The Role of Compositionality in LLM Performance: A Case Study in Adjective-Noun Combinations

Anonymous ACL submission

Abstract

Compositionality is considered central to human language abilities, but is it also important for the performance of large language models (LLMs)? We investigate this with a case study of adjective–noun compositionality. First, we test whether two common factors considered to boost overall LLM performance, scaling and instruction tuning, also enhance compositional behavior across three tasks. We then conduct a layer-wise representational analysis to examine whether internal representations support compositional task performance. Our findings challenge the view that compositionality drives performance in LLMs. However, compositional structure emerges reliably within representations and suggests its role as a stable representational scaffold in models.

1 Introduction

Compositionality —the ability of systems to generate complex meanings from simpler parts— is a key factor in human language processing (Phillips, 2018; Zuberbühler, 2020; Ellison and Reinöhl, 2024) and cognition (De Beule and Bergen, 2006; Reverberi et al., 2012; Sumers et al., 2023). Today, large language models (LLMs) are considered highly performant systems with superior natural language abilities (Min et al., 2023; Zhang et al., 2024b). With the rise of LLMs, debates on compositional understanding of models have again risen to the forefront (Zečević et al., 2023; Saba, 2023; Chen et al., 2024; Dziri et al., 2024; McCurdy et al., 2024): if compositionality is key to language, does it also play a central role in how LLMs achieve their impressive performance?[1]

While earlier literature has often disagreed on compositional abilities of neural networks (Fodor and Pylyshyn, 1988; Smolensky, 1987), there seems to be one point of agreement: if compositionality plays a key role in any system, then it should ideally learn compositional strategies as a way to improve performance, i.e., more performant models should also be more compositional (Fodor and Pylyshyn, 1988; Smolensky, 1987; Van Gelder, 1990; Chalmers, 1993). In this work, we aim to understand whether compositionality plays a role in the performance of LLMs with a case study in the domain of adjective-noun compositionality. We identify two strategies considered to typically improve performance- scaling (Kaplan et al., 2020; Hoffmann et al., 2022) and instruction tuning (Wei et al., 2021; Ouyang et al., 2022). First, we investigate whether such scaled or tuned models are also more compositional by a behavioural analysis of models on 3 relevant tasks. Second, we conduct a representational analysis of models to determine whether the representations are useful in task performance and how this varies across model variants.

We discuss our findings, which suggest that compositionality, while not central to performance, may still serve as a latent but meaningful part of how these models internally represent language.

2 Related Work

Compositionality is a long-standing concept in linguistics and cognitive science (Frege, 1892; Partee et al., 1995). With the rise of connectionism in the late twentieth century, debates centered on whether compositionality was significant for the working of neural networks (Fodor and Pylyshyn, 1988; Van Gelder, 1990; Smolensky, 1987; Chalmers, 1993; Aizawa and Aizawa, 2003). Early empirical work has investigated compositional abilities in neural models trained on compositional tasks like SCAN (Lake and Baroni, 2018), COGS (Kim and Linzen, 2020a), or PCFG (Hupkes et al., 2020). However, the earlier tests for neural models are a

[1] Non-compositional LLMs have been argued to at least be conceivable (Block, 1981; Bender et al., 2021).

poor fit for modern LLMs for the following reasons: a) they are based on a train-test paradigm that is not suitable for testing pretrained LLMs, and b) many of them are based on excerpts of English language (Lake and Baroni, 2018; Kim and Linzen, 2020b; Hupkes et al., 2020), which already forms a part of pretraining most LLMs today.

More recently, there has been a multitude of empirical work in the investigation of "compositional abilities" of LLMs (SHAO et al., 2023; Kauf et al., 2024; Lu et al., 2024; Xu et al., 2024; Zhang et al., 2024a; Dziri et al., 2024). Most of these current evaluations, however, often conflate compositionality with multi-hop reasoning or simple combinatorial tasks, which do not reflect the cognitive richness of true compositional meaning construction (Fodor and Pylyshyn, 1988; Smolensky, 1987). Also, most such work, to the best of our knowledge, does not investigate how significant compositional knowledge is for improving general model performance on tasks, which is our focus in this study.

