

# How Do In-Context Examples Affect Compositional Generalization?

Shengnan An<sup>\*†</sup>, Zeqi Lin<sup>‡</sup>, Qiang Fu<sup>‡</sup>, Bei Chen<sup>‡</sup>,  
Nanning Zheng<sup>†</sup>, Jian-Guang LOU<sup>‡</sup>, Dongmei Zhang<sup>‡</sup>

<sup>†</sup> Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University

<sup>‡</sup> Microsoft Corporation

{an1006634493@stu, nnzheng@mail}.xjtu.edu.cn

{Zeqi.Lin, qifu, beichen, jlou, dongmeiz}@microsoft.com

## Abstract

Compositional generalization—understanding unseen combinations of seen primitives—is an essential reasoning capability in human intelligence. The AI community mainly studies this capability by fine-tuning neural networks on lots of training samples, while it is still unclear whether and how in-context learning—the prevailing few-shot paradigm based on large language models—exhibits compositional generalization. In this paper, we present COFE, a test suite to investigate in-context compositional generalization. We find that the compositional generalization performance can be easily affected by the selection of in-context examples, thus raising the research question what the key factors are to make good in-context examples for compositional generalization. We study three potential factors: similarity, diversity and complexity. Our systematic experiments indicate that in-context examples should be structurally similar to the test case, diverse from each other, and individually simple. Furthermore, two strong limitations are observed: in-context compositional generalization on fictional words is much weaker than that on commonly used ones; it is still critical that the in-context examples should cover required linguistic structures, even though the backbone model has been pre-trained on large corpus. We hope our analysis would facilitate the understanding and utilization of in-context learning paradigm.

## 1 Introduction

Compositional generalization is an essential capability of human intelligence. It means to understand and producing novel expressions by recombining known components in language (Chomsky, 1957; Montague, 1974; Fodor and Lepore, 2002). Taking examples in Figure 1, after learning the combination “*baby in a room*”, human intelligence can easily generalize to “*Jackson in a room*”. On exploring this human-like capability

<sup>\*</sup>Work done during an internship at Microsoft Research.

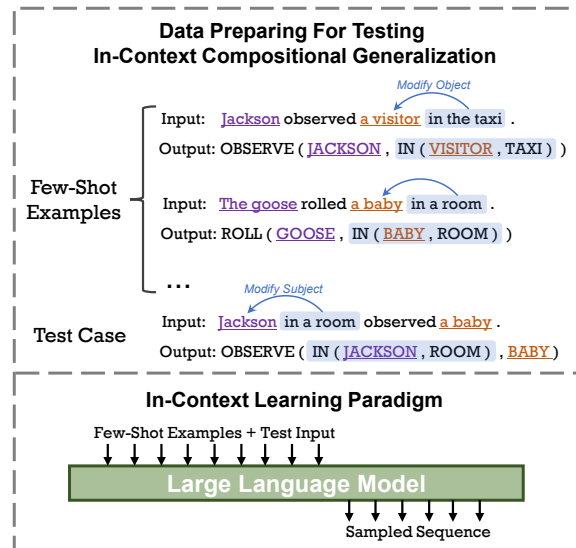


Figure 1: Test compositional generalization under in-context learning. This case belongs to *Phrase Recombination* in COFE. The phrases modify the objects in examples but are recombined with subject in test input.

in deep learning models, several benchmarks such as SCAN (Lake and Baroni, 2018), CFQ (Keysers et al., 2019) and COGS (Kim and Linzen, 2020) have been proposed based on semantic parsing<sup>1</sup> tasks. In these benchmarks, the training set cover all the primitives while lacking certain combinations, and the test set focuses on these missing combinations. By fine-tuning generic neural models on these benchmarks, much work reported that these models exhibit poor compositional generalization (Furrer et al., 2020; Shaw et al., 2021; Bogin et al., 2022).

Recently, in-context learning with large language models exhibits impressive performance on various tasks (Brown et al., 2020; Rae et al., 2021; Wei et al., 2022). By conditioning on few-shot in-context examples, the pre-trained language model, with extremely large model size and pre-trained

<sup>1</sup>Semantic parsing means translating natural language (NL) expressions into semantic representations (i.e., logical forms).

corpus, can perform downstream tasks without any update on pre-trained parameters.

