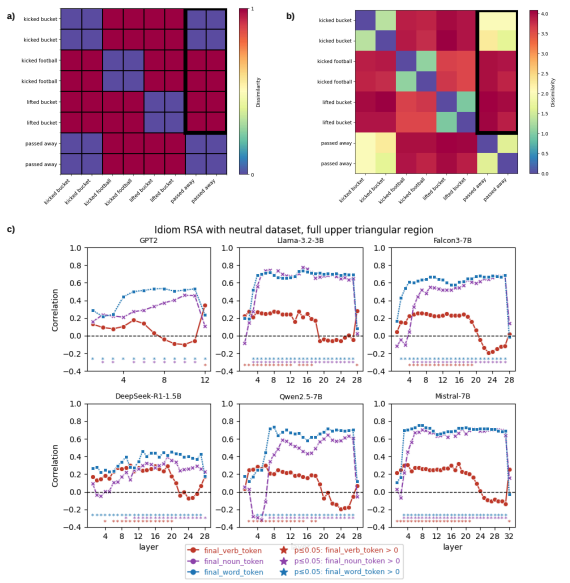

Figure 1: Idiom processing experiment: (a) Ground truth RDM (a dissimilarity score of 0 indicates shared meaning) with the restricted outlined region; (b) Example empirical RDM for Llama hidden state representations with the restricted outlined region, averaged across all idiom groups; (c) RSA results across six autoregressive LLMs showing Spearman correlation for three different token positions. Asterisks represent layers where the correlation is statistically greater than 0.

the correlation at three different token positions averaged across the 60 idioms in (c).

The RSA results demonstrate that idiomatic processing emerges early across all six models. Negative correlations between the noun token representations and the ground truth indicate lexical similarity rather than specific idiom processing, however this shifts and the models form stable idiomatic representations throughout the middle and later layers. As expected, the final token representation of each sentence displays the strongest correlation, highlighting the early emergence as correlation peaks between layers 4–8 for all models except Deepseek, where the correlation increases gradually across the layers. The correlation between the verb token representations and ground truth appear to decrease during the middle and later layers of processing, suggesting that the model integrates the embeddings from the verb with the rest of the phrase rather than representing the meanings of each constituent word separately.

Llama, Falcon, Qwen and Mistral produce the highest correlations (≈0.6), whilst GPT and DeepSeek peak at around 0.4. The final word token representations for all models produces statistically

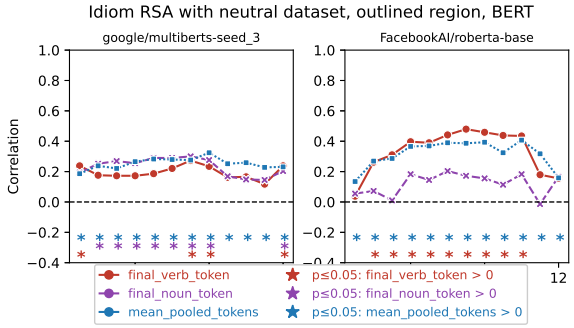

Figure 2: RSA results across BERT models showing Spearman correlation for verb, noun and mean–pooled token positions. Asterisks represent layers where the correlation is statistically greater than 0.

different results, whilst there are very limited asterisks for the final verb token. Spikes in correlation may be a result of specific idioms where the verbs are highly associated with figurative use. To investigate this, Figure 14 in Appendix D identifies the idioms that on average produce the highest correlations across the models, as well as the idioms which are consistently negatively correlating with the ground truth.

The bottom–ranked idioms "get a life" and "run a mile" are negatively correlated, suggesting that the LLMs focus more on lexical processing rather than extracting the figurative meaning. These idioms may appear more frequently in literal contexts, therefore requiring specific context to cement the idiomatic interpretation and steer away from individually processing each word. Contrastingly, the highest ranked idioms show strong correlations with the ground truth, indicating that the models are capable of extracting their phrasal semantics. These idioms may frequently appear in figurative context, leading to high idiomatic predictability and limiting the need to process each token individually to resolve the full meaning.

