From Words to Worlds: Compositionality for Cognitive Architectures

Ruchira Dhar [1]   **Anders Søgaard** [1]

Abstract

Large language models (LLMs) are very performant connectionist systems, but do they exhibit more compositionality? More importantly, is that part of why they perform so well? We present empirical analyses across four LLM families (12 models) and three task categories, including a novel task introduced below. Our findings reveal a nuanced relationship in learning of compositional strategies by LLMs – while scaling enhances compositional abilities, instruction tuning often has a reverse effect. Such disparity brings forth some open issues regarding the development and improvement of large language models in alignment with human cognitive capacities.

1. Introduction

Compositionality is a widely studied aspect of human cognition. Fodor & Pylyshyn (1988) claimed that non – symbolic connectionist representations were inadequate for compositional understanding. The question turns on whether compositionality is acquired (Smolensky, 1987; Chalmers, 1993), or whether compositionality is merely a functional property (Van Gelder, 1990). Symons & Calvo (2014), building on the arguments against connectionism in (Fodor & Pylyshyn, 1988), argue that even if connectionist systems can stumble across an implementation of compositionality, this does not *explain* systematicity in their behaviour nor does it render them suitable cognitive architectures (see Appendix A for related work and Appendix B for further details). To serve as models of cognition or "cognitive architectures", connectionist systems should ideally:

i) be compositional, i.e., have compositional representations and behaviour.

ii) be compositional in a way that explains their behavior

and performance, i.e., they should learn compositional strategies as a way to improve performance.
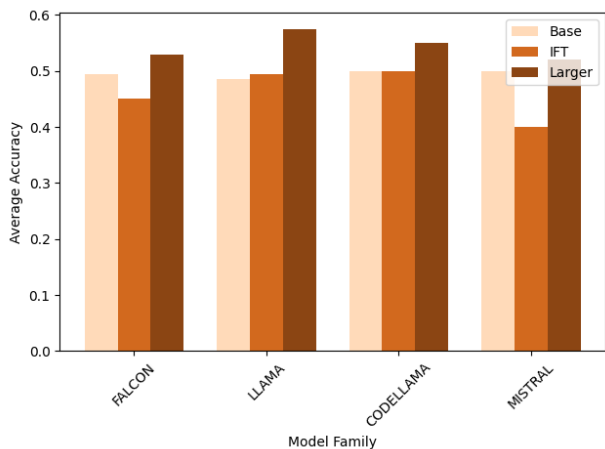

Figure 1. Model Accuracy trends for two setups (combined) with the ANTAILS Dataset.

LLMs are now increasingly seen as possible models of human language (Mahowald et al., 2024; Hu et al., 2024) or cognition (Kauf et al., 2023; Hardy et al., 2023; Marjieh et al., 2023; Lamprinidis, 2023; Aw et al., 2023), and it is therefore crucial to review the Fodor & Pylyshyn (1988) challenge from the perspective of LLMs. While there is work on measuring compositional abilities of LLMs (Dziri et al., 2024; Li et al., 2024; Zhang et al., 2024; Wang et al., 2024), our focus is not to benchmark models for compositionality, but to examine its explanatory value in predicting performance and validating models as cognitive architectures.

Scaling and instruction tuning are widely assumed to improve model alignment and generalization performance across a multitude of tasks ranging from natural language inference and textual entailment (Wei et al., 2022) to MMLU and BigBench (Longpre et al., 2023) – but are these improved performances a result of improved compositionality? Focusing on the domain of adjective – noun (Adj – N) composition, we propose three task types that can evaluate different aspects of compositional behaviour in LLMs and

[1]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark. Correspondence to: Ruchira Dhar <rudh@di.ku.dk>.

consider the impact of model size and instruction tuning on the compositional behaviour of such models. Finally, we discuss the importance of compositionality as a theoretical construct in validating connectionist cognitive architectures.

2. Measuring Compositionality

Over the years, several benchmarks have been developed to test compositionality of neural network models – SCAN (Lake & Baroni, 2018), Lookup Table Composition (Liska et al., 2018), COGS (Kim & Linzen, 2020), and PCFG Set (Hupkes et al., 2020). However, we face a few issues when trying to leverage such datasets for testing today's LLMs-models pretrained on large amounts of text:

a) They are based on a train – test paradigm that is not easily applicable pretrained LLMs.

b) LLMs are trained on very large quantities of texts and may, as a consequence, have seen the test set expressions before.

c) There is no congruence on what aspects of compositionality (Sun et al., 2023b) we test with these methods.

