From Fine-Tuning-Based Metrics to Quantifying the Utilization of Linguistic Compositionality: Issues on the Proxy Chain

Anonymous ACL submission

Abstract

Compositional generalization tests, widely used in research on compositionality, essentially use the performance of fine-tuned models under prescribed compositional structures as a proxy for the utilization of linguistic compositionality. However, this proxy raises two unresolved issues: (1) how prescribed compositional structures affect quantification, and (2) whether the model compositionality as defined by fine-tuning is correlated with the utilization of linguistic compositionality. In this work, we experiment with Topographic Similarity, a quantitative method without prescribed compositional structures, and analyze the development of model compositionality during fine-tuning and its relationship to linguistic compositionality based on two key concepts of linguistic compositionality: structure and component. Our key findings are: (1) Whether or not compositional structures are prescribed significantly affects the quantification results, indicating that the spontaneous development of model compositionality during fine-tuning does not necessarily follow the expected direction; (2) the development of model compositionality during fine-tuning is not entirely symbolic and is influenced by linguistic factors. The analysis of these influences provides an explanation, to some extent, for how fine-tuning-based model compositionality can serve as a proxy for the utilization of linguistic compositionality.

1 Introduction

Compositionality is a property that languages possess to some degree, expressed as "the meaning of a complex expression is determined by its structure and the meanings of its components" (Szabó, 2004; Pagin and Westerståhl, 2010). Humans possess the ability to understand and generate linguistic expressions based on language compositionality. Through independent perception of linguistic components, we achieve comprehension and generation of arbitrary new combinations formed from known components. This powerful out-of-distribution generalization capability is commonly referred to as compositional generalization (Hupkes et al., 2022).

In the field of natural language processing, whether language models understand and generate based on linguistic compositionality has long been a contentious issue. For language models, directly quantifying their degree of leveraging linguistic compositionality is quite challenging. Researchers typically extend the mapping from language to meaning in linguistic compositionality to the input-to-output mapping in specific language tasks. They then characterize a model's compositionality through its ability of compositional generalization, known as the compositional generalization test (Lake and Baroni, 2018; Kim and Linzen, 2020). Specifically, this test constructs a special training-test split for specific downstream tasks, ensuring that the components have never appeared in the training set. The model's performance on the test set is then evaluated after fine-tuning on the training set (Keysers et al., 2020). Essentially, compositional generalization tests utilize the results of fine-tuning-based generalization tests under prescribed compositional structures as a proxy for model compositionality on specific tasks, and then treat model compositionality as a proxy for the model's utilization of linguistic compositionality; this chain of proxies can be represented as: fine-tuning-based metrics → model compositionality → utilization of linguistic compositionality. Although compositional generalization tests are widely used in the field of compositionality research, there is a lack of study on the validity of this proxy chain, and two issues require further explanation:

(I1) There can be a discrepancy between fine-tuning based metrics under prescribed compositional structures and model compositionality. Compositional generalization tests essentially require language models to learn the compositional structures specified by the test designer during fine-

tuning and to generalize accordingly (i.e., there are fixed, predefined correct answers). Although these prescribed compositional structures are typically intuitive, they are not supported by any fine-tuning data. Under these circumstances, the model developing compositional structures that do not align with human intuition should not be considered an error, yet it is treated as such in compositional generalization tests; merely observing whether the model generalizes according to the prescribed compositional structure may fail to accurately characterize the model's compositionality. We need to determine whether this potential discrepancy exists in practical scenarios.

(I2) The relationship between model compositionality as defined by fine-tuning and a model's utilization of linguistic compositionality remains unclear. Although there has been extensive research on the evaluation of the compositionality of language models in specific downstream tasks requiring fine-tuning (Hu et al., 2023; Xu and Wang, 2024), it is still unclear whether model compositionality as defined by fine-tuning can appropriately serve as a proxy for a model's utilization of linguistic compositionality. Similar to debates surrounding language learning capabilities (Kallini et al., 2024), a contentious issue is whether the compositionality development of pre-trained language models during fine-tuning is entirely symbolic (i.e., whether fine-tuning the same model with training data that provides linguistically distinct mappings of identical symbolic essence would lead the model to develop the same compositionality). If this development is entirely symbolic, such model compositionality is independent of the model's utilization of linguistic compositionality and cannot serve as a proxy for that utilization; if this development is not entirely symbolic, we need to investigate the relationship between such model compositionality and the utilization of linguistic compositionality to determine whether it can appropriately serve as a proxy.

In this work, we introduce the concept of Topographic Similarity (Brighton and Kirby, 2006) to quantify the development of model compositionality during fine-tuning; unlike compositional generalization tests, this concept does not impose any requirements regarding the correctness of the output. In a specific conversion task, we analyze the development of model compositionality during fine-tuning under two quantification methods. To explain I1, we compare the quantitative results provided by the Topographic Similarity and the compositional generalization test; To explain I2, we explore the two key concepts of linguistic compositionality—structure and component. By constructing mappings that share the same symbolic essence but differ linguistically as fine-tuning data, we seek to identify linguistic factors influencing the development of model compositionality during fine-tuning, thereby uncovering the relationship between model compositionality as defined by fine-tuning and the utilization of linguistic compositionality. Our research indicates that, for I1, the quantification methods with and without prescribed compositional structures yield significant differences in the quantification of model compositionality, indicating that the spontaneous development of compositionality during model fine-tuning may not necessarily align with the prescribed direction; for I2, the development of model compositionality during fine-tuning is not entirely symbolic and is influenced by linguistic factors. To some extent, this influence helps explain why fine-tuning-based model compositionality can be used as a proxy for the utilization of linguistic compositionality.