Behind the impressive performance of in-context learning, we are curious whether this prevailing paradigm can take a step towards compositional generalization. To investigate this, we first take an initial exploration: for each test case in COGS, we select in-context examples from its training set and ensure that all primitives in each test case are covered by the equipped in-context examples. Our initial exploration suggests that compositional generalization can be easily affected by in-context examples: with only covering primitives, davinci 175B lags behind fine-tuned GPT2-Large with 24.2% accuracy (similar to the observation in Qiu et al. (2022)); with also covering some local structures (inspired by Bogin et al. (2022)), davinci outperforms fine-tuned GPT2-Large with 3.9% accuracy. Based on these initial observations, we raise and investigate the question: *How do in-context examples affect compositional generalization?*

We construct the test suite COFE (based on COGS) to facilitate our systematic investigation. Taking the coverage of primitives as a basic principle in COFE, we further define and inject three factors in selecting in-context examples: similarity, diversity, and complexity. Similarity is considered as the matching of hidden structures behind concrete expressions. Diversity reflects whether the context presents repeated patterns or not. Complexity portrays the amount of information contained in each example. By controlling these factors in constructing COFE, we can systematically investigate how would in-context examples influence the performance on compositional generalization.

Our experiments demonstrate that all three factors matter for in-context compositional generalization. We leverage six large language models in GPT series: davinci, code-cushman-001, code-cushman-002, text-davinci-002, text-chat-davinci-002, and code-davinci-002. The observations are consistent across models: to better perform compositional generalization, all backbone models prefer in-context examples with higher structural similarity to the test case, higher diversity among different examples, and lower complexity in each individual example. Furthermore, beyond the influence from these factors, in-context compositional generalization still faces two challenges. One is that in-context learning has difficulty recombining fictional words (e.g., random tokens) rather than com-

monly used ones. The other one is that in-context examples are still required to cover the linguistic structures in NL expressions, even though the backbone model has been pre-trained on large corpus.

Our contributions are three-fold: 1) to answer the research question posed, we investigate three factors in selecting in-context examples and draw consistent conclusions across models; 2) we construct COFE to conduct our systematic investigation, and will release it to facilitate further exploration of in-context compositional generalization; 3) we also point out two remaining challenges that in-context learning still struggles to handle. We hope our analysis would provide insights on how to select proper in-context examples, and to shed light on the future research of in-context compositional generalization. COFE is publicly available at <https://github.com/microsoft/ContextualISP/tree/master/cofe>.

## 2 In-Context Compositional Generalization

In-context compositional generalization refers to understand and produce novel combinations through recombining the building blocks presented by in-context examples. We first introduce some basic settings for testing this desired capability, then show our initial observations.

### 2.1 Principles for Measuring In-Context Compositional Generalization

To measure in-context compositional generalization under a test suite, each test case and its equipped in-context examples should satisfy two principles.

- **Combination held-out principle:** to test generalization on certain combinations, in-context examples should exclude these combinations while test cases contain them.
- **Primitive coverage principle:** the **primitives** contained in each test case should be fully covered by in-context examples. Primitives are the minimum indivisible units in expressions. In this work, we mainly consider primitives as **lexical items** (e.g., the noun “*baby*” and the verb “*observed*” in Figure 1).

We say that a model exhibits in-context compositional generalization if it performs well on a test suite that satisfies these two principles.