Figure 2 shows the results for the standard RSA processing with BERT models. As these models are bi–directional, the correlations across all three token positions for MultiBERTs are consistent showing that information is distributed across the phrase, however there is a noticeable difference between the token representations for RoBERTa, with the noun tokens producing correlations close to 0. The mean pooled representations across each input sequence fall in between the verb and noun token positions of the idiom phrase, balancing out the literal influence from the noun token. The corre-

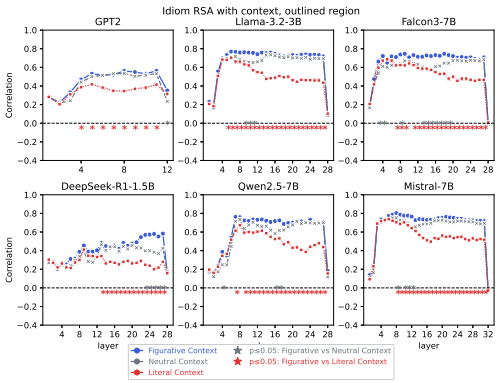

Figure 3: Results for the idiom processing RSA experiment with figurative, neutral and literal context, showing Spearman correlation strengths for the final word token representations. Asterisks reflect layers where the differences in correlations across context conditions are statistically significant.

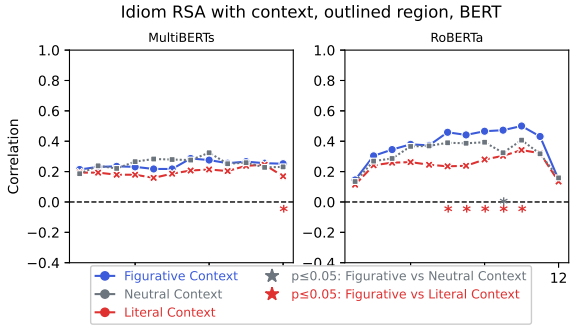

Figure 4: RSA results across BERT models showing Spearman correlation for verb, noun, and mean–pooled token positions. Asterisks reflect layers where the differences in correlations across context conditions are statistically significant.

lation for MultiBERTs remains lower than the other LLMs, whilst the final verb token representations of RoBERTa produce similar correlations to GPT and Deepseek. These lower correlations highlight that whilst the smaller BERT models may be able to identify idioms, their internal representations struggle to capture their phrasal level semantics and may be less robust to unseen or semantically flexible idioms.

4.2 Idiom processing RSA with context

Objective: Investigate how exposure to context affects LLM representations of idiomatic information, and whether literal interpretation can be enforced.

Hypothesis: Figurative context should increase the Spearman correlation, however the introduction of literal context should inhibit the extraction of idiomatic information.

The results for the autoregressive LLMs are shown in Figure 3, with significant difference between the Figurative and Literal context plotted. Only the final word token representations are visualised as they produce the highest correlations.

Interestingly, whilst there is a statistically significant difference across the figurative and literal representations for all models but GPT, adding figurative context to the data doesn't significantly increase the original correlation. However, forcing the models to interpret the phrases literally does reduce correlation, altering how models represent their information in the hidden state space. The representations begin to move away from reflecting idiomatic information, with some literal processing influencing the models.

Repeating the context experiment with the BERT models shows a similar pattern (Figure 4), where literal context reduces the correlation significantly for RoBERTa. For MultiBERTs, it is interesting that both figurative and literal context yield lower correlations than the neutral context dataset.

5 Localising figurative information

The results from the RSA experiment suggest early emergence of idiomatic information, however it is still unknown how this information is extracted from the textual input and represented within LLMs. AP is a fine–grained tool that can be used to identify components within the transformer mechanism of the smaller GPT model and the 7 billion Qwen model that contribute to idiom comprehension.

Objective: Localise idiomatic processing to a certain set of components within the transformer mechanism of LLMs and discover how processing varies across layers.

Hypothesis: There exists a set of components within the LLM that are responsible for controlling the shift from literal to idiomatic processing.

Patching is carried out on the hidden state representations extracted after the last linear layer of each transformer layer and at the attention heads within the transformer. A diagram of the patching method is included in Appendix A.