2 Compositionality Quantification

2.1 Task Definition

We construct a conversion task where the input is natural language text generated based on a template (Lee et al., 2025) with 12 components:

`In the [place.ADJ] [place.N], I [action.VI] beside the [animal.N] and the [plant.N], while the [person.ADJ] [job.N] [action.ADV] [action.VT] the [object.ADJ] [texture.ADJ] [object.N].`

Each component c_k enclosed by [] has 50 distinct candidate words $\{w_{k,1}, w_{k,2}, ..., w_{k,50}\}$, with each candidate word corresponding to a unique 5-digit random number code $\{e_{k,1}, e_{k,2}, ..., e_{k,50}\}$. For each component, we provide DeepSeek-V3.2 with a description of the component's function and request it to generate 50 commonly used candidate words. When generated candidate words have already served as candidates for other components, we require the model to regenerate new candidate words, thereby ensuring no overlap among candidate words for different components (see appendix A for the detailed vocabulary). The task output is the sequence formed by concatenating the corresponding number codes for each component in order (separated by spaces).

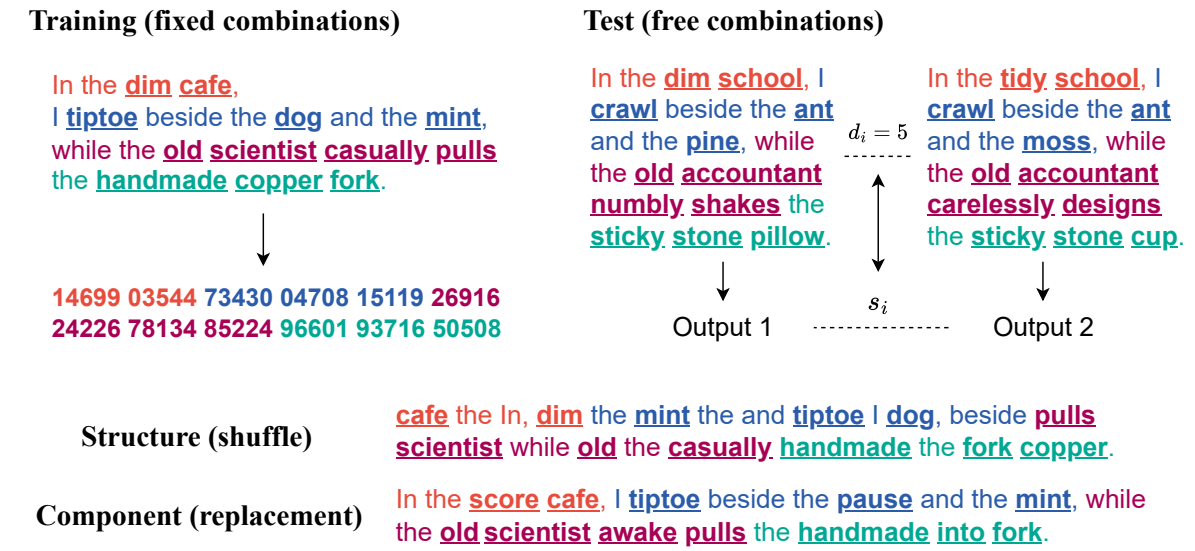

Figure 1: Illustration of compositionality quantification (Topographic Similarity) and how inputs change under two types of settings. Each **component** has 50 candidate words. In the training data, the combination of components within a semantic group (indicated by color) is fixed. In the test data, components can be freely combined.

By selecting one candidate word for each component, we can obtain 50^{12} distinct instances. However, during training set construction, for each semantic group represented by a color in the template, we restrict the selection of components to 50 fixed combinations, and any candidate word appears in exactly one combination. For example, for components c_1, c_2 belonging to the same semantic group, the tuple (c_1, c_2) has only 50 possible values in the training set: $\{(w_{1,1}, w_{2,1}), (w_{1,2}, w_{2,2}), ..., (w_{1,50}, w_{2,50})\}$.

In such a training set, any word always appears alongside fixed other components within the semantic group. Consequently, there is no evidence that a word independently determines an output component (i.e., the training set provides no evidence for compositionality). When fine-tuning a language model using this training set, the extent to which the model perceives a word as independently determining an output component indicates the degree of its spontaneous compositionality development. We randomly construct $N = 20,000$ distinct samples in this manner to form the training set for fine-tuning the pre-trained model.