Some work considers compositional multi-hop reasoning (Lu et al., 2024; Xu et al., 2024; Dziri et al., 2024; Shao et al., 2022), but we focus on meaning construction from constituent representations.
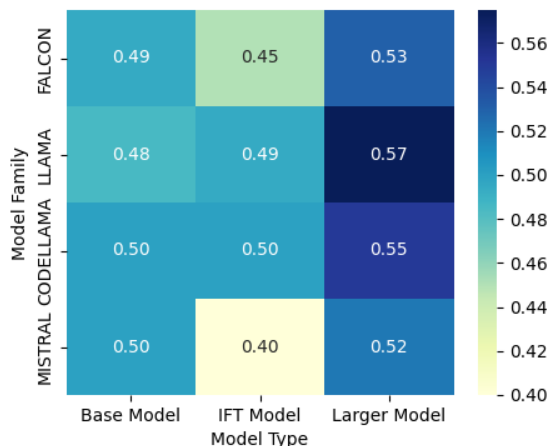

Figure 2. Heatmap for three model types and four model families on the ANTAILS dataset.

We take inspiration from Hupkes et al. (2020)'s tripartite distinction between aspects of compositionality, and introduce a test for each such aspect:

Substitutivity: This involves the ability to understand the relatedness of words such that substituting a synonym in a complex expression should not be taken to alter the meaning of the complex expression. We test this with the **ANTAILS Dataset**, largely based on the AddOne Dataset (Pavlick & Callison-Burch, 2016). For a given sentence with a noun (N) like *The runner set a record*, we substitute N with an adjective – noun combination like *The runner set a new record* and test the model to see whether it can understand the entailment pattern. The model here has to maintain it's understanding of entailment patterns with adjective substitution.

Systematicity & Globalism: This involves the ability to recombine known parts and rules and being able to productively use the parts in new contexts where constituents can have different behaviours (Carnap, 1988). We test this with the **PLANE Dataset** proposed by Bertolini et al. (2022) that tests adjective – noun entailment in a situation where the entailment pattern for an AN – N and AN – H (where AN is the adjective-noun combination, N is the noun and H is a hypernym of N) combination is already given and the model is tested on entailment of AN – AH combination. This requires the model to employ systematicity (since the AN entailment pattern needs to be recombined in the AN – AH statement) and also globalism (since the entailment pattern of the AN – AH combination needs to be inferred differently from the AN – N and AN – H combinations).

Over-generalization: This involves the ability to distinguish between compositional and non – compositional phenomena by measuring the distance of adjective – noun combinations vs exocentric compounds. We test this with a new task type using a handcrafted toy dataset- the **COMPCOMB Dataset** – which is a novel contribution of this work. Each data point consists of a triple – a noun, an adjective that goes with the noun, and an exocentric compound which contains the noun. For example, (coat, trenchcoat and turncoat)- when we take the word *"coat"*, we know that *"trenchcoat"* (a special type of coat) is closely related to it but the exocentric compound *"turncoat"* (a betrayer) is not since it is semantically different. This tests over – generalization since the model needs to be able to distinguish between genuine compounds and combinations by avoiding generalization on the basis of surface forms.

3. Evaluating Models

Our aim is to determine whether models' compositional abilities can indicate their trends of performance. There are two types of changes that have been shown to consistently impact the performance of models:

Scaling Parameters: Research on LLM scaling laws – Kaplan et al. (2020) and Hoffmann et al. (2022) – show that

Table 1. ANTAILS Experiment in Setup 1

MODEL FAMILY	BASE MODEL	IFT MODEL	LARGER MODEL
FALCON	0.50±0.01	0.46±0.01	0.54±0.03
LLAMA 2	0.50±0.01	0.54±0.04	0.60±0.01
CODELLAMA	0.50±0.01	0.50±0.01	0.55±0.02
MISTRAL	0.50±0.01	0.30±0.04	0.51±0.02

Table 2. ANTAILS Experiment in Setup 2

MODEL FAMILY	BASE MODEL	IFT MODEL	LARGER MODEL
FALCON	0.49±0.01	0.44±0.02	0.52±0.02
LLAMA 2	0.47±0.03	0.45±0.05	0.55±0.05
CODELLAMA	0.50±0.01	0.50±0.01	0.55±0.02
MISTRAL	0.50±0.03	0.50±0.20	0.53±0.07

Figure 3. Model accuracy trends for PLANE dataset.

model performance for large language models get better with size i.e an increase in the number of parameters.