A growing line of research has also investigated compositionality in neural models by probing their internal representations. (Hewitt and Manning, 2019) probe representations for syntactic compositionality, (Murty et al., 2023) introduces tree projection methods to characterize intrinsic compositional structure in transformers, and (Kumon and Yanaka, 2025) analyze internal mechanisms that support compositional generalization. Such work demonstrates that linguistic properties can often be detected in hidden states, even when they are not directly reflected in model outputs. Our study differs in focus: rather than probing for the presence of compositional structure inside models, we want to ask whether these representations are useful for performing compositional tasks. Such a shift would allow us to assess not whether compositionality exists in LLMs' internals, but whether it contributes to the kinds of generalization behaviors that matter for the compositional task at hand.

This gap between the presence of compositionality and the lack of clarity on its practical relation to LLM performance motivates our investigation.

3 Experimental Setup

To study compositionality, we focus on the domain of adjective–noun (AN) phrases. This choice is motivated both linguistically and practically. From a linguistic perspective, AN constructions are among the most frequent and semantically transparent instances of phrasal composition, making them a classic testbed for compositional semantics (Partee et al., 1995; Kamp and Partee, 1995; Guevara, 2010; Hartung et al., 2017; Bertolini et al., 2021). Crucially, meaning composition in this domain can be operationalized through inference: if the composed phrase preserves or alters entailment relations with its parts, the model must reason about semantic structure rather than rely on lexical associations (Lebowitz, 1983; Geiger et al., 2019; Donatelli and Koller, 2023). This inference-based framing has been central in both formal semantics (Partee et al., 1995) and recent NLP evaluations of compositionality (Pavlick and Callison-Burch, 2016; Bertolini et al., 2022), and it motivates our focus on tasks that explicitly test whether models maintain meaning under composition.

Compositionality in natural language can manifest in many forms, such as substitutivity, systematicity, and overgeneralization (Hupkes et al., 2020). We adopt one adjective-noun (AN) based task on meaning inference for each of these aspects: *Substitutivity* tested with the **AddOne Dataset** (Pavlick and Callison-Burch, 2016), *Systematicity* tested with the **PLANE Dataset** (Bertolini et al., 2022), and *Overgeneralization* tested with a handcrafted toy dataset—the **COMPCOMB Dataset**, which we introduce in this work (more details in Appendix C). For more details on the three aspects and how the tasks represent them, refer to Appendix A [2].

3.1 Behavioural Analysis

We consider two factors typically considered to boost performance of LLMs: **scaling parameters**, where LLM scaling laws show that model performance get better with size i.e an increase in the number of parameters (Kaplan et al., 2020; Hoffmann et al., 2022; Biderman et al., 2023; Chowdhery et al., 2023), and **instruction tuning**, which has been shown to improve general performance of LLMs, especially for generalization to unseen tasks and alignment with human behaviour (Wei et al., 2022; Ouyang et al., 2022; Longpre et al., 2023; Chung et al., 2024; Sun et al., 2024). If compositionality plays a role in model performance, larger and instruction-tuned models should also be more compositional and perform better in the three relevant compositional tasks.

[2] We will release all code and task data upon publication.

	LLaMA2			CodeLLaMA			Qwen2.5-Coder			Gemma2		
Dataset	**7B**	**7B-it**	**13B**	**7B**	**7B-it**	**13B**	**7B**	**7B-it**	**14B**	**2B**	**2B-it**	**9B**
AddOne (F1)	77.97	18.01	78.49	78.49	75.36	54.92	77.58	20.87	12.62	78.49	29.61	59.76
PLANE (F1)	36.70	33.73	35.35	33.33	42.67	33.33	33.33	34.57	38.14	33.33	33.33	36.84
CompComb (Acc)	66.67	72.59	45.93	95.56	71.11	93.33	77.04	73.34	71.11	77.04	73.33	45.19

Table 1: The results of all models on Addone, PLANE, and CompComb datasets.

Methodology We conduct evaluation across 4 families of models— `LLama-2` , `Codellama`, `Qwen2.5-Coder`, and `Gemma2`– with 3 variants of base, instruction-tuned and scaled model in each family (model details in Appendix B). For AddOne and PLANE, we use a zero-shot prompting and log proability setup. Accuracy is based on which completion receives higher likelihood calculated accordingly. For COMPCOMB, we report average accuracies across three prompting setups.