2.2 Topographic Similarity

Following Sevestre and Dupoux (2025), we employ the concept of Topographic Similarity to quantify model compositionality during fine-tuning. We first generate $M = 2,000$ text inputs based on the template, where candidate word selection no longer imposes combination constraints as in training set construction (i.e., the values of each component c_k are sampled completely independently and are unrelated to the other components). For the i-th input x_i, we randomly select an input distance $d_i \in [1, 12]$, forming the input x_i^* by replacing d_i components in x_i with other candidate words. We then measure the distance between the model's outputs for inputs x_i and x_i^* (i.e., the number of differing number codes) as s_i. Finally, we compute the Pearson correlation coefficient between D and S as a quantitative metric for compositionality:

$$r(D, S) = \frac{\sum_{i=1}^{M}(d_i - \bar{D})(s_i - \bar{S})}{\sqrt{\sum_{i=1}^{M}(d_i - \bar{D})^2}\sqrt{\sum_{i=1}^{M}(s_i - \bar{S})^2}}$$

where $\bar{D}$ and $\bar{S}$ denote the mean values of $\{d_i\}$ and $\{s_i\}$, respectively. When the model exhibits maximum compositionality, each component in the input independently determines a corresponding number code in the output, resulting in a correlation coefficient of 1. A lower correlation coefficient indicates lower model compositionality. When using Topographic Similarity to quantify compositionality, there is no concept of output correctness.

2.3 Compositional Generalization Tests

We also use compositional generalization tests to quantify model compositionality during the fine-

tuning. In compositional generalization tests, the correctness of the output must be defined. We use an intuitive compositional structure to define correctness: when the k-th component c_k of the input takes the value $w_{k,t}$, the correct answer for the k-th number code of the output is $e_{k,t}$. Under this definition, we use the model's accuracy across M samples as a quantitative measure of compositionality. We observe that models are never able to achieve a completely correct output on a single sample; therefore, we use the number code rather than the sample as the unit of accuracy (i.e., accuracy is defined as the number of correct number codes divided by the total number of number codes, rather than the number of completely correct samples divided by the total number of samples).

In Section 2.1, we note that the development of model compositionality in this task is spontaneous: any given $w_{k,t}$ always appears alongside fixed components within the same semantic group; thus, there is no evidence that the input $w_{k,t}$ independently determines any specific component of the output. In compositional generalization tests, the required compositional structure is that the input $w_{k,t}$ independently determines that the k-th number code of the output is $e_{k,t}$. This compositional structure is one possible form of compositionality development and is not supported by the training data; when we use compositional generalization tests, we are essentially stipulating that the model's compositionality development must conform to the prescribed compositional structure.

3 Main Experiments

In this section, we present the main experimental procedures and present experimental results quantifying compositionality using Topographic Similarity. The experiments are first conducted under the original setting; drawing on two key concepts of linguistic compositionality (structure and component) , we generate training data that shares the same symbolic essence as the original setting but differs linguistically, and compare it with the experiments under the original setting. Identical symbolic essence refers to the following: for datasets A and B, if there exists a one-to-one mapping from words to symbols and a transformation rule that does not involve specific symbols, such that the two datasets are identical, then A and B are considered to have the same symbolic essence.

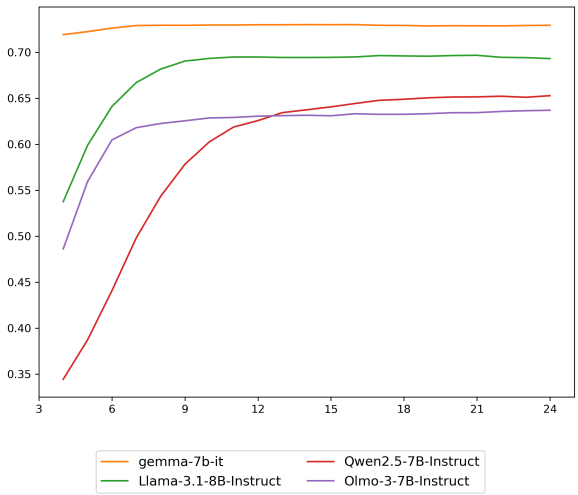

Figure 2: Curves showing how the model compositionality quantified by Topographic Similarity varies over epochs under the original setting.

3.1 Original Setting

We conduct experiments on the following pretrained language models: gemma-7b-it (Gemma-Team, 2024), Llama-3.1-8B-Instruct (Llama-Team, 2024), Qwen2.5-7B-Instruct (Qwen-Team, 2024), and Olmo-3-7B-Instruct (Olmo-Team, 2025). All results are averaged across three independent runs with different random seeds. For training, we use a fixed learning rate of 1e-6 and a batch size of 1024. We train for 30 epochs, evaluating at the end of each epoch. Each point on the curve represents the average result over the most recent 5 epochs. Three runs of experiments take approximately 7 to 9 hours when using 8 A800 GPUs. Figure 2 shows the topographic similarity of different models under the original setting.

3.2 Structure

Regarding structure, we seek to investigate whether the compositionality development of language models during fine-tuning changes when the symbolic essence of training data remains identical but the linguistic structures differ.