5.1 Hidden state patching

For this experiment, the dataset of idiom and literal pairs is used, where three pairs of sentences

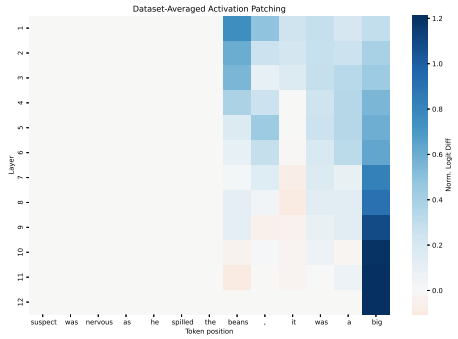

Figure 5: Normalised logit difference averaged across the dataset after hidden state patching with GPT.

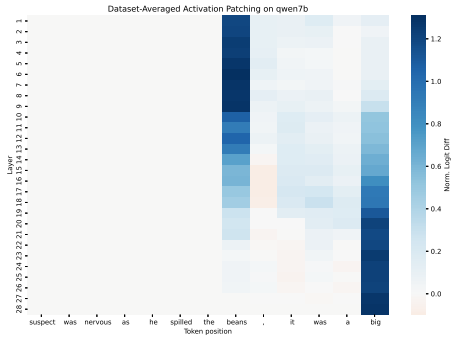

Figure 6: Normalised logit difference averaged across the dataset after hidden state patching with Qwen.

are passed through the LLM and the hidden state representations are extracted, along with the logit difference between a continuation word that follows the idiomatic sentence, and a continuation word that follows the literal sentence.

For each sentence pair, the representations from the clean run using the idiomatic prompt are patched into the model during the forward pass of the corrupt run with the literal prompt. This is known as denoising, often used as a discovery tool to find a minimal set of components responsible for a task. This AP is completed layer–wise and token–wise and the normalised logit difference averaged across each pair is plotted to identify where idiomatic information from the clean run has an effect.

The normalised logit difference after patching the hidden state representations is shown in Figure 5. Token representations which cause the model to shift from a literal interpretation to a figurative interpretation should reflect strong positive normalised logit differences, as the model favours the idiomatic continuation tokens. The final token at layer 12 shows the strongest positive logit difference which is to be expected, as the LLM has processed the full idiomatic prompt and predicted the next token to follow this context. However, as the activations from the token “beans” is patched into the corrupt prompt, early layers already begin to shift from literal to figurative processing. As later tokens are patched, this begins to move into middle layers before the last token creates a strong figurative interpretation through the model.

A similar pattern occurs with the Qwen model, with the strongest logit difference of 1.2 in the early layers as the activations of “beans” are patched into the corrupt prompt. These results align with

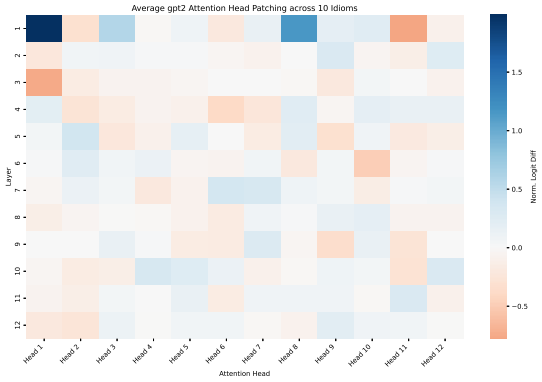

Figure 7: Normalised logit difference for the attention head patching experiment with GPT, averaged across the full dataset.

Gomes (2025) and the RSA results which present early emergence of idiomatic processing. This was also repeated for Llama, seen in Appendix B. Once again, the noun token produces a strong positive logit difference in the early layers, with following tokens providing a weaker but still positive logit difference in early to middle layers.

5.2 Attention head patching

For a more fine–grained approach, AP at the attention heads highlights if particular attention heads are responsible for passing the idiomatic meaning through the model.

The logit difference for the attention heads within GPT is shown in Figure 7. Similar to the hidden state patching, strong positive logit differences indicate that the particular attention head restores information which shifts the model towards an idiomatic interpretation. The strongest attention head is identified at layer 1, head 1, with secondary attention heads at layer 1, head 8 and layer 1 head 3. This further supports the theory that processing

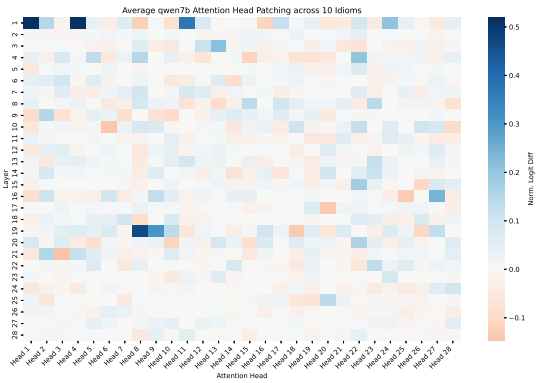

Figure 8: Normalised logit difference for the attention head patching experiment with Qwen.