Results Across the three tasks, we observe distinct and revealing trends in how compositional performance changes with instruction tuning and model scale. As shown in Table 1 and Figure 1, we observe that performance trends differ significantly across the 3 tasks—for AddOne and COMPCOMB, compositional performance declines consistently with both instruction tuning and increased model size while for PLANE, scaling and instruction tuning don't show significant change in compositional abilities. This non-monotonic relationship suggests that the mechanisms driving general improvements in LLMs may not align with those required for robust compositional generalization. Thus, compositionality may not play a central role in driving overall performance gains in LLMs.

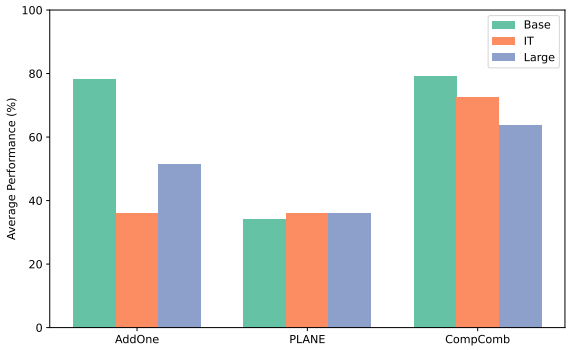


Figure 1: The average performance across different model category (Base, Instruction Tuning, and Large model size) on three tasks. We report the weighted F1 score on AddOne and PLANE, and Accuracy on CompComb.

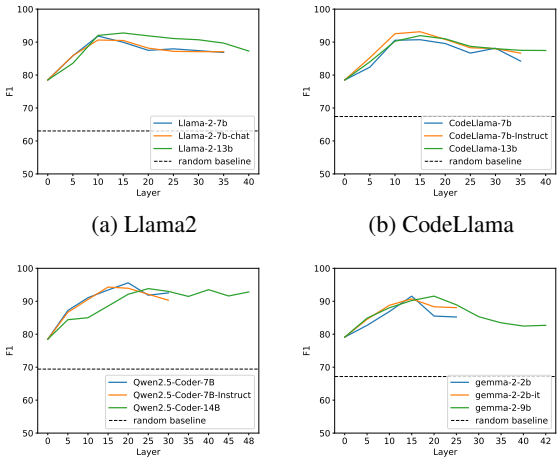
3.2 Representational Analysis

Earlier work has shown that external performance and internal representations can be disparate in LLMs (Lampinen et al., 2024; Orgad et al., 2025). We conduct a layer-wise analysis across multiple models to examine where compositional signals arise in the network and whether their presence aligns with improvements in behavioral performance on our tasks. This allows us to test if compositionality functions as a representational scaffold within LLMs, even when it does not translate into stronger task-level generalization.

Methodology For each model evaluated in subsection 3.2, we extract hidden representations from every fifth transformer layer—i.e., layers 0, 5, 10, 15, 20, etc.—depending on the depth of the model. This sampling allows us to efficiently capture trends across the entire architecture. For tasks with discrete output labels (AddOne and PLANE), we train a simple linear classifier (logistic regression) on representations extracted from each layer to predict the correct label (e.g., entailment, non-entailment). We train it for three epochs using a batch size of 8. We use the Adam optimizer with a learning rate of 1e-3. For COMPCOMB, which involves analyzing cosine distances between token-level embeddings, we compute pairwise similarity metrics directly from multiple layers' hidden states. For all experiments, we further include a random baseline (where representations are randomly initialised) for comparison.

Results *First,* representations perform well above chance baseline (dashed line), showing that compositional information is inherently encoded. This signal remains stable across base, instruction-tuned, and scaled models—neither enhanced nor disrupted by them (in Figure 2,3,4). *Second,* for AddOne and PLANE, compositional knowledge peaks in intermediate layers, suggesting a key role for these

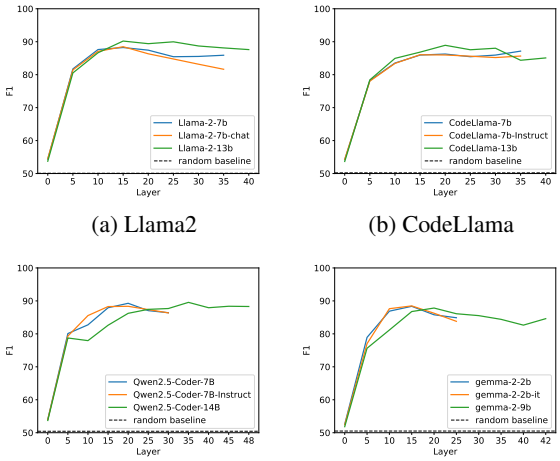
layers in encoding abstractions. Overall, compositionality emerges early and consistently across model variants, serving as a scaffold for organizing linguistic knowledge even if it does not directly drive performance.