Following Kallini et al. (2024), we employ LocalShuffle to construct instances that lose natural language structure. Under the shuffle setting, for each original instance, we shuffle all words within each semantic group, causing the input to lose its typical natural language structure. The shuffling pattern remains consistent across different instances within the same semantic group, and the order of the output number codes is adjusted to match the input component sequence after shuf-

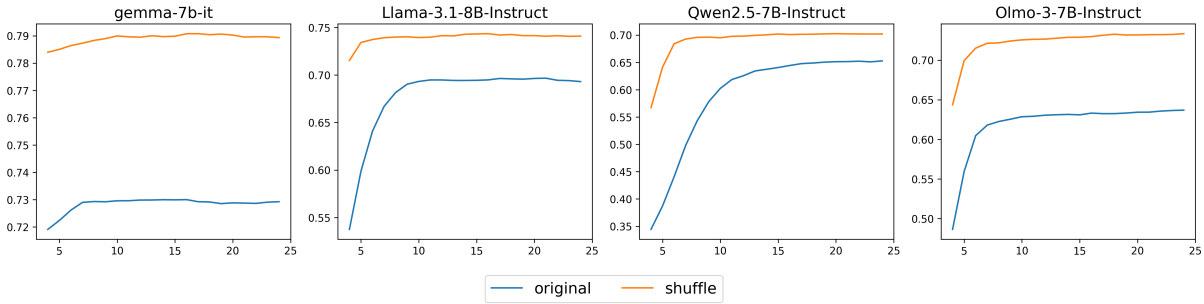

Figure 3: Curves showing how the model compositionality quantified by Topographic Similarity varies over epochs under the original and shuffle settings.

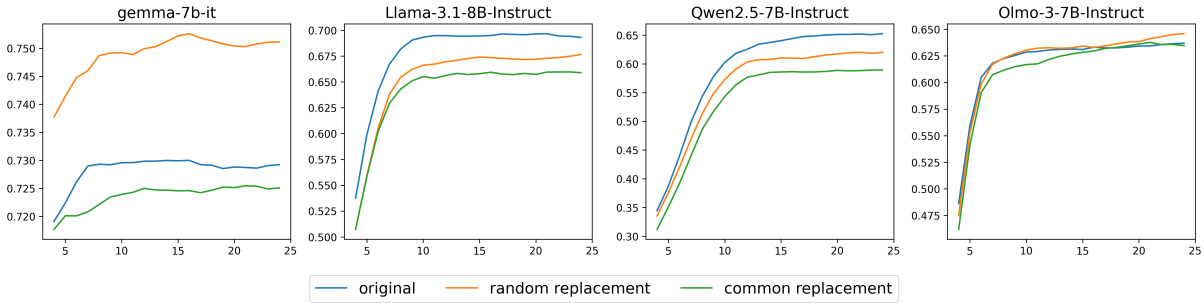

Figure 4: Curves showing how the model compositionality quantified by Topographic Similarity varies over epochs under the original and replacement settings.

fling. This ensures that the symbolic essence of the training set remains identical between the original and shuffled settings. We observe whether there are differences in the compositionality development of models during fine-tuning under the original and shuffled settings.

Figure 3 presents the experimental results. We find that all pre-trained models exhibit higher compositionality under the shuffle setting compared to the original setting. This indicates that even when the training set provides mappings of identical symbolic essence, structural alterations still influence compositionality development during fine-tuning. When the structure deviates from the natural language structure common in pre-training, the model develops higher compositionality, exhibiting a greater degree of component independence.

3.3 Component

Regarding component, we seek to investigate whether the compositionality development of language models during fine-tuning changes when the symbolic essence of training data remains identical but the linguistic components differ.

We first employ a random replacement setting to construct instances with uncommon component combinations. Under the random replacement setting, for each original instance, we select one word component in each semantic group and replace it with a word from a shared vocabulary of the four evaluated models (containing 13,740 words). The shared vocabulary is derived from the intersection of the vocabularies from the four evaluated pre-training models. We identify words within this vocabulary intersection that can function as independent new tokens following a blank space. These words are then used to construct a vocabulary for the random replacement setting, excluding any component candidate words used to build the original instances. Different instances share identical component selections within the same semantic group and employ the same replacement word for the same original word. This ensures that the symbolic essence of the training set remains identical between the original and random replacement settings. We observe whether there are differences in the compositionality development of models during fine-tuning under the two settings.

Figure 4 presents the experimental results (blue and orange lines). We observe that random replacement settings influence the compositionality development of models, but the direction varies across different models: compositionality increases on gemma-7b-it and Olmo-3-7B-Instruct, while com-

positionality decreases on Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct.

This result leads us to consider whether there are factors other than the commonness of component combinations that influence compositionality development in component replacement. Through observation of instances, we find that under the random replacement setting, there is a high probability of selecting uncommon words from the large vocabulary. Based on frequency and compositionality theory (Kirby, 2001; Sevestre and Dupoux, 2025), we hypothesize that the commonness of the replacement words themselves may also affect compositionality development. To verify this, we supplement our experiments with a common replacement setting. The common replacement setting follows the same construction as the random replacement setting, but the vocabulary used to select replacement words is a subset of the vocabulary in the random replacement setting, containing only 1,312 common words. To obtain the vocabulary, we first require DeepSeek-V3.2 to provide a list containing 3,000 distinct common words. We then take the intersection of this list with the vocabulary used for the random replacement setting to obtain the vocabulary for the common replacement setting.