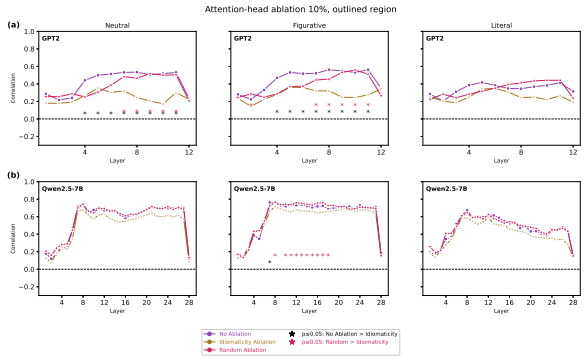

Figure 9: Targeted idiomaticity ablation with 10% of heads across three context conditions for: (a) GPT, and (b) Qwen. Asterisks show statistically significant results.

begins as early as the first transformer layer. Attention heads positively contributing to idiomatic interpretation are distributed across the model rather than being concentrated in a given layer, however the strength in logit difference decreases in later layers. This suggests that whilst the processing does continue throughout the model, the earlier layers contribute more to resolving the idiomatic information.

In Appendix B, Figure 12 shows the results of attention head patching with Llama. The first attention heads in the first layer produce strong positive logit differences, where the influence is spread throughout the network of attention heads in each layer.

Qwen shows a similar pattern of high positive logit difference for the first few heads in layer 1, however the strength in both positive and negative logit difference is weaker than GPT and Llama, as seen in Figure 8. There is a noticeable strong idiomatic signal at layer 19, head 8, which is anomalous with the other results where strong idiomatic processing was displayed in the early layers, which could serve as a refinement attention head, responsible for reinforcing the idiomatic interpretation. However, the overall reduced strength of the attention heads implies that Qwen uses distributed attention heads throughout layers to shift between literal and idiomatic interpretations.

6 Idiomaticity ablation

AP identifies specific components that are able to restore the idiomatic interpretations after replacing literal activations with idiom activations, leading to positive logit differences. Ablation is then implemented to further test whether the attention heads contribute to the representation of figurative meaning by repeating the RSA experiment whilst zeroing the specific attention heads in the forward pass and plotting the correlation. Random ablation repeated across five runs is included as a control.

Objective: Use identified components to intervene the forward pass and disrupt the encoding of idiomatic information.

Hypothesis: With the targeted idiomaticity ablation, the RSA correlations reduce as the hidden state space no longer represents the shared idiomatic meaning.

Figure (a) in Figure 9 highlights that when only 10% of attention heads are ablated in GPT, there is a sharp decline in correlation for the targeted idiomaticity ablation across all three contextual conditions. In contrast, random ablation does not have the same effect, with only slight changes in correlation. Figure (b) shows that ablating 10% of Qwen heads produces a clear decrease in correlation after idiomaticity ablation across all three context conditions.

7 Discussion

The results from the RSA experiment suggest that LLMs geometrically represent idiomatic meaning in their hidden state space. The key theme identified is the early emergence of idiomatic processing, with added context dynamically reshaping the hidden state space through later layers of processing. This early extraction could be the result of exposure to frequently occurring idioms and VNCs commonly used in figurative contexts. This aligns with psycholinguistic theories of idiom processing, where surrounding context reinforces or suppresses the representation of idiomatic information.

The hidden state and attention patching results

indicate that idiom processing begins in the early layers, with attention heads in layer 1 causally contributing to a shift from literal to idiomatic interpretations, leading to the evidence of idiomatic meaning present hidden state space of the first transformer layer. The smaller GPT model highlights attention heads with a strong idiomatic signal which may reflect a localised circuit for restoring idiomatic meaning. Contrastingly, the larger Qwen model reflects a more distributed representational structure, where heads exhibit weaker positive signals overall, with many that have little impact on the literal to idiomatic shift. This makes the larger LLMs robust and less dependent on localised internal circuits. Negative logit differences indicate that, despite replacing literal token activations with activations from an idiomatic prompt, there are certain heads that maintain a literal interpretation through the model and suppress figurative activations, which mirrors a hybrid model of processing with both literal and figurative interpretations taking place.