Category	In-Context Examples	Test Case	Illustration of Combination
<b>Primitive Substitution</b>	input: <u>shark</u> output: SHARK input: <u>A girl</u> drew the boy . output: DRAW ( GIRL , BOY , NONE )	input: <u>The shark</u> drew a boy . output: DRAW ( SHARK , BOY , NONE )	
<b>Primitive Structural Alternation</b>	input: The goose <u>baked</u> . output: BAKE ( GOOSE , NONE , NONE ) input: A teacher <u>noticed</u> a chicken . output: NOTICE ( TEACHER , CHICKEN , NONE )	input: A teacher <u>baked</u> the chicken . output: BAKE ( TEACHER , CHICKEN , NONE )	
<b>Phrase Recombination</b>	input: Logan mailed Stella <u>the cake in the pile</u> . output: MAIL ( LOGAN , IN ( CAKE , PILE ) , STELLA ) input: The goose rolled <u>a baby in a room</u> . output: ROLL ( GOOSE , IN ( BABY , ROOM ) , NONE )	input: <u>A visitor in the pile</u> rolled a resident . output: ROLL ( IN ( VISITOR , PILE ) , RESIDENT , NONE )	
<b>Longer Chain</b>	input: The boy admired that Noah confessed that \ Emma was given a cookie . output: ADMIRE ( BOY , NONE , NONE ) \ CCOMP CONFESS ( NOAH , NONE , NONE ) \ CCOMP GIVE ( NONE , COOKIE , EMMA )	input: The girl wished that a crocodile declared that \ the boy admired that Emma liked that \ Evelyn was passed a drink . output: WISH ( GIRL , NONE , NONE ) \ CCOMP DECLARE ( CROCODILE , NONE , NONE ) \ CCOMP ADMIRE ( BOY , NONE , NONE ) \ CCOMP LIKE ( EMMA , NONE , NONE ) \ CCOMP PASS ( NONE , DRINK , EVELYN )	
<b>Deeper Nesting</b>	input: Noah appreciated a girl in a house \ beside the chair . output: APPRECIATE ( NOAH , \ IN ( GIRL , \ BESIDE ( HOUSE , CHAIR \ ) ) , NONE )	input: A dog painted the girl beside the chair \ in a house beside a road on a dish . output: PAINT ( DOG , \ BESIDE ( GIRL , \ IN ( CHAIR , \ BESIDE ( HOUSE , \ ON ( ROAD , DISH \ ))) ) , NONE )	

Figure 2: Five categories of aiming combinations. The key parts in combinations are marked with underlines and colors (blue in NL-side and purple in code-side). The last column follows the notations defined in Section 3.2.

## 2.2 COGS (Under In-Context Learning)

COGS is a compositional generalization benchmark designed for the fine-tuning paradigm: based on a semantic parsing task, the training set of COGS covers all primitives in this task, while several combinations of primitives in the test set are excluded from the training set. We term these excluded combinations as **aiming combinations**.

We measure in-context compositional generalization based on COGS, by converting it from the original fine-tuning paradigm to the in-context learning paradigm. For each COGS test case, we select in-context examples from the training set  $\mathcal{B}$ , ensuring that the two principles are satisfied. Note that, for each test case, there are usually different collections of in-context examples satisfying the two principles. Our basic setting is to use a random one among them, and we show that this casual strategy could lead to an underestimation of in-context compositional generalization (Section 2.3).

To facilitate testing on more complex logical forms, we reconstruct some target-side clauses from the chain structure into the nested-function format (illustrated in Figure 2). This reconstruction follows An et al. (2023) and is similar to the conversion from Lambda calculus to FunQL in Geo domain (Zelle and Mooney, 1996; Kate et al., 2005; Zettlemoyer and Collins, 2012). Moreover, to improve human readability, we omitted two types of details: the special marker for definite descriptions and the Skolem constants. These details do not affect the testing of compositional generalization.

Apart from these omitted details, the logical forms in COFE unambiguously represent the main semantics in the domain of COGS, such as semantic roles, modifications, and orders among clauses and modifications. More details about COFE logical forms are contained in Appendix A.

**Categories of aiming combinations.** The aiming combinations in COGS can be divided into five categories, of which two are **low-level combinations** (i.e., focusing on specific primitives) and three are **high-level combinations** (i.e., focusing on high-level structures), illustrated in Figure 2.

- Primitive Substitution (*PrimSubs*): Compose a primitive (e.g., “shark”) with a grammatical role (e.g., “subject”).
- Primitive Structural Alternation (*PrimAlte*): Compose a primitive (e.g., “baked”) with a sentence structure (e.g., “subj. verb obj.”).
- Phrase Recombination (*PhraReco*): Compose a prepositional phrase (e.g., “A in B”) with a grammatical role (e.g., “subject”).
- Longer Chain (*LongChain*): Extend the tail of the logical form with CCOMP clauses  $\in \mathbf{Y}_S^1$ . The max recursive times of CCOMP clauses in  $\mathcal{B}$  is 2, while in test case it is 12.
- Deeper Nesting (*DeepNest*): Expand the arguments in functions with IN/ON/BESIDE clauses  $\in \mathbf{Y}_S^1$ . The max recursive times in  $\mathcal{B}$  and test cases are the same with *LongChain*.

Note that *PrimSubs* and *PrimAlte* are low-level combinations while others are high-level ones.