The green line in Figure 4 shows the results of the supplementary experiment. We find that all pre-trained models exhibit higher compositionality under the random replacement setting than under the common replacement setting, indicating that low word commonality leads to high compositionality. Comparison between the original setting and the common replacement setting mitigates the influence of word commonality. The results show that all pre-trained models exhibit lower or essentially unchanged compositionality the common replacement setting. This implies that even when the training set provides mappings of identical symbolic essence, changes in component still influence compositionality development during fine-tuning. The influencing factors include both the commonness of components and the commonness of component combinations. Uncommon components lead models to develop higher compositionality during fine-tuning, while uncommon component combinations result in lower compositionality.

4 Analysis

In this section, we first apply the experimental framework described in Section 3 to the quantification of model compositionality based on compositional generalization tests, and present the experimental results. By comparing the results of the two different quantification methods, we provide an explanation for issue I1 in Section 1; through a comprehensive analysis and discussion of the experimental results, we provide an explanation for issue I2 in Section 1.

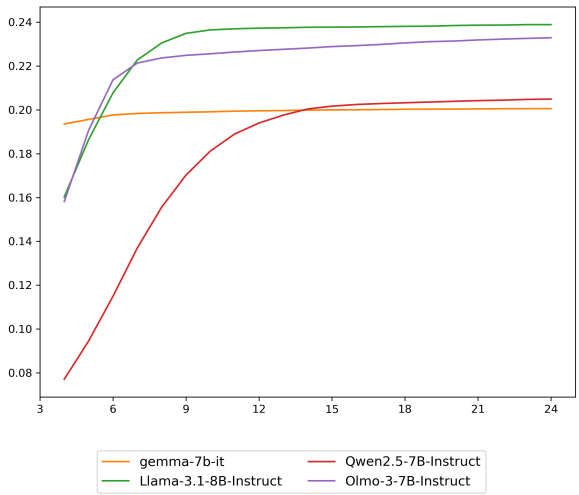

Figure 5: Curves showing how the model compositionality quantified by compositional generalization tests varies over epochs under the original setting.

4.1 Compositional Generalization Tests

Figure 5 shows the model's performance in the compositional generalization tests under the original setting. There is a significant difference in model ranking between the compositionality estimation provided by the compositional generalization tests and that provided by Topographic Similarity. As discussed in Section 2.3, the compositional generalization tests predefine a compositional structure, which means it focuses only on the compositionality development of models that aligns with the expected direction. During fine-tuning, models may exhibit spontaneous compositionality developments that do not align with this expected direction, leading to significant differences in the results between the two quantification methods.

Figures 6 and 7 illustrate the compositionality development of the models during fine-tuning quantified by the compositional generalization tests under different settings. Regarding structure, the models still exhibit higher compositionality under the shuffle setting than under the original setting, similar to the results obtained when using Topographic Similarity. Regarding components, the

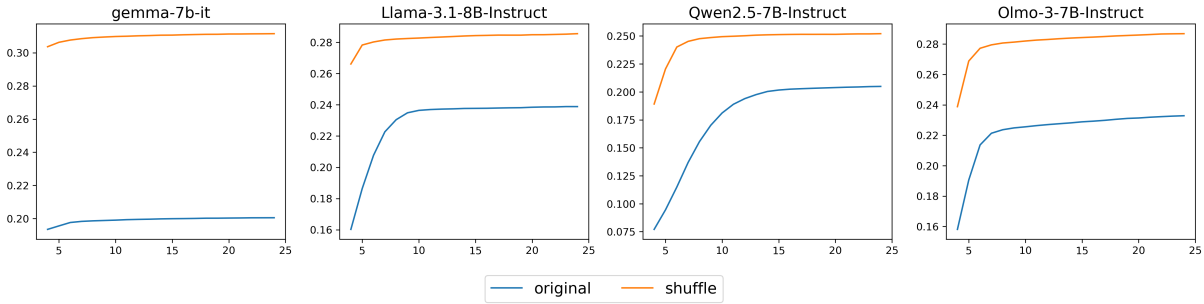

Figure 6: Curves showing how the model compositionality quantified by compositional generalization tests varies over epochs under the original and shuffle settings.