Using targeted idiomaticity ablation, the results provide causal evidence to support the theory that there are attention heads responsible for extracting and representation idiomatic information. The decrease in correlation with the ground truth RDM indicates that the hidden state space becomes less aligned with figurative semantic information due to neutralising targeted "idiom heads". The biggest impact is evident with GPT, further supporting the concept that GPT relies on sparse specific attention head circuits for cleanly representing idiomatic information. Comparing effects across the three context conditions highlights that the ablated idiom heads are also important for context disambiguation between potential literal and idiomatic interpretations. The decrease in correlation for Qwen is strongest in the middle layers of the model, as opposed to the early layer disruption with GPT. This observation supports early suggestion that large models like Qwen refine and reinforce idiomatic meaning in later stages of processing, unlike GPT which depends on localised circuits in early layers. The idiomaticity ablation is repeated with 25% attention heads ablated (Appendix C), with the effect of removing idiom heads strongest for Qwen. Interestingly, GPT shows similar effects for both targeted and random attention heads. Due to limited numbers of localised circuits, ablating 25% of heads may overshoot and also remove heads that contribute to overall semantic processing.

These findings indicate that idiom comprehension requires a multi-stage mechanism. Early layers identify possible idiomatic candidates, whilst also processing the expressions at an individual token level. Both literal and figurative meanings are maintained through the model, however additional context helps to strengthen the probability of a particular interpretation. The later model layers refine and finalise figurative meaning, organising representations to reflect their phrasal semantics.

8 Limitations and future work

The main limitation of this study was the size of the datasets used, as these were carefully selected to fit within the constraints of the RSA experiment. Motivated by careful psycholinguistic experimental control, the syntax was strictly fixed for the idioms chosen, and so it is unclear whether the conclusions made here extend to idioms of different structures and other figurative compositional expressions. AP was only carried out on hidden state representations and attention heads, but this does not explore interactions between heads and the tokens that they attenuate to. Future work could extend this framework to consider a wider range of idioms, varying in length, syntax, and decomposability. Using knowledge editing methods on the identified attention heads, it may be possible to enhance idiomatic understanding and improve comprehension of low and negatively correlated idioms without fully retraining the LLMs.

9 Conclusion

This work combined the exploration of RSA with the causal mechanistic tool of AP to localise where idiomatic information is represented. The results provide evidence of a sophisticated framework for processing idioms, extracting phrasal semantics and representing figurative meaning. Idiom comprehension can be summarised into a multi-stage process: early token–level processing, identifying potential idiomatic components whilst maintaining literal interpretations followed by refinement in middle layers, taking into account surrounding context, and finally refining and organising the idiomatic meanings in the hidden state. Smaller models such as GPT contain localised internal circuits that contribute to idiom processing, whereas this is more distributed within the layers of larger models, making them more robust to interventions and reducing dependency on individual components.

References

C. Cacciari and P. Tabossi. 2014. *Idioms: Processing, structure, and interpretation*. Psychology Press.

F. De Luca Fornaciari, B. Altuna, I. Gonzalez-Dios, and M. Melero. 2024. A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

R. W. Gibbs Jr and N. P. Nayak. 1989. Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive psychology*, 21(1):100–138.

R. W. Gibbs Jr, N. P. Nayak, and C. Cutting. 1989. How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of memory and language*, 28(5):576–593.

A. Gomes. 2025. Anatomy of an Idiom: Tracing Non-Compositionality in Language Models. *Preprint*, arXiv:2511.16467.

H. Haagsma, J. Bos, and M. Nissim. 2020. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

R. Hashempour and A. Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80.

K. G. Hashiloni, O. Hefetz, and K. Bar. 2025. Easy as PIE? identifying multi-word expressions with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23782–23801.

R. Hubbard, N. Bulkes, and V. T. Lai. 2023. Predictability and decomposability separately contribute to compositional processing of idiomatic language. *Psychophysiology*, 60(8):e14269.

S. Kendrick, M. Ormerod, H. Wang, and B. Devereux. 2025. Investigating noun-noun compound relation representations in autoregressive large language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 253–263.