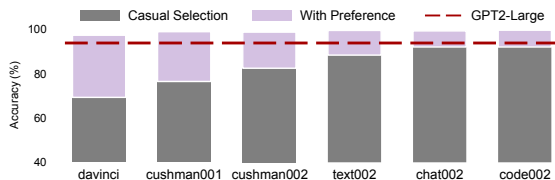


Figure 3: Initial observations on *PrimSubs*: casual selection leads to low performance while adding preference brings considerable gains.

### 2.3 In-Context Learning vs Fine-Tuning

Compositional generalization under the fine-tuning paradigm has been widely studied (Furrer et al., 2020; Shaw et al., 2021; Bogin et al., 2022), while there is little observation under in-context learning. To first get a general sense about in-context compositional generalization, we conduct an initial exploration to compare with a fine-tuning baseline.

**Models and setups.** We test in-context compositional generalization with six large models in GPT series: davinci, code-cushman-001 (cushman001), code-cushman-002 (cushman002), text-davinci-002 (text002), text-chat-davinci-002 (chat002), and code-davinci-002 (code002). The sampling temperature is 0 (i.e., greedy decoding), and the max decoding length is 500. The reported metric is exact-match accuracy. To set a fine-tuning baseline, we take GPT2-Large with 0.7B parameters. We fine-tune it on the whole  $\mathcal{B}$  and test without in-context examples. We set learning rate as  $1e-5$  and batch size as 8 during fine-tuning, and set beam size as 5 for inference. Appendix B includes more details.

**Casual selection leads to low performance of in-context compositional generalization.** For selecting in-context examples, we first take a *casual selection*: while satisfying the primitive coverage principle, we randomly select 10 examples without other preference. We conduct initial exploration on *PrimSubs* category. Figure 3 shows that under the casual selection, all six models lag behind the fine-tuned GPT2-Large on *PrimSubs*. In particular, although the size of davinci is more than 200 times that of GPT2-Large, there is a 24.2% accuracy gap between davinci and the fine-tuned GPT2-Large. These observations are close to Qiu et al. (2022).

However, we suppose the potential of in-context learning is still not fully revealed. Specifically, the selection of in-context examples does not yet take full advantage of available examples in  $\mathcal{B}$ . In next try, while still following the primitive coverage

principle, we consider injecting some additional preference in the selection of in-context examples.

**Preference in selection could bring huge improvement on *PrimSubs*.** Inspired by Bogin et al. (2022) that suggests the influence of unobserved local structures, we consider to prioritize examples that have similar hidden structures to the test case. Figure 3 shows that with this preference in selection, results on *PrimSubs* hugely change: davinci now outperforms the fine-tuned GPT2-Large; code-davinci-002 even performs near-perfectly. These changes strongly suggest that the selection of in-context examples can significantly affect in-context compositional generalization.

Based on these initial results, to further reveal the potential of in-context learning, we perform in-depth investigations on how the selection of in-context examples affects compositional generalization.

## 3 Factors Under In-Context Examples

To facilitate our systematic investigation, we construct **COFE** (C**OM**positional generalization with **F**EW-shot examples), which is derived from COGS. For selecting in-context examples in constructing COFE, we identify, inject, and control three potential factors: similarity, diversity, and complexity.

### 3.1 Conceptual Definitions

We first give conceptual definitions of our considered factors and discuss our intuitions behind them.

**Similarity** has been widely considered as the main factor in selecting in-context examples (Liu et al., 2022; Shin et al., 2021; Rubin et al., 2021; Poesia et al., 2021). The primitive coverage principle can be regarded as a basic *lexical similarity* on the surface of expressions. Beyond this surface similarity, we consider that the **structural similarity** hidden behind expressions could be a beneficial factor. From the view of syntactic structure, the recombination of primitives is equivalent to the reconstruction of the parse tree. Similar structures would ease the difficulty of recombination because the model does not need to completely reconstruct the entire structure of in-context examples. Moreover, some work has suggested that the challenge of compositional generalization under fine-tuning lies in unobserved structures (Keysers et al., 2019; Shaw et al., 2021; Bogin et al., 2022).

**Diversity** concerns the **repetitiveness** among in-context examples. It portrays the property among in-context examples. Specifically, the context is under low diversity if it contains many repeating patterns among in-context examples, otherwise it is under high diversity. Under in-context learning, the low diversity can easily lead to biased observations on the full task space, as there are only few examples for the model to learn. Thus, we suppose that the low diversity among examples could block in-context compositional generalization. Moreover, some work also demonstrated that the diversity in training data could affect compositional generalization under fine-tuning (Oren et al., 2021).