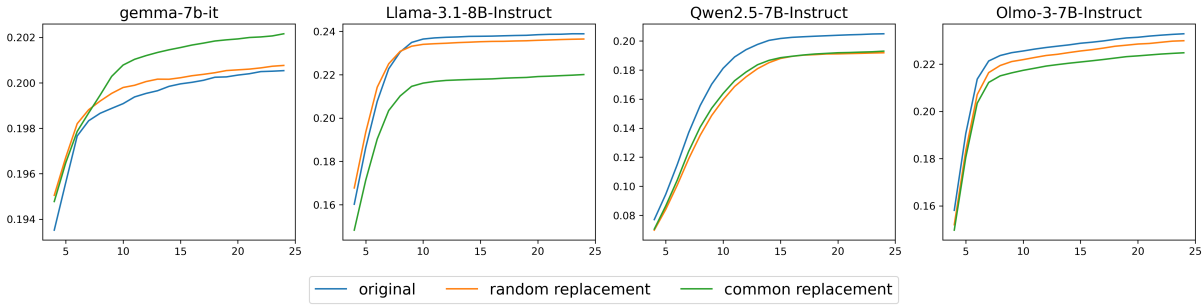

Figure 7: Curves showing how the model compositionality quantified by compositional generalization tests varies over epochs under the original and replacement settings.

model still exhibits different compositionality under the component-replacement setting compared to the original setting, and conclusions similar to those obtained using Topographic Similarity can be drawn for models other than gemma-7b-it. This indicates that even when a natural compositional structure is predefined, the compositionality development of the model during fine-tuning is not entirely symbolic, and in some cases, it still exhibits patterns consistent with those observed when using quantification methods that do not predefine a compositional structure.

4.2 Discussion of the Phenomenon

In the field of emergent communication, prior research has demonstrated that limited data exposure due to low commonality leads neural networks to develop higher compositionality (Sevestre and Dupoux, 2025). However, the differences in compositionality observed across various settings emerge during fine-tuning that is symbolically identical. The training data used for fine-tuning exhibits no differences in the degree of data exposure. Therefore, these compositionality differences cannot be attributed to data exposure differences during fine-tuning.

Past research indicates that pre-training on natural language leads to language models exhibiting poorer learning capabilities for languages that share the same symbolic essence as natural language but are linguistically impossible (Kallini et al., 2024). We therefore conjecture that the influence of structure and component factors on the compositionality development of language models during fine-tuning stems from characteristics inherent to language models that arise from pre-training on natural language text. This characteristic manifests during fine-tuning, preventing compositionality development from being entirely symbolic. We further find that the relationship between fine-tuning-based model compositionality and the settings is consistent with the general patterns observed in linguistic compositionality utilization:

Uncommon linguistic structures lead to high model compositionality. Our empirical results demonstrate that models exhibit heightened compositionality when processing infrequent linguistic structures. This phenomenon is consistent with human linguistic intuition, which suggests that novel or complex syntax necessitates a departure from heuristic-based pattern matching. In human cognition, the absence of familiar structural templates compels a more rigorous application of systematic rules (Hupkes et al., 2020) to decode meaning. Sim-

ilarly, the compositionality development of models suggests that when shortcut solutions facilitated by high-frequency structures are unavailable, the model makes greater use of compositionality.

When linguistic structures are common, uncommon linguistic components lead to high combinability, while uncommon combinations of linguistic components lead to low compositionality. The observed divergence in model compositionality regarding rare components versus rare combinations resonates with the nature of human linguistic productivity (Chomsky, 1956; Hupkes et al., 2020). First, the high compositionality observed with uncommon components aligns with the intuition that novel lexical items naturally trigger a mechanism of component independence. In human cognition, when a word is unfamiliar but its syntactic role is clear, we rely on the systematic rules of the language to integrate it. Conversely, the compositionality degradation on rare combinations of common components reflects a conflict between compositionality and frequency-based priors. In humans, highly unusual pairings of common words often cause semantic interference because they defy our statistical expectations of the world. In such cases, we tend to interpret these combinations as having some special meaning as a whole, rather than explaining them through linguistic compositionality. The compositionality exhibited by the model aligns with these tendencies.

The correlation between the observed patterns of fine-tuning-based model compositionality and linguistic compositionality provides some support for the former serving as a proxy for the latter.

4.3 Discussion of the Controversy

A contentious issue regarding the use of a particular metric as a proxy for a model's degree of utilization of linguistic compositionality is that the concept of "degree of utilization of linguistic compositionality" itself lacks a fundamental quantitative definition. When using a metric as a proxy for a property, one validation method involves specifying quantitative values for the property that is considered absolutely correct, and then assessing the numerical correlation between the metric and the values (for example, verifying the correlation between a metric and expert annotations). However, regardless of whether the subject is a human or a model, there is still no universally accepted quantitative method for measuring the degree of utilization of linguistic compositionality that can be considered absolutely correct. This means that when considering the use of a particular metric as a proxy for a model's degree of utilization of linguistic compositionality, we are typically unable to rely on numerical quantitative comparisons and must instead focus more on qualitative analysis based on numerical data (Elmoznino et al., 2025; Xu et al., 2026). Therefore, in this work, when discussing the relationship between the fine-tuning-based model compositionality and the utilization of linguistic compositionality, we have adopted a numerically based qualitative analysis rather than a specific quantitative analysis.