Instruction Tuning: Several works (Wei et al., 2022; Ouyang et al., 2022; Chung et al., 2024) have shown the advantage of instruction finetuning (IFT) as a method to improve general performance of LLMs, especially for generalization to unseen tasks and alignment with human behaviour.

Can these changes in performance of LLMs be explained by their compositional behaviour? To investigate this, we conduct analysis and evaluation across 4 families of models – Falcon (Almazrouei et al., 2023), LLama (Touvron et al., 2023), Codellama (Roziere et al., 2023), and Mistral (Jiang et al., 2023).

3.1. ANTAILS Dataset

For this dataset, we test 3 models for each model family – the base model of 7B (Base), an instruction tuned version of the same (IFT), and a larger model – with two different kinds of setups: one involving a two – choice question scenario where we determine accuracy by fixed rank precision (P@k) to evaluate the model output (Setup 1) and another in which we use the log probabilities of the model for two completions (entails vs does not entail) as an indication of the model's judgement (Setup 2). Furthermore, for both setups we include two prompt variations for the evaluation and the result table shows the average accuracy for each model across the prompt variations.

Results: We observe that for all families of models, the Larger Model always performs better than the Base Model (Figure 1& 2). However, the impact of instruction tuning is inconsistent with performance decreasing for the two models, remaining the same for Codellama, and increasing for Llama.
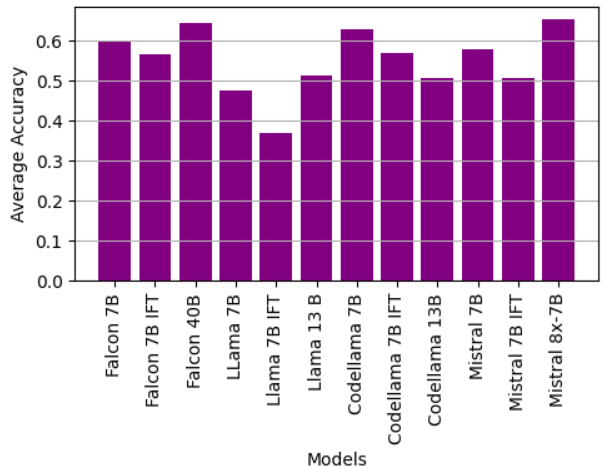
3.2. PLANE Dataset

For this dataset, we also have a setup that is exactly similar to the previous one. Since the dataset is divided by types of adjectives, we also present the results classified by the different adjectival categories.

Results: Similar to the ANTAILS dataset, we observe in Figure 3 that the overall model performance, within a model family, across two setups improves with size and worsens with instruction tuning (Tables 3 and 4). However, in the case of within – family comparison in the Codellama family of models, the larger model (13B) is worse than the base (7B) indicating that training a general LM with code and scaling it might not always have positive impacts on compositional reasoning. Similar trends were also observed by MA et al. (2024), where introduction of code at pretraining stage gives worse performance in logical reasoning tasks.

Figure 4 shows the comparative analysis of results across different adjective classes – I (Intersective), N (Subsective), and O (Intensional). Most models perform worse for subsective adjectives. Interestingly, Redolfi & Melloni (2024) notes that children also acquire subsectives the slowest during the period of language acquisition.

3.3. COMPCOMB Dataset

For each datapoint this dataset, we evaluate the accuracy of model in terms of comparative cosine distance analysis of the embeddings/hidden states of models. If the dist(N, AN) ¡ dist (N,H) we consider the model accurate since it is able to capture the semantic similarity of N – AN as compared to N –H. We do this for two types of embedding for each model – for Setup 1, we use the initial embeddings from the
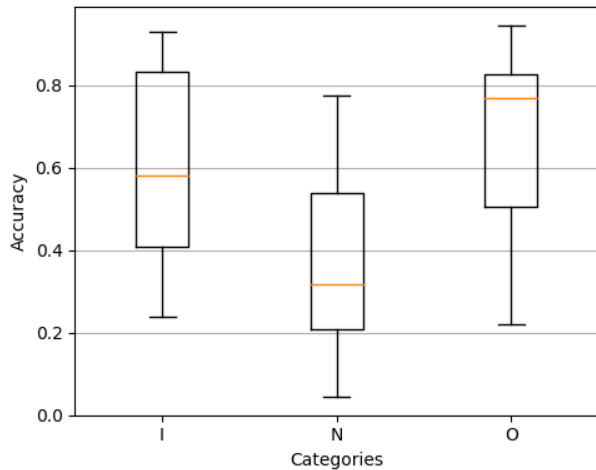
Figure 4. Average accuracies of models across 3 classes of adjectives.