(a) Llama2 (b) CodeLlama

(c) Qwen2.5-Coder (d) Gemma2

Figure 2: Layer-wise results (weighted F1 score) on AddOne dataset.



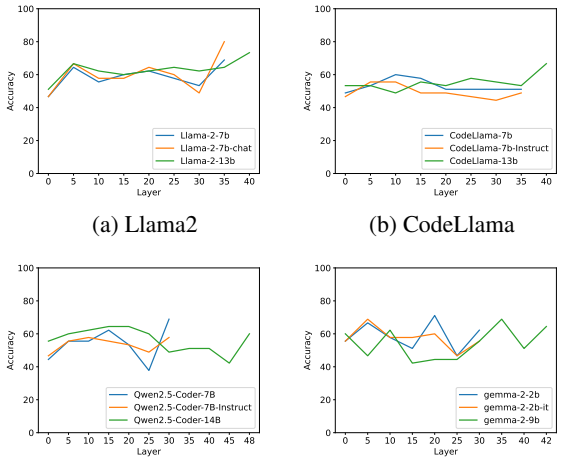
(a) Llama2 (b) CodeLlama

(c) Qwen2.5-Coder (d) Gemma2

Figure 3: Layer-wise results (weighted F1 score) on PLANE dataset.

4 Discussion

Our results paint a nuanced picture of the role compositionality plays in large language models. Scaling and instruction tuning, two key drivers of LLM performance, do not consistently improve compositional behavior. This suggests that compositionality is not a primary mechanism behind performance gains, echoing long-standing concerns about its limits in connectionist systems(Russin et al., 2024a; Zhao et al., 2024; McCurdy et al., 2024). Yet our representational analysis reveals that compositional structure emerges reliably and stably across models, especially in intermediate layers which have also been previously shown to encode rich linguistic abstractions (Tenney et al., 2019; Liu et al., 2020) and strongly align with neural activity observed in human language processing (Antonello and Cheng, 2024; Cheng et al., 2025). This reinforces the view that compositionality is important in how models represent language. In this sense, compositionality does play a role in LLMs—not for directly optimizing task-level performance—but at the very least for shaping how such models internalize and "understand" language. Compositionality may not be what matters most for LLM performance, but that doesn't mean it doesn't matter at all.



(a) Llama2 (b) CodeLlama

(c) Qwen2.5-Coder (d) Gemma2

Figure 4: Layer-wise results (accuracy) on CompComb dataset.

This gap between internal structure and external behavior also raises an open question for future work: can compositional representations be leveraged to improve generalization? Future work may explore approaches such as in-context learning or targeted fine-tuning on compositional data (Li et al., 2023; Russin et al., 2024b; Han and Padó, 2024), potentially aligning models' internal organization with task-level performance. Such interventions could clarify whether compositionality can be made functionally relevant, while also advancing toward models whose reasoning is more interpretable and cognitively grounded (Piantadosi et al., 2016; Martin, 2020).

Limitations

There are some limitations that future work should address to deepen and broaden this understanding. *First*, our evaluation is limited in scope and focuses only on the domain of adjective-noun compositionality. Moreover, we cover a limited set of model families, which are unilingual. Broader studies with sub-word or sentence-level compositionality and multilingual models would enable more generalizable conclusions on the overall role of compositionality in LLMs.

Second, while we identify where compositional signals emerge within models, we do not establish whether these representations play a causal role in inference. It remains unclear whether how these representations signals actively contribute to generalization. Future work should apply causal probing techniques—such as activation patching or path attribution—to determine whether and how compositional representations are functionally utilized during inference.

Last, our findings show that compositionality is not necessary for strong general performance—but we do not directly assess whether it can be functionally advantageous in settings where generalization is more challenging. Tasks that require multi-step reasoning or logical inference may depend more heavily on structured representations. Exploring whether models with better compositional behaviour also perform better in such settings would help clarify whether compositionality is a practical asset in such cases. Understanding when and where it matters will be key to evaluating its broader role in future model development.