**Complexity** reflects the amount of information contained in each individual in-context example. The higher complexity means that the example could provide more information to the model, but these information could be redundant. In addition, the difficulty in directly learning from complex examples has been flagged at the intersection of cognitive science and machine learning (Elman, 1993; Bengio et al., 2009). Such difficulty may be more severe for in-context learning, since the parameters of the model cannot be updated to fit these complex examples. Thus, we suppose that too high complexity might hinder performance.

### 3.2 Incorporate Three Factors Into Test Suite

To inject these factors in selecting in-context examples, we design a *matching score* based on the parse trees behind concrete expressions. Formally, considering the primitive coverage, **structural similarity**, **diversity** and **complexity**, the matching score of two parse trees  $\mathbf{T}$  and  $\mathbf{T}'$  is defined as follows,

$$\text{Match}(\mathbf{T}, \mathbf{T}') = w_p \cdot |\mathbf{P}(\mathbf{T}) \cap \mathbf{P}(\mathbf{T}')| + w_s \cdot |\mathbf{S}(\mathbf{T}) \cap [\mathbf{S}(\mathbf{T}') - \mathbf{S}(\mathcal{C})]| - w_c \cdot \text{depth}(\mathbf{T}'), \quad (1)$$

in which  $\mathbf{P}(\cdot)$  contains primitives,  $\mathbf{S}(\cdot)$  contains *partial structures* (defined later),  $\mathcal{C}$  contains already selected examples,  $\mathbf{S}(\mathbf{T}') - \mathbf{S}(\mathcal{C})$  means to exclude already covered parts in  $\mathbf{S}(\mathcal{C})$  from  $\mathbf{S}(\mathbf{T}')$ , and  $\text{depth}(\cdot)$  reflects the complexity of the tree.

The meaning of three factors in Equation 1 is that: the structural similarity means covering  $\mathbf{S}(\mathbf{T})$ , the high diversity means to avoid repeatedly covering the same element in  $\mathbf{S}(\mathbf{T})$ , and the low complexity is to prioritize low-depth structures.

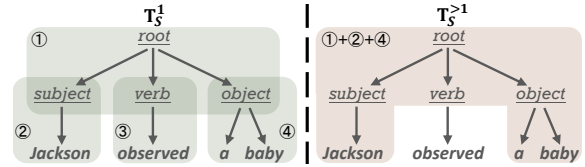


Figure 4:  $\mathbf{T}_S^1$  and  $\mathbf{T}_S^{>1}$  in the parse tree of the expression “*Jackson observed a baby*”.  $\mathbf{T}_S^1$  contains four one-depth sub-structures. We only illustrate one combination  $\in \mathbf{T}_S^{>1}$  composed from ①, ② and ④  $\in \mathbf{T}_S^1$ .  $\mathbf{T}_L$  are in **bold** and  $\mathbf{T}_N$  are with underlines.

Based on this matching score, the overall ranking score between the test case  $(\mathbf{X}, \mathbf{Y})$  and a candidate  $(\mathbf{X}_c, \mathbf{Y}_c)$  is calculated as follows,

$$\text{score}_c = \text{Match}(\mathbf{X}, \mathbf{X}_c) + \text{Match}(\mathbf{Y}, \mathbf{Y}_c), \quad (2)$$

in which both the matching of source side (i.e., NL expressions) and target side (i.e., logical forms) are considered. Poesia et al. (2021) has demonstrated the importance of target-side similarity in semantic parsing and code generation tasks, and this work will further investigate the necessity of source-side matching. In the following, we will give a more detailed description of notations in Equation 1.

**Detailed description:** Figure 4 shows an illustration of notations. Considering an expression  $e$  with the parse tree  $\mathbf{T}$ ,  $\mathbf{T}_L$  represents leaf nodes (e.g., “*Jackson*”) and  $\mathbf{T}_N$  contains internal nodes (e.g., “*subject*”).  $\mathbf{T}_S^1$  contains one-depth sub-structures in  $\mathbf{T}$ . Each  $\mathbf{T}_S^1 \in \mathbf{T}_S^1$  (e.g., ① in Figure 4) contains one parent node (e.g., “*root*”) and a set of child nodes (e.g., “*subject*”, “*verb*” and “*object*”).  $\mathbf{T}_S^{>1}$  contains deeper sub-structures that are composed from several one-depth sub-structures in  $\mathbf{T}_S^1$  (e.g., ①+②+④ in Figure 4). In Equation 1, the primitives  $\mathbf{P}(\mathbf{T}) = \mathbf{T}_L$ , and the partial structures  $\mathbf{S}(\mathbf{T}) = \mathbf{T}_S^1 \cup \mathbf{T}_S^{>1}$ . Note that aiming combinations  $\subset \mathbf{S}(\mathbf{T})$ . Appendix E includes more details.