Table 3. PLANE Experiments in Setup 1

MODELS	**ACCURACIES**		
	I	**N**	**O**
FALCON-7B	0.95±0.01	0.07±0.07	0.69±0.13
FALCON-7B-INSTRUCT	0.70±0.16	0.03±0.02	0.85±0.05
FALCON-40B	0.80±0.16	0.14±0.1	0.98±0.02
LLAMA-2-7B-HF	0.03±0.03	0.98±0.02	0.01±0.01
LLAMA-2-7B-CHAT-HF	0.25±0.05	0.53±0.50	0.12±0.12
LLAMA-2-13B-HF	0.04±0.03	0.95±0.05	0.14±0.07
CODELLAMA-7B-HF	0.78±0.02	0.25±0.04	0.81±0.01
CODELLAMA-7B-INSTRUCT-HF	0.55±0.08	0.35±0.05	0.60±0.15
CODELLAMA-13B-HF	0.18±0.01	0.90±0.01	0.03±0.01
MISTRAL-7B-v0.1	0.05±0.01	0.80±0.15	0.60±0.02
MISTRAL-7B-INSTRUCT-v0.1	0.64±0.01	0	0.50±0.05
MIXTRAL-8X7B-v0.1	0.98±0.03	0.04±0.02	0.98±0.02

embedding layer (EL) while in Setup 2, we access the last hidden state of the model (LHS).

Results: The accuracy of models across all families increases with size (Table 5). In Figure 5, we notice that while the embedding layer still shows over – generalization for larger models, the last hidden state representation has much better performance. For instruction tuned models, the performance of the embedding layer varies.

Refer to appendices C, D & E for additional details on models, datasets and task setups.

4. Conclusion

Cognitive architectures should arguably be performant and exhibit compositionality, and the induction of compositional strategies should be explanatory of their performance. LLMs, as candidate cognitive architectures, are clearly per-

Table 4. PLANE Experiments in Setup 2

| **MODELS**	**ACCURACIES**		
	I	**N**	**O**
FALCON-7B	0.91±0.05	0.07±0.03	0.90±0.09
FALCON-7B-INSTRUCT	0.97±0.02	0.06±0.01	0.80±0.12
FALCON-40B	0.87±0.06	0.17±0.10	0.91±0.05
LLAMA-2-7B-HF	0.45±0.06	0.51±0.14	0.88±0.01
LLAMA-2-7B-CHAT-HF	0.55±0.07	0.45±0.11	0.32±0.02
LLAMA-2-13B-HF	0.89±0.01	0.15±0.02	0.91±0.05
CODELLAMA-7B-HF	0.88±0.02	0.20±0.01	0.86±0.11
CODELLAMA-7B-INSTRUCT-HF	0.65±0.07	0.39±0.06	0.89±0.10
CODELLAMA-13B-HF	0.47±0.08	0.65±0.05	0.82±0.05
MISTRAL-7B-v0.1	0.77±0.21	0.27±0.21	0.98±0.02
MISTRAL-7B-INSTRUCT-v0.1	0.48±0.02	0.50±0.09	0.92±0.02
MIXTRAL-8X7B-v0.1	0.58±0.10	0.49±0.15	0.85±0.15
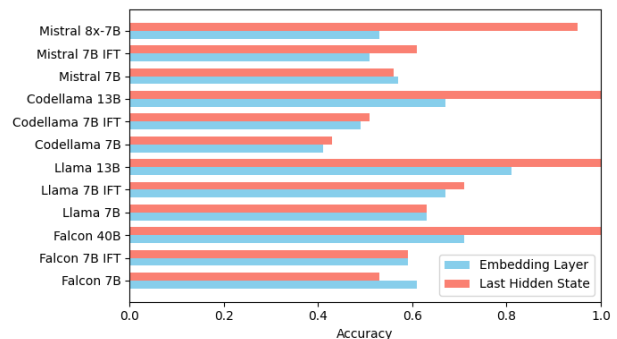
Figure 5. Model Accuracy trends for two setups with the COMP-COMB Dataset

formant, behave compositionally (as seen by their performance on the ANTAILS and PLANE datasets), and their representations appear compositional (as seen by their performance on the COMPCOMB dataset). However, when it comes to how explanatory the induction of compositional strategies are of performance improvements, we observe different patterns for different LLMs:

1) Scaling models improves their generalization capabilities (Hendrycks et al., 2020; Desai & Durrett, 2020) and overall performance (Kaplan et al., 2020; Hoffmann et al., 2022). Compositional behaviour also improves with scaling across model families. This could indicate that the induction of compositional strategies is explanatory of improvements with scaling.