Ethical Considerations

We do not anticipate any risks in our work. Any data used does not contain any personal details or offensive content. In this study, our use of all artifacts is consistent with the license and intended use of all datasets and models. The AddOne Dataset (Pavlick and Callison-Burch, 2016) is available on HuggingFace onthis page and is free for all use. The PLANE dataset (Bertolini et al., 2022) is also available on HuggingFace onthis page and has license of Creative Commons Attribution 2.0. Our toy dataset COMPCOMB will be released upon publication along with details of license and intended use. All models used are also from HuggingFace and their details, along with licenses, are summarized in Appendix B. All code and data will be released upon publication for greater reproducibility.

References

Kenneth Aizawa and Kenneth Aizawa. 2003. The systematicity arguments applied to connectionism. *The Systematicity Arguments*, pages 151–174.

Richard Antonello and Emily Cheng. 2024. Evidence from fMRI supports a two-phase abstraction process in language models. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Lorenzo Bertolini, Julie Weeds, and David Weir. 2022. Testing large language models on compositionality and inference with phrase-level adjective-noun entailment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4084–4100.

Lorenzo Bertolini, Julie Weeds, David Weir, and Qiwei Peng. 2021. Representing syntax and composition with geometric transformations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3343–3353, Online. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Ned Block. 1981. Psychologism and behaviorism. *Philosophical Review*, 90(1):5–43.

Rudolf Carnap. 1988. *Meaning and necessity: A study in semantics and modal logic*, volume 30. University of Chicago Press.

David J Chalmers. 1993. Connectionism and compositionality: Why fodor and pylyshyn were wrong. *Philosophical Psychology*.

Yulong Chen, Yang Liu, Jianhao Yan, Xuefeng Bai, Ming Zhong, Yinghao Yang, Ziyi Yang, Chenguang Zhu, and Yue Zhang. 2024. See what LLMs cannot answer: A self-challenge framework for uncovering LLM weaknesses. In *First Conference on Language Modeling*.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. 2025. Emergence of a high-dimensional abstraction phase in language transformers. In *The Thirteenth International Conference on Learning Representations*.

Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Joachim De Beule and Benjamin K Bergen. 2006. On the emergence of compositionality. In *The Evolution of Language*, pages 35–42. World Scientific.

Lucia Donatelli and Alexander Koller. 2023. Compositionality in computational linguistics. *Annual Review of Linguistics*, 9(1):463–481.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, and 1 others. 2024. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36.

T Mark Ellison and Uta Reinöhl. 2024. Compositionality, metaphor, and the evolution of language. *International Journal of Primatology*, 45(3):703–719.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Gottlob Frege. 1892. On sense and reference. *Translations from the philosophical writings of Gottlob Frege*, 2:56–85.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. Posing fair generalization tasks for natural language inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4485–4495, Hong Kong, China. Association for Computational Linguistics.

Emiliano Raul Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 workshop on geometrical models of natural language semantics*, pages 33–37.

Sungjun Han and Sebastian Padó. 2024. Towards understanding the relationship between in-context learning and compositional generalization. *arXiv preprint arXiv:2403.11834*.

Matthias Hartung, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano. 2017. Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 54–64.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Hans Kamp and Barbara Partee. 1995. Prototype theory and compositionality. *Cognition*, 57(2):129–191.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Comparing plausibility estimates in base and instruction-tuned large language models. *arXiv preprint arXiv:2403.14859*.

Najoung Kim and Tal Linzen. 2020a. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Najoung Kim and Tal Linzen. 2020b. Cogs: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 9087–9105.

Ryoma Kumon and Hitomi Yanaka. 2025. Analyzing the inner workings of transformers in compositional generalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the*

Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8529–8540, Albuquerque, New Mexico. Association for Computational Linguistics.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Andrew Kyle Lampinen, Stephanie C.Y. Chan, and Katherine Hermann. 2024. Learned feature representations are biased by complexity, learning order, position, and more. *Transactions on Machine Learning Research*.

Michael Lebowitz. 1983. Generalization from natural language text. *Cognitive Science*, 7(1):1–40.

Yingcong Li, Kartik Sreenivasan, Angeliki Giannou, Dimitris Papailiopoulos, and Samet Oymak. 2023. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. *Advances in Neural Information Processing Systems*, 36:22021–22046.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and 1 others. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2024. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36.