W. Kintsch. 1988. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, 95(2):163.

N. Kriegeskorte, M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P. A. Bandettini. 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141.

M. Lepori and T. R. McCoy. 2020. Picking BERT's brain: Probing for linguistic dependencies in contextualized embeddings using representational similarity analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3637–3651.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

S. Oh, X. Huang, M. Pink, M. Hahn, and V. Demberg. 2026. Tug-of-war between idioms' figurative and literal interpretations in LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2942–2958.

M. Ormerod, J. M. del Rincón, and B. Devereux. 2024. How is a "kitchen chair" like a "farm horse"? Exploring the representation of noun-noun compound semantics in transformer-based language models. *Computational Linguistics*, 50(1):49–81.

K. Stowe, P. Utama, and I. Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

D. A. Titone and C. M. Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of pragmatics*, 31(12):1655–1674.

J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

F. Zhang and N. Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.

A Activation patching overview

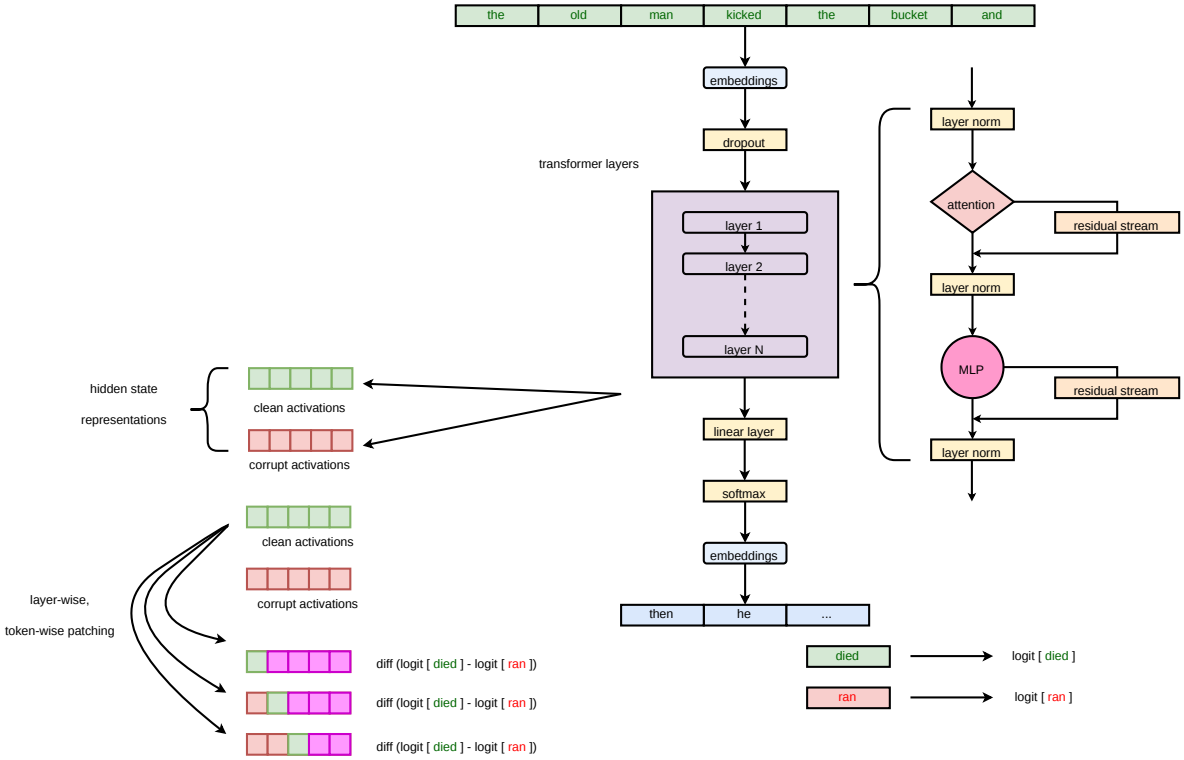

Figure 10: Overview of hidden state patching. A clean prompt is passed through the model, followed by a corrupt prompt, before the clean activations are injected into the corrupt prompt and the logit difference between a correct and an incorrect continuation word is plotted.