2) Instructing finetuning has been shown to improve alignment and result in performance gains across several task types (Wei et al., 2022; Ouyang et al., 2022). However, we see that compositional performance does not always improve with instruction tuning. Performance gains from instruction tuning do not correlate with improved compositional behaviour.

Table 5. COMPCOMB Experiments for two setups

MODELS	ACCURACY (EL)	ACCURACY (LHS)
FALCON-7B	0.61	0.53
FALCON-7B-INSTRUCT	0.59	0.59
FALCON-40B	0.71	1
LLAMA-2-7B-HF	0.63	0.63
LLAMA-2-7B-CHAT-HF	0.67	0.71
LLAMA-2-13B-HF	0.81	1
CODELLAMA-7B-HF	0.41	0.43
CODELLAMA-7B-INSTRUCT-HF	0.49	0.51
CODELLAMA-13B-HF	0.67	1
MISTRAL-7B-V0.1	0.57	0.56
MISTRAL-7B-INSTRUCT-V0.1	0.51	0.61
MIXTRAL-8X7B-V0.1	0.53	0.95

In sum, while scaling often leads to more compositional models, instruction tuning does not show similar trends. Recent work (Ghosh et al., 2024) has shown that instruction tuning sometimes degrades performance. Our results indicate that one source of error may be reduced compositionality. Performance is multi – faceted, and compositionality may be explanatory of some performance gains, not others. If we think cognitive architectures should learn compositional strategies (Fodor & Pylyshyn, 1988; Symons & Calvo, 2014), *and* that LLMs could potentially be cognitive architectures (Lamprinidis, 2023; Sumers et al., 2023; Zhao et al., 2023), we must evaluate if the compositionality of LLMs is explanatory of their performance and be precise about what (relevant) performance is at play. This work is, to the best of our knowledge, the first step in that direction.

Limitations

The focus on adjective-noun combinations in tasks might provide a limited view of the models' overall compositional abilities. Broader investigation across various domains is necessary to understand models' capabilities, limitations, and behavior trends in scaled versus instruction – tuned models. Additionally, incorporating error analysis and interpretability techniques will uncover underlying mechanisms and biases in model outputs, guiding improvements and ensuring more transparent interpretations and application of results. We plan on incorporating such changes in future iterations of this work.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Alabdulakreem, A., Arnold, C. M., Lee, Y., Feenstra, P. M., Katz, B., and Barbu, A. Securellm: Using compositionality to build provably secure language models for private, sensitive, and secret data. *arXiv preprint arXiv:2405.09805*, 2024.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Aw, K. L., Montariol, S., AlKhamissi, B., Schrimpf, M., and Bosselut, A. Instruction-tuning aligns llms to the human brain. *arXiv preprint arXiv:2312.00575*, 2023.

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

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

Chalmers, D. J. Connectionism and compositionality: Why fodor and pylyshyn were wrong. 1993.

Chen, J., Pan, X., Yu, D., Song, K., Wang, X., Yu, D., and Chen, J. Skills-in-context prompting: Unlocking compositionality in large language models, 2024. URL https://openreview.net/forum?id=s1ByDEbpI8.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Cummins, R. *Representations, targets, and attitudes*. MIT press, 1996.

Desai, S. and Durrett, G. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*, 2020.

Drozdov, A., Schärli, N., Akyürek, E., Scales, N., Song, X., Chen, X., Bousquet, O., and Zhou, D. Compositional semantic parsing with large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=gJW8hSGBys8.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Le Bras, R., et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36, 2024.

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

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

Ghosh, S., Evuru, C. K. R., Kumar, S., Aneja, D., Jin, Z., Duraiswami, R., Manocha, D., et al. A closer look at the limitations of instruction tuning. *arXiv preprint arXiv:2402.05119*, 2024.

Hardy, M., Sucholutsky, I., Thompson, B., and Griffiths, T. Large language models meet cognitive science: Llms as tools, models, and participants. In *Proceedings of the annual meeting of the cognitive science society*, volume 45, 2023.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., and Levy, R. Language models align with human judgments on key grammatical constructions. *arXiv preprint arXiv:2402.01676*, 2024.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67: 757–795, 2020.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., Fedorenko, E., and Lenci, A. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47 (11):e13386, 2023.