5 Conclusion

In this work, we investigate two issues on the proxy chain that uses fine-tuning-based metrics to quantify a model's utilization of linguistic compositionality (fine-tuning-based metrics $\rightarrow$ model compositionality $\rightarrow$ utilization of linguistic compositionality). We introduce Topographic Similarity, a method for quantifying compositionality that does not require predefined compositional structures, and compare it with the traditional compositional generalization test, which does require such predefined structures. For the two key concepts in linguistic compositionality—structure and component—we construct fine-tuning datasets that share the same symbolic essence but differ linguistically, and use the quantification results to examine whether there is a correlation between the fine-tuning test results and the model's utilization of linguistic compositionality. Regarding the two issues on the proxy chain, our main findings are as follows: (1) whether a predefined compositional structure has a significant impact on the quantification results, indicating that the spontaneous development of compositionality in models during fine-tuning may not necessarily follow the expected direction; (2) the development of compositionality in models during fine-tuning is not entirely symbolic and is influenced by linguistic factors; this influence demonstrates a connection between fine-tuning-based model compositionality and the model's utilization of linguistic compositionality. Our work offers a new perspective on methods for quantifying model compositionality and, to some extent, provides an explanation for why fine-tuning-based model compositionality can serve as a proxy for the utilization of linguistic compositionality.

Limitations

Although we have attempted to explain issues I1 and I2 on the proxy chain in our work, there are still some limitations in these two sections.

For I1, the spontaneous direction of model compositionality development remains unclear. Our comparison of quantification methods using both predefined and non-predefined compositional structures has shown that the development of model compositionality may follow a spontaneous direction that does not align with predefined expectations. However, it is not yet clear what specific compositional structures this spontaneous development will take. At present, we cannot determine whether such compositional structures can be explicitly characterized through the output or whether it is necessary to probe the model's internal state. We will continue to investigate this issue.

For I2, we cannot explain it through quantitative correlation analysis. As discussed in Section 4.3, the lack of a substantive quantitative definition means that we must rely solely on qualitative analysis based on numerical data when explaining I2. We will continue to explore whether viable quantitative analytical methods exist.

Ethics Statement

We comply with the license to use language models for scientific research purposes only. The datasets we use do not contain any information that names or uniquely identifies individual people or offensive content.

We use the translation tool DeepL for our writing, which may provide translations based on LLMs. We also use Gemini for polishing. Nevertheless, we meticulously review the content to avoid potential issues.

References

Henry Brighton and Simon Kirby. 2006. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, 12(2):229–242.

N. Chomsky. 1956. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124.

Eric Elmoznino, Thomas Jiralerspong, Yoshua Bengio, and Guillaume Lajoie. 2025. A complexity-based theory of compositionality. *Preprint*, arXiv:2410.14817.

Gemma-Team. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Chengang Hu, Xiao Liu, and Yansong Feng. 2023. DiNeR: A large realistic dataset for evaluating compositional generalization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14938–14947, Singapore. Association for Computational Linguistics.

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 2020*, pages 5065–5069. ijcai.org.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2022. State-of-the-art generalisation research in NLP: a taxonomy and review. *CoRR*, abs/2210.03050.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14691–14714. Association for Computational Linguistics.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9087–9105. Association for Computational Linguistics.

S. Kirby. 2001. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan,*

Stockholm, Sweden, July 10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, pages 2879–2888. PMLR.

Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, and Emily Cheng. 2025. Geometric signatures of compositionality across a language model's lifetime. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5292–5320. Association for Computational Linguistics.

Llama-Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Olmo-Team. 2025. Olmo-3-7b-instruct.

Peter Pagin and Dag Westerståhl. 2010. Compositionality i: Definitions and variants. *Philosophy Compass*, 5(3):250–264.

Qwen-Team. 2024. Qwen2.5: A party of foundation models.

Jean-Baptiste Sevestre and Emmanuel Dupoux. 2025. Frequency & compositionality in emergent communication. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27250–27262, Suzhou, China. Association for Computational Linguistics.

Zoltán Gendler Szabó. 2004. Compositionality. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Ziyao Xu, Cong Wang, and Houfeng Wang. 2026. Investigating more explainable and partition-free compositionality estimation for llms: A rule-generation perspective. *Preprint*, arXiv:2604.27340.

Ziyao Xu and Houfeng Wang. 2024. SPOR: A comprehensive and practical evaluation method for compositional generalization in data-to-text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 604–621, Bangkok, Thailand. Association for Computational Linguistics.

A Candidate Words

Tables 1 and 2 show the candidate words for each component.