## 4 Experiments and Analysis

### 4.1 Experimental Settings and Hyper-Parameters

We take a greedy-search algorithm to sequentially select 10 examples for each test case. Models and setups follow our initial explorations in Section 2.3. For the investigation of each factor, hyper-parameters in Equation 1 are set as follows<sup>2</sup>.

<sup>2</sup>Appendix C contains our detailed implementations.

Table 1: Results with (and without) structural similarity. Grey boxes mark the significantly better performances compared to the fine-tuned GPT2-Large.

Model	Setting	PrimSubs	PrimAlte	PhraReco	LongChain	DeepNest	Avg. Acc
code-davinci-002	Primitive Coverage	92.2	77.1	60.8	62.1	12.3	60.9
	+ Structural Similarity	99.8	99.7	65.3	87.0	26.0	75.6
text-chat-davinci-002	Primitive Coverage	92.2	75.4	47.0	65.0	6.3	57.2
	+ Structural Similarity	99.5	99.3	53.4	87.7	18.9	71.8
text-davinci-002	Primitive Coverage	88.5	66.4	38.7	46.5	2.9	48.6
	+ Structural Similarity	99.7	99.4	39.4	80.2	12.7	66.3
code-cushman-002	Primitive Coverage	82.6	55.6	21.3	29.3	5.0	38.8
	+ Structural Similarity	98.9	99.0	28.5	64.0	15.1	61.1
code-cushman-001	Primitive Coverage	76.6	60.7	16.9	5.0	1.0	32.0
	+ Structural Similarity	99.1	98.4	20.7	11.1	8.9	47.6
davinci	Primitive Coverage	69.4	52.3	9.4	2.3	0.2	26.7
	+ Structural Similarity	97.5	95.4	12.3	13.4	1.4	44.0
Fine-Tuning Baseline	-	93.6	97.9	14.0	5.4	0.0	42.2

In all settings, we prioritize the matching of primitives (i.e.,  $|\mathcal{P}(\mathbf{T}) \cap \mathcal{P}(\mathbf{T}')|$  in Equation 1) since the primitive coverage principle should be firstly satisfied. Concretely, we set  $w_p = 100$  and ensure  $w_p \gg w_s$  and  $w_c$  in all settings.

For investigating structural similarity<sup>3</sup>, we set  $w_s = 1$  and  $w_c = 0$ , and exclude  $S(\mathcal{C})$  term.

For investigating the effect of higher diversity, we add the  $S(\mathcal{C})$  term and keep other settings.

For complexity, we set  $|w_c| \cdot \max(\text{depth}(\mathbf{T}')) < w_s$ , such that the of preference of complexity will not influence the priority of structural similarity. Concretely, as  $\max(\text{depth}(\mathbf{T}')) = 12$  in COFE, we set  $w_c = 0.01$  for the low-complexity experiments and  $w_c = -0.01$  for the high-complexity experiments, and exclude  $S(\mathcal{C})$  term.

Some basic statistics for COFE under full similarity setting are listed in Table 2, and Appendix C.5 contains statistics under other settings. These statistics show that the primitive coverage principle is well satisfied, since the cover rates of  $\mathbf{T}_L$  are almost 100%. Note that the coverage on  $\mathbf{T}_S^1 \cup \mathbf{T}_S^{>1}$  must be lower than 100% since the aiming combination must be excluded.

## 4.2 Similarity

### Structural similarity brings significant gains.

Table 1 shows the performance with structural similarity. Compared to the results without structural similarity (i.e., only with the coverage on primitives), there are considerable gains on all five categories and across all six models. These gains clearly demonstrate that beyond primitive coverage, the structural similarity under in-context examples

are essential for compositional generalization.