Kim, N. and Linzen, T. COGS: A compositional generalization challenge based on semantic interpretation. In

Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main. 731. URL https://aclanthology.org/2020. emnlp-main.731.

Lake, B. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/ v80/lake18a.html.

Lake, B. M. and Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

Lamprinidis, S. Llm cognitive judgements differ from human. In *International Conference on Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications*, pp. 17–23. Springer, 2023.

Li, Z., Jiang, G., Xie, H., Song, L., Lian, D., and Wei, Y. Understanding and patching compositional reasoning in llms. *arXiv preprint arXiv:2402.14328*, 2024.

Liska, A., Kruszewski, G., and Baroni, M. Memorize or generalize? searching for a compositional RNN in a haystack. *CoRR*, abs/1802.06467, 2018. URL http: //arxiv.org/abs/1802.06467.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pp. 22631–22648. PMLR, 2023.

Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.-W., Wu, Y. N., Zhu, S.-C., and Gao, J. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

MA, Y., Liu, Y., Yu, Y., Zhang, Y., Jiang, Y., Wang, C., and Li, S. At which training stage does code data help LLMs reasoning? In *The Twelfth International Conference on Learning Representations*, 2024. URL https:// openreview.net/forum?id=KIPJKST4gw.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., and Griffiths, T. What language reveals about perception: Distilling psychophysical knowledge from large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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

Pavlick, E. and Callison-Burch, C. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In Erk, K. and Smith, N. A. (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2164–2173, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1204. URL https://aclanthology.org/P16-1204.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Pinker, S. and Prince, A. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193, 1988.

Plunkett, K. and Juola, P. A connectionist model of english past tense and plural morphology. *Cognitive Science*, 23 (4):463–490, 1999.

Redolfi, M. and Melloni, C. Processing adjectives in development: Evidence from eye-tracking. *Journal of Child Language*, pp. 1–24, 2024.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Remez, T., Rapin, J., et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

Shao, N., Cai, Z., Liao, C., Zheng, Y., Yang, Z., et al. Compositional task representations for large language models. In *The Eleventh International Conference on Learning Representations*, 2022.

SHAO, N., Cai, Z., xu, H., Liao, C., Zheng, Y., and Yang, Z. Compositional task representations for large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6axIMJA7ME3.

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

Socher, R., Manning, C. D., and Ng, A. Y. Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-2010 deep learning and unsupervised feature learning workshop*, volume 2010, pp. 1–9. Vancouver, 2010.

Sumers, T. R., Yao, S., Narasimhan, K., and Griffiths, T. L. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

Sun, J., Shaib, C., and Wallace, B. C. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270*, 2023a.

Sun, K., Williams, A., and Hupkes, D. The validity of evaluation results: Assessing concurrence across compositionality benchmarks. In Jiang, J., Reitter, D., and Deng, S. (eds.), *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 274–293, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.19. URL https://aclanthology.org/2023.conll-1.19.

Symons, J. and Calvo, P. Systematicity: an overview. 2014.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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

Wang, S., Wei, Z., Choi, Y., and Ren, X. Can llms reason with rules? logic scaffolding for stress-testing and improving llms. *arXiv preprint arXiv:2402.11442*, 2024.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Xu, Z., Shi, Z., and Liang, Y. 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*, 2024.

Zhang, M., He, J., Lei, S., Yue, M., Wang, L., and Lu, C.-T. 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)*, pp. 11996–12000. IEEE, 2024.

Zhao, Z., Song, S., Duah, B., Macbeth, J., Carter, S., Van, M. P., Bravo, N. S., Klenk, M., Sick, K., and Filipowicz, A. L. More human than human: Llm-generated narratives outperform human-llm interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, pp. 368–370, 2023.

A. Related Work

Most of the earlier work on testing compositionality in connectionist systems was centered around two main types:

1. Testing compositional abilities: Most works (Lake & Baroni, 2018; Liska et al., 2018; Kim & Linzen, 2020; Hupkes et al., 2020) have a training and testing paradigm where models were considered to be performing compositional generalization if they were able to successfully handle unseen test sequences.

2. Enhancing compositional abilities: This area of research was focused on what enhancements to connectionist models – architectures, training methods, or data – could provide improved compositionality. Some like Socher et al. (2010) involved combining syntactic parse trees with connectionist architectures to learn compositional functions, allowing models to be 'compositional by design' while other work like Lake & Baroni (2023) proposed a novel method for training neural networks via a series of compositional tasks that endows them with systematic generalization capabilities.

