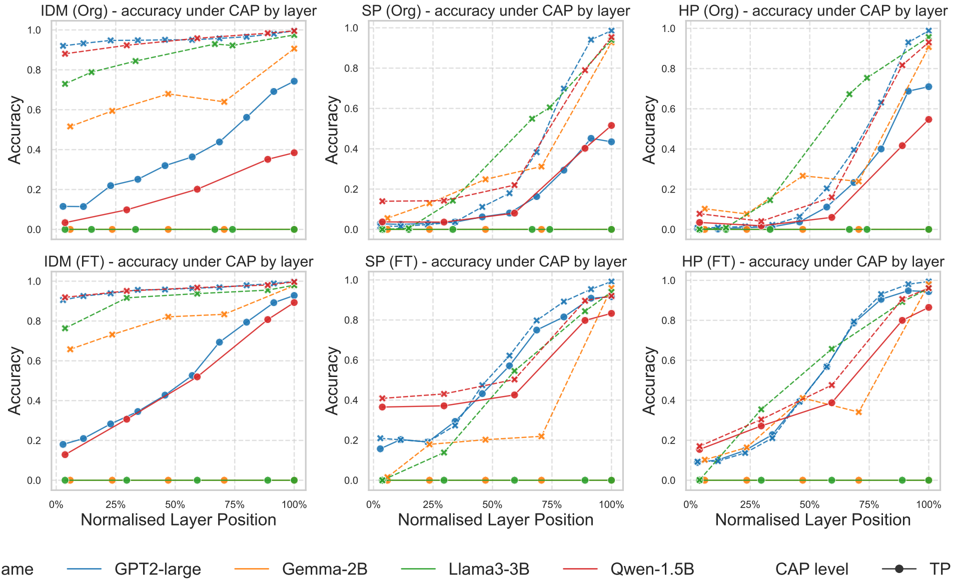

Figure 2: Average grouped accuracy of CAP across different aggregation functions for normalised layer positions (0%-100%) is shown for word-level CAP (TW) and phrasal-level CAP (TP). Sub-figures (a)-(c) illustrate the CAP effect on the original (Org) models, while sub-figures (d)-(f) show its impact on the fine-tuned (FT) models. Fine-tuning consistently improves performance, particularly in the middle to late layers (25%-100%), while early layers (0%-25%) show more variability and lower accuracy across models.

of Information Gain $IG_{Y,R_l(X)}$, and the aggregation of a pair of input token representations $R_l(X_i), R_l(X_j)$ in terms of their Mutual Information $I(R_l(X_i), R_l(X_j))$.

$IG_{Y,R_l(X)}$ quantifies the amount of information gained about the predicted token Y from the observation of the $R_l(X)$, for which the expectation is the mutual information $I(Y, R_l(X))$ of Y and $R_l(X)$, which is equivalent to the reduction in entropy of Y achieved by learning the state of $R_l(X)$: $IG_{Y,R_l(X)}(Y, r) = H(Y) - H(Y|r)$.

During training, $R_l(X)$ will be adjusted in a way that reduces the uncertainty about Y, meaning it will promote the maximisation of $IG_{Y,R_l(X)}$ for any given layer l, which can be expressed as:

$$IG_{Y,X} = max(\sum_l IG_{Y,R_l(X)}) \qquad (5)$$

where $IG_{Y,X}$ represents the information gain of Y w.r.t. input token X.

When looking at two input tokens X_i, X_j, the higher the mutual information $I(R_l(X_i), R_l(X_j))$ is, the lower the impact that aggregating $R_l(X_i)$ and $R_l(X_j)$ would have over $IG_{Y,X}$, as those variables share more of the same information. Intuitively, that would apply to linguistic composition, e.g., tokens that form a word and thus have a stronger dependence when observed together.

However, as the model's ability to predict Y is contingent on the accumulated information of all layers, and Equation 5 is independent of layer order, there is an intrinsic incentive to delay the aggregation of information (to later layers), as

$$IG_{R_{l_p}(X),R_{l_q}(X)} < IG_{R_{l_p}(X),R_{l_r}(X)}, \quad \forall p < q < r, \qquad (6)$$

where p, q and r are layer indices, i.e., subsequent layers have more information about the inputs than previous ones. This can be explained in that optimising Equation 5 can be achieved by retaining at each $R_{l_p}(X)$ only the necessary information to maximise $\sum_{i,j} IG_{R_{l_q}(X_i),MHA(R_{l_p}(X_j))}$, where $MHA(R_{l_p}(X_j))$ is the multi-head attention weighted representation. Such an objective implies minimising the mutual information $I(R_{l_p}(X_i), R_{l_p}(X_j))$, i.e., reducing redundancy across tokens from the same layer. Therefore, token dependencies will tend to be modelled by aggregation paths spanning multiple layers, with more layers allowing for more complex and longer paths. This is in line with the findings of Mechanistic Interpretability studies (Elhage et al., 2021; Conmy et al., 2023). Equation 6 also implies that the earlier an aggregation is done, the larger the impact it will have on $IG_{Y,X}$, which explains the empirical results. The effects of $I(R_l(X_i), R_l(X_j))$ on LLMs are further compounded by the tokenisation objective (e.g., BPE, WordPiece), which *minimises* $I(X_i, X_j)$, i.e., token redundancy, as a means of reducing the vocabulary size, leading to longer aggregation paths.