Andrea E Martin. 2020. A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, 32(8):1407–1427.

Kate McCurdy, Paul Soulos, Paul Smolensky, Roland Fernandez, and Jianfeng Gao. 2024. Toward compositional behavior in neural models: A survey of current views. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9323–9339, Miami, Florida, USA. Association for Computational Linguistics.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D Manning. 2023. Characterizing intrinsic compositionality in transformers with tree projections. In *The Eleventh International Conference on Learning Representations*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Peter Pagin and Dag Westerståhl. 2010a. Compositionality ii: Arguments and problems. *Philosophy Compass*, 5(3):265–282.

Peter Pagin and Dag Westerståhl. 2010b. Pure quotation and general compositionality. *Linguistics and Philosophy*, 33:381–415.

Barbara Partee and 1 others. 1995. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany. Association for Computational Linguistics.

Steven Phillips. 2018. What underlies dual-process cognition? adjoint and representable functors. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 40.

Steven T Piantadosi, Joshua B Tenenbaum, and Noah D Goodman. 2016. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123(4):392.

Geoffrey K Pullum and Barbara C Scholz. 2010. Recursion and the infinitude claim. In *Recursion and human language*, pages 113–137. De Gruyter Mouton.

Carlo Reverberi, Kai Görgen, and John-Dylan Haynes. 2012. Compositionality of rule representations in human prefrontal cortex. *Cerebral cortex*, 22(6):1237–1246.

Jacob Russin, Sam Whitman McGrath, Danielle J Williams, and Lotem Elber-Dorozko. 2024a. From frege to chatgpt: Compositionality in language, cognition, and deep neural networks. *arXiv preprint arXiv:2405.15164*.

Jacob Russin, Ellie Pavlick, and Michael J Frank. 2024b. Curriculum effects and compositionality emerge with in-context learning in neural networks. *ArXiv*, pages arXiv–2402.

Walid S Saba. 2023. Stochastic llms do not understand language: towards symbolic, explainable and ontologically based llms. In *International conference on conceptual modeling*, pages 3–19. Springer.

NAN SHAO, Zefan Cai, Hanwei xu, Chonghua Liao, Yanan Zheng, and Zhilin Yang. 2023. Compositional task representations for large language models. In *The Eleventh International Conference on Learning Representations*.

Paul Smolensky. 1987. Connectionist ai, symbolic ai, and the brain. *Artificial Intelligence Review*, 1(2):95–109.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. 2023. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2024. Evaluating the zero-shot robustness of instruction-tuned language models. In *The Twelfth International Conference on Learning Representations*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Tim Van Gelder. 1990. Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, 14(3):355–384.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Zhuoyan Xu, Zhenmei Shi, and Yingyu Liang. 2024. Do large language models have compositional ability? an investigation into limitations and scalability. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*.

Min Zhang, Jianfeng He, Shuo Lei, Murong Yue, Linhan Wang, and Chang-Tien Lu. 2024a. Can llm find the green circle? investigation and human-guided tool manipulation for compositional generalization. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11996–12000. IEEE.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA. Association for Computational Linguistics.

Jun Zhao, Jingqi Tong, Yurong Mou, Ming Zhang, Qi Zhang, and Xuanjing Huang. 2024. Exploring the compositional deficiency of large language models in mathematical reasoning through trap problems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16361–16376, Miami, Florida, USA. Association for Computational Linguistics.

Klaus Zuberbühler. 2020. Syntax and compositionality in animal communication. *Philosophical Transactions of the Royal Society B*, 375(1789):20190062.

A Tasks and Datasets

In this appendix section, we provide a detailed explanation of the different aspects of compositionality we evaluate (Hupkes et al., 2020) along with the task used.