Recent work has shifted the focus to testing compositional generalization in pretrained models via tasks that require no further training. There is some research that focuses on prompting to enable better results in compositional tasks (Drozdov et al., 2023; Chen et al., 2024).

However, much of recent work (Li et al., 2024; Alabdulakreem et al., 2024; SHAO et al., 2023; Zhang et al., 2024) interprets compositionality to be multi – hop reasoning which is not "true" compositionality which was originally discussed as a feature of human language and cognition (Frege, 1892; Fodor & Pylyshyn, 1988).

B. The Compositionality Debate- Symbolism vs Connectionism

The concept of compositionality has a long history in linguistic and cognitive science – it was perhaps first discussed in detail by Frege (1892) in the context of how natural language expressions were assigned meanings. Partee et al. (1995) formulated the so-called *principle of compositionality*:

> The meaning of a complex expression is determined by the meanings of its constituent parts and the rules used to combine them.

Compositionality has long been considered a cornerstone of human cognitive capabilities and was notably discussed in Fodor & Pylyshyn (1988) as the reason for the systematicity of human thought – how the ability to think a thought is linked to the ability to also have related thoughts. Non –

symbolic connectionist representations were, in the view of Fodor and Pylyshyn, inadequate and unviable. Instead, they claimed that: (i) only classical or symbolic representations can give rise to compositional and, in turn, systematic behaviour; and (ii) neural networks do not have classical representations and thus they cannot exhibit compositional understanding or behaviour.

A central tenet of Fodor & Pylyshyn (1988) was also not how connectionist systems could not behave compositionally but why they could not serve as viable cognitive architectures. The then highly debated topic of acquisition of English past tense is discussed to point out that even though Plunkett & Juola (1999) finds a way to show that connectionist systems can simulate this pattern, (Pinker & Prince, 1988) is correct in asserting that it does not in any way *explain* the actual cognitive process. Symons & Calvo (2014) further elucidate the point claiming that even if connectionist systems can stumble across an implementation of compositionality, this does not *explain* systematicity in their behaviour nor does it render them suitable cognitive architectures.

Smolensky (1987) and Chalmers (1993) were instrumental in challenging the prevailing skepticism towards connectionist networks by asserting that these networks have the capacity to embody classical representations through their intricate connection weights and activation patterns, thereby exhibiting compositional behavior. This assertion stems directly from the inherent expressivity of connectionist models, which allow them to capture complex relationships and hierarchies within data.

Van Gelder (1990) suggested that neural networks displayed *functional compositionality* instead of the traditional 'concatenative' compositionality discussed by Fodor. According to Van Gelder, neural networks demonstrate a form of compositionality where the functions computed by individual neurons or layers combine in a compositional manner to produce complex behaviors, without necessarily relying on explicit concatenation of discrete symbols.

Cummins (1996) argued against the dichotomy between classical and non-classical representations, contending that the distinction fails to hold ground since classical representations themselves can exhibit non – compositional characteristics. This perspective underscores the complexity and fluidity inherent in the nature of representations, suggesting that compositionality is not necessarily tied to a specific type of representation but rather emerges from the interactions and transformations within a system.

C. Models

The models used here are all based on the transformer architecture but are decoder – only models. For each model family, we use 3 variants:
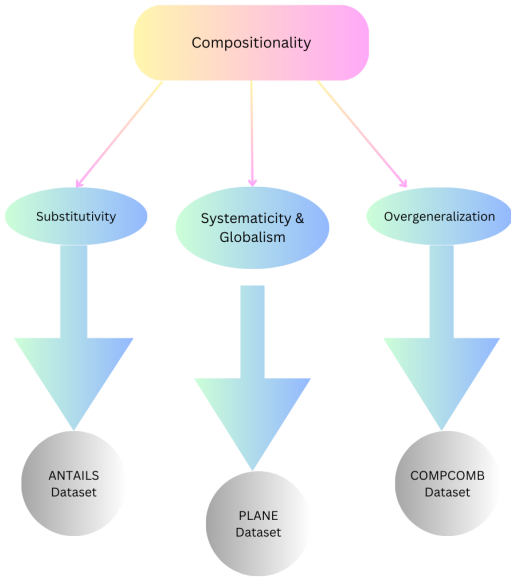
Figure 6. Testing Compositionality: Experimental Setup for LLMs