6 Related work

Compositionality, the principle that the meaning of complex expressions is derived from their parts and structure, is foundational in linguistics, cognitive science, and AI (Fodor, 1975; Montague and Thomason, 1975; Tull et al., 2024). In neural models, compositionality enables generalisation and interpretability, yet remains difficult to diagnose and enforce (Donatelli and Koller, 2023). Several studies investigate how and where compositional representations emerge in transformer models. Carvalho et al. (2025) observed similar effects in adjective-noun phrase probing, while Haslett (2024) found that models struggle to segment or represent morphemes, especially in non-Latin scripts, suggesting breakdowns in both form and meaning composition. The logit lens (Nostalgebraist, 2020) demonstrated that transformers build predictions progressively where early layers make initial guesses and deeper layers refine guesses with broader context. (Dai et al., 2022) show feed-forward layers act as key-value memories, combining information for complex predictions. MEMIT (Meng et al., 2023) and PMET (Li et al., 2025) show how controlled inferences can be built by manipulating models' components. Some nuance emerges in later-layer behaviours. DecompX (Modarressi et al., 2023) traced token representations layer-by-layer and observed partial shifts toward integration. Yu and Ettinger (2020) tested model encoding and found that transformers mainly encode individual word content rather than true phrase-level meaning. While some models appear more compositional under certain conditions, general trends remain unclear. For example, Dankers et al. (2022) demonstrate that models can show unexpectedly high or low compositionality depending on the data and task, suggesting exposure and framing affect outcomes as much as architecture. Petty et al. (2024) show that deeper Transformers tend to generalise more compositionally than shallower ones, though the benefits diminish beyond a certain depth. This highlights that architectural depth, not just scale, may shape compositional ability, though with diminishing returns. In multi-step reasoning tasks, models often fall back on shallow pattern matching rather than true decomposition (Dziri et al., 2023).

Prior work has primarily relied on synthetic tasks to assess compositional generalisation, focusing on properties such as systematicity, productivity, and substitutivity (Hupkes et al., 2020; Lake and Baroni, 2018), these setups often abstract away from the complexities of natural language. More recent studies using natural data are often limited to small domains such as semantic parsing or machine translation (Lake and Baroni, 2018; Kim and Linzen, 2020), and typically lack insight into internal representations.

In contrast to prior works focused on final outputs or synthetic tasks, CAP is a method for probing compositional structure within LLMs using real inputs. It intervenes directly on hidden activations, merging token-level representations into word- or phrase-level constituents at various depths. This allows us to evaluate where semantic composition occurs and how robust LLMs to structured perturbations. Unlike surface-level probes, CAP provides a targeted, activation-level lens on how meaning is constructed and distributed across model layers and linguistic units.

7 Conclusion

This work systematically analyses the robustness of transformer-based LLMs to compositional perturbations. Motivated by studies highlighting an unexpected gap between linguistic compositionality and LLM representations, we characterised the impact of compositional aggregation at each inference step and provided an information-theoretical explanation. Our findings indicate a pattern where token dependencies are modelled by aggregation paths spanning multiple layers, and complex token structure learning comes at the cost of higher sensitivity to perturbations at inputs and earlier layers. Based on the relation between information gain from input to predicted token and mutual information between token representations, we postulate that compositional semantic representations cannot be isolated to any particular (intermediate) stage of a standard transformer model. These insights suggest that future compositional-aware models should explore specialised architectures or training objectives. Natural extensions include analysing encoder-based and encoder-decoder transformers and investigating final token representations to further understand internal compositional mechanisms.

Limitations

Several limitations are acknowledged in our paper. First, the WordNet dataset may not fully represent language diversity across all domains. Second, the employed transformer models are decoder-based only and could be subject to biases from their training data. Third, our findings depend on the Benepar parsing model, which may introduce inaccuracies in linguistic analysis. Additionally, while our tasks provide an indirect signal of meaning preservation, incorporating explicit reconstruction tasks in future work could offer complementary insight into how CAP affects the retention of input-level information. Finally, the applicability of our results to other languages has not been tested. Expanding CAP to multilingual settings and testing with alternative parsers or models trained with different positional encodings would further validate the generality of our findings.

Ethical Statement

The proposed framework aims to have a positive impact on improving the critical understanding of the mechanisms involved in language interpretation in transformers. A more complete understanding of these mechanisms requires coordination with other interpretability methods.

References

Danilo S Carvalho, Edoardo Manino, Julia Rozanova, Lucas Cordeiro, and André Freitas. 2025. Montague semantics and modifier consistency measurement in neural language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5515–5529.

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.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Lucia Donatelli and Alexander Koller. 2023. Compositionality in computational linguistics. *Annual Review of Linguistics*, 9(Volume 9, 2023):463–481.

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

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *MIT Press google schola*, 2:678–686.

JA Fodor. 1975. The language of thought.

James Fodor, Simon De Deyne, and Shinsuke Suzuki. 2024. Compositionality and Sentence Meaning: Comparing Semantic Parsing and Transformers on a Challenging Sentence Similarity Dataset. *Computational Linguistics*, pages 1–52.

Gottlob Frege. 1892. Über sinn und bedeutung [on sense and reference]. *Zeitschrift für Philosophie Und Philosophische Kritik*, 100:25–50.

David A. Haslett. 2024. How much semantic information is available in large language model tokens? Preprint available on OSF.

Zhibo Hu, Chen Wang, Yanfeng Shu, Hye-Young Paik, and Liming Zhu. 2024. Prompt perturbation in retrieval-augmented generation based large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1119–1130.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? (extended abstract). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 5065–5069. International Joint Conferences on Artificial Intelligence Organization. Journal track.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press.