### More precise structural similarity brings larger gains.

As mentioned in Section 3.2, the structural similarity considers to match  $S(\mathbf{T})$  which contains two parts,  $\mathbf{T}_S^1$  and  $\mathbf{T}_S^{>1}$ . Specifically, we regard that  $\mathbf{T}_S^1$  describes the *rough structure* of  $\mathbf{T}$ , and  $\mathbf{T}_S^{>1}$  determines a more *precise structure*. Based on the results in Table 1, we are curious about whether a rough structural similarity is enough. To verify this, we remove  $\mathbf{T}_S^{>1}$  from  $S(\mathbf{T})$ , which means that now we do not restrict the selected in-context examples to match precise structures in test cases. Figure 5 shows that the performances on four categories significantly drop with only a rough structural similarity, indicating that matching the precise structure of test case is still required for in-context examples. The only exception lies in *PhraReco*. It suggests that similarity is not the only influential factor for in-context compositional generalization. In Section 4.3, we will show that the low diversity and high complexity potentially cause this exception.

### With structural similarity, low-level combinations are almost solved while high-level combinations still have large room for improvement.

Specifically, for code-davinci-002, which exhibits the best performance among all backbone models, it performs near-perfectly on low-level combinations (i.e., *PrimSubs* and *PrimAlte*) while still does not achieve >95% accuracy on high-level combinations (i.e., *PhraReco*, *LongChain* and *DeepNest*). Although in-context learning greatly exceeds the fine-tuning baseline on high-level combinations, we suppose there is still potential for improvement. Compared to low-level combinations, han-

<sup>3</sup>This setting is named *full similarity setting*.

Table 2: Basic statistics of COFE.

Statistics	Number of Instances	Average Coverage				Average Length		
		$T_L$	$T_N$	$T_S^1$	$T_S^{>1}$	Context	Case Input	Case Output
Test Cases	4,785	99.7%	100%	88.9%	49.3%	297.7	17.8	33.7
- PrimSubs	1,100	100%	100%	79.8%	45.1%	236.7	7.1	11.5
- PrimAlte	700	100%	100%	96.6%	59.7%	269.4	7.9	13.8
- PhraReco	1,000	100%	100%	84.4%	19.8%	254.0	10.7	16.9
- LongChain	1,000	99.8%	100%	97.8%	76.7%	370.6	32.4	76.7
- DeepNest	985	98.8%	100%	89.0%	48.6%	356.4	29.0	46.3
Example Bank	24,155	-	-	-	-	-	7.5	10.5

ding high-level ones requires more creation than imitation, thus just considering similarity for in-context examples is not enough. In the following, we will further investigate these high-level combinations from the view of diversity and complexity.

### 4.3 Diversity and Complexity

**High diversity brings considerable gains on *PhraReco*.** Figure 6 shows how diversity among in-context examples affects generalization on high-level combinations. It shows that increasing the diversity could bring considerable gains in *PhraReco*, while not affecting the other two categories. For the performance on *PhraReco*, the improvements from higher diversity are in line with our speculations in Section 3.1, that low diversity leads to biased observations, thus blocking high-level structural generalization. For *LongChain* and *DeepNest*, beyond biased structures, their difficulty also lies in *length generalization*, thus just increasing structural diversity brings less effect to them.

**Low complexity brings considerable gains on *PhraReco*.** Figure 7 shows how the complexity in each individual example affects generalization on high-level combinations. For *PhraReco*, there are  $\sim 10\%$  gains in accuracy when the high complexity setting is changed to low complexity setting. We suppose the reason behind this gain is that simple examples could reduce the learning difficulty for the model. Moreover, simple examples also contain less redundant information thus would not confuse the model<sup>4</sup>. For *LongChain* and *DeepNest*, there is still less change on performance. Note that the max depth in these two categories is 13 while the max depth in the whole example bank is only 3. Therefore, changing the complexity of in-context examples would bring negligible influence for test cases in *LongChain* and *DeepNest*.

<sup>4</sup>Note that low and high complexity settings keep the same coverage rate on  $S(T)$ , as demonstrated in Appendix C.5.

Table 3: Results under different prompt orders (full similarity setting).  $\Delta$  represents the max difference in performance for each model.