Component	Candidate Words
place.ADJ	quiet, noisy, clean, dirty, bright, dim, modern, traditional, spacious, cramped, comfortable, plain, luxurious, shabby, safe, dangerous, peaceful, bustling, tidy, messy, warm, cool, humid, dry, ventilated, stuffy, fresh, polluted, barren, advanced, backward, lifeless, deserted, private, historic, newly-built, dilapidated, cozy, ornate, crowded, public, terrifying, dedicated, versatile, abandoned, vacant, heavily-guarded, fireproof, oppressive, relaxing
place.N	library, supermarket, school, hospital, park, restaurant, cinema, mall, bank, postoffice, airport, station, cafe, bar, gym, pool, museum, gallery, theater, stadium, zoo, arcade, bookstore, market, office, apartment, house, classroom, hotel, dormitory, garage, basement, attic, kitchen, bedroom, bathroom, church, temple, courthouse, prison, clinic, laboratory, factory, farm, garden, playground, parking, lobby, elevator, stairwell
action.VI	stand, sit, lie, sleep, run, walk, jump, laugh, cry, crawl, fly, rest, wait, think, dream, sing, dance, play, work, study, read, write, breathe, yawn, sneeze, cough, shiver, tremble, smile, shout, whisper, spin, roll, wriggle, complain, pray, meditate, slip, glide, stomp, bend, stretch, linger, hide, squat, kneel, tiptoe, blink, bow, wander
animal.N	dog, cat, cow, pig, chicken, duck, goose, horse, sheep, goat, rabbit, mouse, bear, lion, tiger, leopard, wolf, fox, elephant, giraffe, zebra, hippopotamus, rhinoceros, crocodile, shark, whale, dolphin, eagle, pigeon, sparrow, crow, peacock, parrot, snake, lizard, turtle, frog, butterfly, bee, ant, spider, mosquito, cockroach, cricket, kangaroo, panda, koala, penguin, monkey, gorilla
plant.N	apple, banana, grape, strawberry, watermelon, pineapple, mango, peach, lemon, pomelo, papaya, tomato, potato, carrot, cucumber, cabbage, spinach, onion, garlic, broccoli, eggplant, rose, sunflower, lily, orchid, jasmine, peony, chrysanthemum, tulip, daisy, lavender, cactus, oak, maple, willow, pine, palm, birch, cherry, ginkgo, moss, fern, grass, mint, aloe, ivy, lotus, wheat, rice, corn
person.ADJ	kind, intelligent, brave, friendly, diligent, humorous, honest, confident, optimistic, patient, lazy, selfish, rude, arrogant, timid, pessimistic, hypocritical, irritable, stupid, jealous, tall, short, fat, thin, young, old, outgoing, introverted, serious, lively, gentle, strong, fragile, enthusiastic, indifferent, careful, careless, cheerful, melancholy, generous, stingy, cautious, impulsive, reliable, unreliable, creative, conservative, open-minded, steady, easygoing

Table 1: The candidate words for the first six components.

Component	Candidate Words
job.N	doctor, teacher, engineer, lawyer, nurse, police-officer, firefighter, chef, driver, programmer, accountant, salesperson, manager, waiter, architect, artist, musician, writer, journalist, scientist, pilot, soldier, civil-servant, barber, farmer, worker, businessperson, consultant, designer, translator, actor, director, photographer, editor, researcher, dentist, veterinarian, pharmacist, psychologist, mechanic, electrician, plumber, courier, librarian, professor, judge, student, athlete, tailor, welder
action.ADV	happily, sadly, angrily, excitedly, patiently, hurriedly, easily, arduously, carefully, carelessly, bravely, fearfully, confidently, hesitantly, resolutely, gently, roughly, enthusiastically, coldly, sincerely, hypocritically, diligently, lazily, painfully, calmly, nervously, naturally, quickly, slowly, gracefully, clumsily, seriously, dazedly, stealthily, casually, actively, passively, wearily, bewilderedly, contentedly, composedly, anxiously, attentively, distractedly, smugly, desperately, numbly, unhurriedly, restlessly, blankly
action.VT	kicks, throws, licks, eats, scratches, bites, carries, hugs, rubs, hits, holds, squeezes, grabs, punches, pats, touches, pushes, pulls, lifts, moves, rotates, shakes, presses, cuts, destroys, repairs, cleans, washes, heats, observes, examines, analyzes, describes, finds, hides, shows, protects, uses, makes, designs, draws, measures, wraps, isolates, hangs, monitors, scans, kisses, curses, twists
object.ADJ	red, blue, green, yellow, black, white, purple, orange, brown, gray, pink, cyan, transparent, huge, tiny, round, square, triangular, oval, rectangular, fragrant, smelly, odorless, smooth, rough, soft, hard, sticky, new, opaque, worn-out, cheap, expensive, durable, premium, inferior, practical, flashy, classic, fashionable, outdated, user-friendly, eco-friendly, portable, custom-made, handmade, sturdy, lightweight, minimalist, exquisite
texture.ADJ	plastic, wooden, metal, glass, ceramic, leather, rubber, paper, stone, iron, steel, aluminum, copper, gold, silver, cotton, silk, wool, bamboo, titanium, brass, bronze, marble, granite, concrete, sponge, wax, nylon, polyester, silicone, linen, canvas, velvet, carbon-fiber, acrylic, resin, zinc, enamel, memory-foam, rattan, cork, chrome-plated, corduroy, beeswax, nephrite, graphite, mica, pearl, zirconia, silicon-carbide
object.N	necklace, bottle, table, chair, book, pen, mobile-phone, computer, cup, bowl, spoon, fork, knife, lamp, clock, watch, shoes, clothing, pants, hat, bag, wallet, key, glasses, headphones, television, refrigerator, air-conditioner, fan, mirror, toothbrush, towel, pillow, quilt, sofa, bookshelf, vase, plate, pot, pan, kettle, oven, broom, keyboard, hammer, hanger, suitcase, umbrella, basket, scissors

Table 2: The candidate words for the last six components.