Substitutivity: It refers to a model's ability to preserve meaning when a component of a complex expression is replaced with a semantically equivalent or synonymous alternative (Carnap, 1988; Hupkes et al., 2020). In a truly compositional system, meaning should be a function of the meanings of its parts and their combination – so substituting one part with a synonym should not alter the overall meaning of the expression. We evaluate this using the **AddOne Dataset** (Pavlick and Callison-Burch, 2016), which tests whether models recognize entailment relationships between original and modified sentences. For example, given the sentence *The runner set a record*, we substitute the noun *record* with an adjective–noun phrase such as *new record*, producing the modified sentence *The runner set a new record*. While the surface form changes, the core meaning remains unchanged in context. A model with compositional understanding should therefore recognize the entailment:

$$\textit{The runner set a new record} \models \textit{The runner set a record.}$$

Success on this task requires more than detecting synonymy at the word level – it requires the model to integrate the substitution compositionally and infer that the meaning of the larger sentence remains intact. The dataset is released with train (4481 samples), val (387 samples), and test (510 samples) splits.

Systematicity: It refers to a model's ability to recombine known parts and rules to interpret novel expressions (Fodor and Pylyshyn, 1988; Hupkes et al., 2020). We evaluate this using the **PLANE Dataset** (Bertolini et al., 2022), which focuses on adjective–noun (AN) entailment patterns. In this task, the model is provided with two base entailments: one between an adjective–noun phrase and the noun alone (e.g., *red car* $\models$ *car*), and one between the noun and its hypernym (*car* $\models$ *vehicle*). The model is then asked to determine the validity of the composed entailment (*red car* $\models$ *red vehicle*). The dataset is released with train (60012 samples) and test (2016 samples) splits.

To succeed, the model must apply *systematic compositional reasoning*: it must combine the entailment patterns it has already observed to infer the new one. Crucially, this requires understanding how the adjective interacts with both the noun and the hypernym class – particularly when adjectives vary in type (e.g., intersective vs subsective vs intensional). The task thus tests whether the model can generalize meaning compositionally, rather than relying on shallow heuristics or memorized patterns.

Over-generalization: It refers to a model's tendency to incorrectly apply compositional rules to expressions that are, in fact, non-compositional (Hupkes et al., 2020). A robust compositional system should be able to distinguish between genuine compositional combinations and compounds that do not follow regular meaning composition. We test this with a novel task using a handcrafted toy dataset of 45 samples in English language—the **COMPCOMB Dataset**, which we introduce in this work. Each data point consists of a triple: a noun, a compositional adjective–noun (AN) combination involving that noun, and a non-compositional exocentric compound that includes the same noun. For example, in the triple *(coat, trenchcoat, turncoat)*, the word *trenchcoat* refers to a specific type of *coat*, preserving a compositional relationship. In contrast, *turncoat* is an idiomatic expression referring to a traitor, where the meaning does not transparently derive from *coat*.

To succeed, the model must avoid overgeneralizing based on surface forms or string overlap, and instead distinguish between compositional and non-compositional expressions. This task directly tests whether the model can identify and respect the boundaries of compositional generalization. Please note that we will release the dataset upon publication along with the license and intended use for research purposes.

It is important to note that two other dimensions as per Hupkes et al. (2020)—*productivity* and *localism*—are deliberately excluded from our analyses, for reasons rooted in both linguistic theory and the epistemological limitations of empirical verification which are commonly recognized in the literature (Hupkes et al., 2020).

First, productivity—the notion that natural language allows for the generation of an unbounded number of novel expressions from a finite set of primitives—is a foundational concept in generative linguistics (Chomsky, 1956). However, its empirical status is more controversial. As noted by Pullum and Scholz (2010), it is not possible to establish the infinitude of natural language through observation, given the finiteness of human utterances and cognitive resources. Moreover, even if a system generalizes to longer sequences than it was trained on, this does not, in itself, constitute evidence for true linguistic productivity. In formal terms, productivity involves generalization across an unbounded domain:

$$\forall n \in \mathbb{N},\ \exists\, s_n \in \mathcal{L} \text{ such that } \text{length}(s_n) = n,$$

where $\mathcal{L}$ is the set of grammatical sentences. Demonstrating this for either human language or LLM behavior is inherently intractable. Consequently, productivity lacks the falsifiability and granularity required for robust evaluation in our setting.