Table 6. Models used and corresponding Huggingface Hub Links

MODEL NAME	MODEL LINK
FALCON-7B	HTTPS://HUGGINGFACE.CO/TIIUAE/FALCON-7B
FALCON-7B-INSTRUCT	HTTPS://HUGGINGFACE.CO/TIIUAE/FALCON-7B-INSTRUCT
FALCON-40B	HTTPS://HUGGINGFACE.CO/TIIUAE/FALCON-40B
LLAMA-2-7B-HF	HTTPS://HUGGINGFACE.CO/META-LLAMA/LLAMA-2-7B-HF
LLAMA-2-7B-CHAT-HF	HTTPS://HUGGINGFACE.CO/META-LLAMA/LLAMA-2-7B-CHAT-HF
LLAMA-2-13B-HF	HTTPS://HUGGINGFACE.CO/META-LLAMA/LLAMA-2-13B-HF
CODELLAMA-7B-HF	HTTPS://HUGGINGFACE.CO/META-LLAMA/CODELLAMA-7B-HF
CODELLAMA-7B-INSTRUCT-HF	HTTPS://HUGGINGFACE.CO/META-LLAMA/CODELLAMA-7B-INSTRUCT-HF
CODELLAMA-13B-HF	HTTPS://HUGGINGFACE.CO/META-LLAMA/CODELLAMA-13B-HF
MISTRAL-7B-V0.1	HTTPS://HUGGINGFACE.CO/MISTRALAI/MISTRAL-7B-V0.1
MISTRAL-7B-INSTRUCT-V0.1	HTTPS://HUGGINGFACE.CO/MISTRALAI/MISTRAL-7B-INSTRUCT-V0.1
MIXTRAL-8X7B-V0.1	HTTPS://HUGGINGFACE.CO/MISTRALAI/MIXTRAL-8X7B-V0.1

1. Falcon Family: falcon-7B, falcon-7B-instruct, falcon-40B

2. Llama 2 Family: Llama-2-7B-hf, Llama-2-7B-chat-hf, Llama-2-13B-hf

3. Codellama Family : Codellama-7B-hf, Codellama-7B-Instruct-hf, Codellama-13B-hf

4. Mistral Family: Mistral-7B-v0.1, Mistral-7B-Instruct-v0.1, Mixtral-8x7B-v0.1

We provide a summary of models used and their Hugging-face Hub links in Table 6 to enable easy reproduction and use.

D. Datasets

We measure different aspects of compositionality with 3 task types/ datasets:

1. ANTAILS: It is the adjective noun entailment dataset. The dataset is influenced by Pavlick & Callison-Burch (2016) but we found certain discrepancies in the dataset due to which

Figure 7. ANTAILS Dataset

Figure 8. PLANE Dataset

we slightly modify and build our own dataset as shown in Figure 7.

2. PLANE: It adjective – noun hypernym inference pattern testing introduced in Bertolini et al. (2022) and we use the same dataset as shown in Figure 8.

3. COMPCOMB: This is a task (as shown in Figure 9) that centers around measuring the distance of compounds vs adjective – noun combinations in the embedding space of models. It is a novel contribution of this work.

E. Task Setup

To investigate the trends of learning compositional strategies, we investigate two types of models – base models and instruction models – and use similar prompts for all models. Some motivations for our prompting choice setup are the following:

1. For all models, we do a zero – shot prompt setting to attempt an unbiased comparison of general vs instruction tuned models. Works on instruction tuning (Wei et al., 2022;

Figure 9. COMPCOMB Dataset

Sanh et al., 2021) indicate that such models have good zero – shot task performance and thus we chose this prompting mode for all models to try and avoid undue bias.

2. We wanted to use a similar prompt structure across models in our work to maintain uniformity of evaluation. Since using instruction format prompts would disadvantage a non instruction tuned model and research indicates instruction tuning improves general reasoning and performance, we chose to avoid specific prompting methods involving advanced instructions. Non – instruction prompts can effectively serve as robust evaluation tools, helping to assess the model's true understanding and generalization ability beyond the training data (Peng et al., 2023; Sun et al., 2023a).

For the ANTAILS and PLANE datasets, we use two task setups:

1. **Two – Choice QA**: The first setup gives models statements indicating entailment and non entailment as two options and the model choice of option is considered. We avoid using the yes-no setup to prevent possible yes – bias outputs.

2. **Logprob Calculation**: The second setup involves passing in the prompt with dataset samples and calculating the log probabilities of the model for a statement indicating entailment and one indicating non – entailment. The statement assigned higher completion log probability is considered to be the model output.

For both setups, we use two prompts and average the outputs to calculate our results. We observe similar trends across different prompt choices.

For the COMPCOMB dataset, the above task settings do not apply since we directly compare representations of the model from the embedding layer and the last hidden layer.