B Llama Patching Results

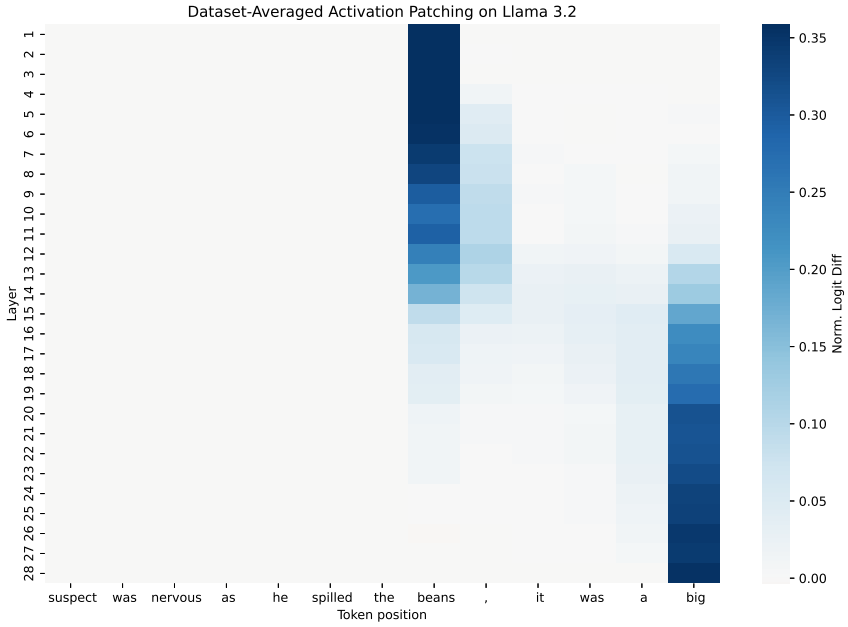

Figure 11: Normalised logit difference averaged across the dataset after hidden state patching with Llama.

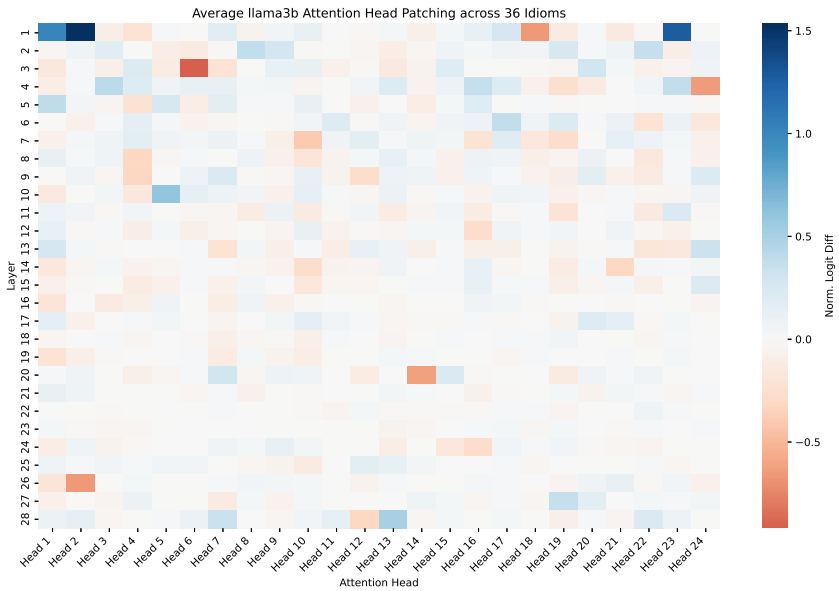

Figure 12: Normalised logit difference for the attention head patching experiment with Llama, averaged across the full dataset.