Second, we exclude localism as a defining criterion of compositionality. Localism refers to the idea that composition should occur via strictly local operations—i.e., the meaning of a complex expression is derived solely from the meanings of its immediate subparts (its "children" in a compositional tree). This notion is aligned with what the relevant linguistic literature Pagin and Westerståhl (2010a,a,b) terms *strong* or *first-level* compositionality. However, this is only one possible interpretation. Under *weak* or *global* compositionality, meanings may be computed from broader structures, potentially referencing distant context or higher-order structures. Crucially, both modes of composition satisfy the formal definition of compositionality:

$$[\![E]\!] = f([\![E_1]\!], \ldots, [\![E_n]\!]),$$

where $[\![E]\!]$ denotes the meaning of expression E, and $E_1, \ldots, E_n$ are its constituents. The nature of

f—whether strictly local or globally informed—is not constrained by the principle of compositionality itself. Empirically, the absence of local compositional operations does not entail non-compositional behavior. A model may validly construct meanings using global strategies and still preserve the structural integrity of compositional mappings. Thus, testing for localism does not test for compositionality per se, but rather for a specific implementation strategy.

For these reasons, we limit our analysis to aspects of compositionality that are both theoretically robust and empirically interpretable. We regard productivity and localism as valuable extensions for future work, but not as necessary conditions for evaluating whether compositionality matters in current large language models.

B Model Details

In this appendix section, we provide details of the models used in our experiments. For each model family, we selected three models for testing:

- Base Model (Base): A minimum parameter model, ranging from 2B to 7B, that serves as the foundational model version for each family.
- Instruction – Finetuned Model (IFT): The same 7B base model, further fine-tuned with instruction-tuning to enhance task performance.
- Scaled Model (Large): A model variant with a higher parameter count, ranging from 9B to 70B, depending on availability within each family. These larger models allow us to investigate how scaling affects compositional behavior.

The diversity in models ensures that our analysis captures how both model complexity and tuning approaches impact compositionality. The names and links to each model are in Table 2

C Methodology Details

In this appendix section, we provide further details of the methodology for experiments.

For subsection 3.1, in **AddOne** and **PLANE** datasets, we evaluate with a comparative log probability setup. The evaluation involves computing the model's log probabilities for two possible completions: one being the correct option and the other the control (incorrect option). The model's preference between the two completions is determined by comparing their log probabilities and the setup focuses on the model's probabilistic confidence in its outputs. The completion with the higher log probability is considered indicative of the model's judgement. For the two datasets here, which involve binary classification tasks, the two completions correspond to entailment and non-entailment outcomes. We report results on the test splits. For the classifier training, we use their provided train splits. For **COMPCOMB**, we use a prompting setup for the behavioural analysis.

For subsection 3.2, in **AddOne** and **PLANE** datasets we use the linear regression task setup as mentioned in subsection 3.2. More specifically for **COMPCOMB**, we examine how well the model's internal representations encode the relationships between related and unrelated adjective-noun and exocentric compound-noun pairs. To evaluate, we extract data from multiple layers. For each layer, we get representations of the model for each word in the triple, and the model is considered to be accurate if its representations for noun and adjective-noun combinations are closer than the noun and semantically unrelated exocentric compound representations. This gives insights into how well the model captures semantic relationships and distinctions between input items (such as distinguishing between a noun and its related and unrelated compounds). This setup allows for an analysis of the model's ability to differentiate semantically related pairs from unrelated ones based on internal representation quality.

Family	Variant	Hugging Face Link	License
LLaMA-2	Base (7B)	Llama-2-7b-hf	LLaMA 2 Community License
	Instruction-tuned (7B)	Llama-2-7b-chat-hf	LLaMA 2 Community License
	Scaled (13B)	Llama-2-13b-hf	LLaMA 2 Community License
CodeLlama	Base (7B)	CodeLlama-7b-hf	LLaMA 2 Community License
	Instruction-tuned (7B)	CodeLlama-7b-Instruct-hf	LLaMA 2 Community License
	Scaled (13B)	CodeLlama-13b-hf	LLaMA 2 Community License
Qwen2.5-Coder	Base (7B)	Qwen2.5-Coder-7B	Apache License 2.0
	Instruction-tuned (7B)	Qwen2.5-Coder-7B-Instruct	Apache License 2.0
	Scaled (14B)	Qwen2.5-Coder-14B	Apache License 2.0
Gemma2	Base (2B)	Gemma-2-2B	Gemma Terms of Use
	Instruction-tuned (2B)	Gemma-2-2B-it	Gemma Terms of Use
	Scaled (9B)	Gemma-2-9B	Gemma Terms of Use

Table 2: Models used in section 3, grouped by family and variant, with links to their Hugging Face repositories and associated licenses.