C Idiomaticity ablation II

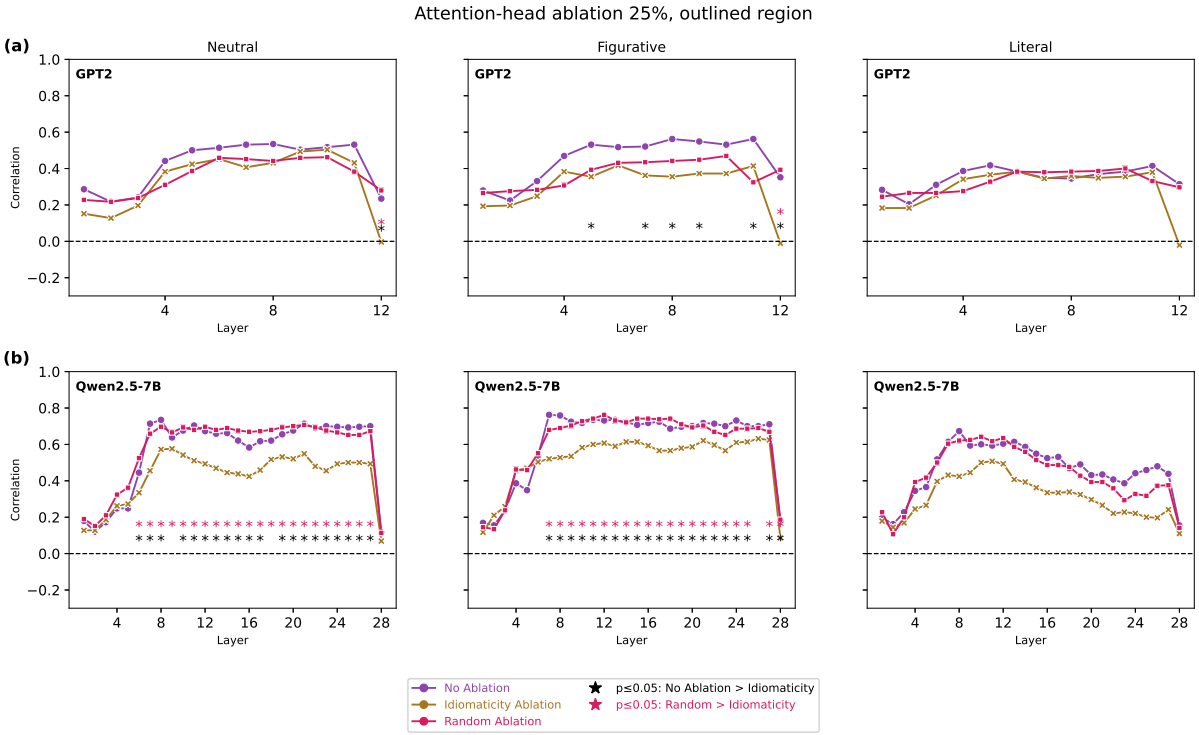

Figure 13: Targeted idiomaticity ablation with 25% of heads across three context conditions for: (a) GPT, and (b) Qwen. Asterisks show statistically significant results.

D Top Idioms

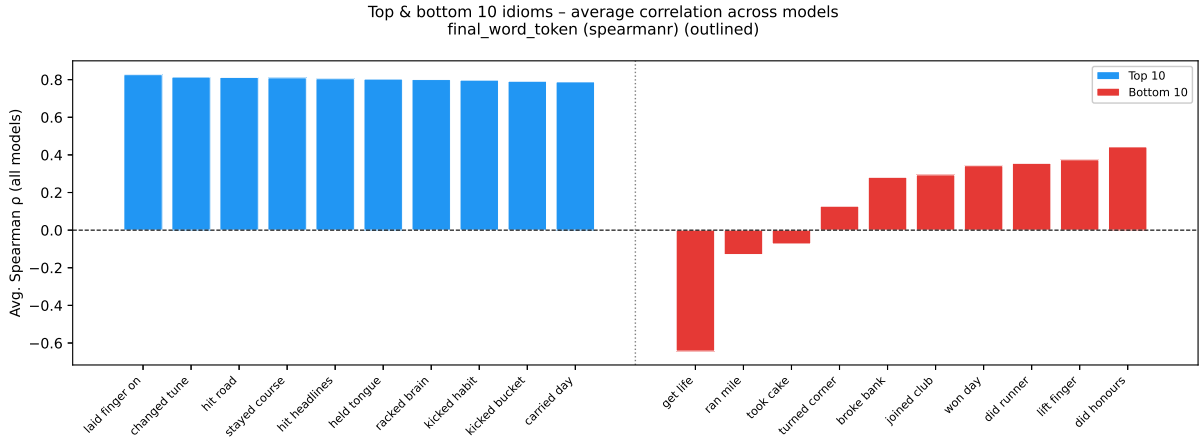

Figure 14: Top 10 and bottom 10 idioms across models based on average Spearman correlation for final word token representations.