Model	Structure Closer	Atom Closer	Random Order	$\Delta$
code-davinci-002	<b>75.6</b>	74.2	74.5	1.4
text-davinci-002	<b>66.3</b>	66.0	<b>66.3</b>	0.3
code-cushman-002	<b>61.1</b>	60.0	60.1	1.1
code-cushman-001	47.6	<b>48.2</b>	47.3	0.9
davinci	<b>44.0</b>	43.6	42.5	1.5

### 4.4 Analysis: Robustness to Prompt Order

Some previous work on in-context learning showed that the order of exemplars in prompt could sometimes hugely influence the performance of LLMs (Zhao et al., 2021; Lu et al., 2022). Here, we examine whether our observations above are sensitive to the prompt order. Based on the full similarity setting (Section 4.2), we consider three different strategies for ordering exemplars: 1) random order; 2) atom closer: exemplars with higher coverage on atomic blocks are placed closer to the test input; 3) structure closer (default): examples with higher similarity on linguistic structures are placed closer to the test input. Implementations of different strategies for prompt order are detailed in Appendix C.3.

Results in Table 3 show that the performance only slightly changes under different prompt orders. These results indicate that the main results revealed by COFE is consistent and reliable. It also indicates that in-context learning could be less sensitive to the prompt order when the in-context examples are chosen properly.

### 4.5 Discussion: Difficulty in *DeepNest*

Among all five categories, in-context learning performs worst on *DeepNest*. Compared to *LongChain* which also test recursive structures, the results on *DeepNest* still lag far behind. There is an interesting observation from the study of error cases (such as Figure 10): in-context learning frequently makes word-level mistakes, while the overall nested struc-

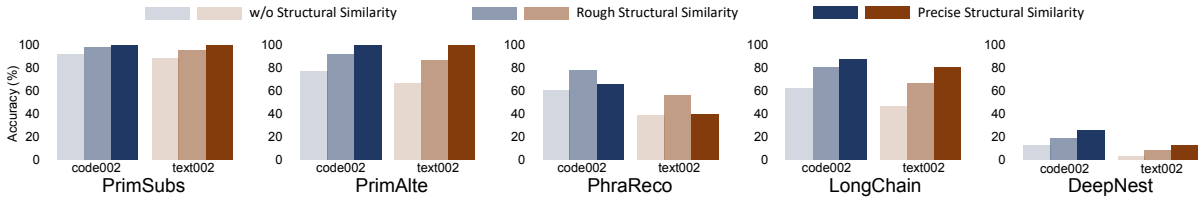


Figure 5: Performance of code-davinci-002 and text-davinci-002 with different levels of structural similarity.

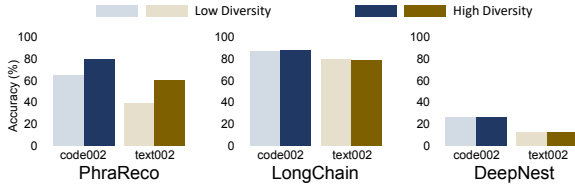


Figure 6: Performance under different diversity settings (on high-level combinations).

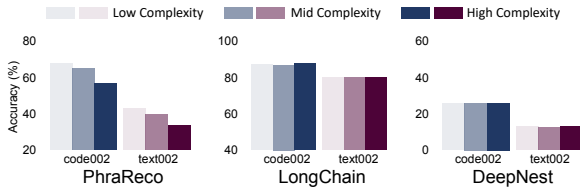


Figure 7: Performance under different complexity settings (on high-level combinations).

ture in the prediction is close to the ground truth. It suggests that **the performance bottleneck in DeepNest is to correctly fill the details in the complex structure**, rather than generating the sketch of the structure. Appendix F.1 provides further analysis.

## 5 Remaining Challenges

Our investigation has revealed a huge potential of in-context learning on performing compositional generalization<sup>5</sup>. Despite this potential, for achieving the ideal in-context compositional generalization, there remains the following two challenges.

**In-context examples are still required to match linguistic structures in NL expressions.** Since all backbone models have been pre-trained on large natural language corpus, we expect that these models could already handle the high variety in NL expressions without further hints from in-context examples. Motivated by this, we conduct experiments on another variant of COFE: the source-side term  $\text{Match}(\mathbf{X}, \mathbf{X}_c)$  is removed from Equation 2, and the coverage of  $\mathcal{S}(\mathbf{X})$  is limited (detailed in Appendix C.6). Figure 8 shows that on all five categories, the performance consistently drops if in-

<sup>5</sup>Appendix G shows the results of assembling factors.

context examples do not match the NL-side structure. It suggests that even having been pre-trained on large corpus, in-context learning still struggles to effectively recognize the semantic equivalence among different linguistic structures behind NL expressions (detailed in Appendix F.3).

**In-context learning has difficulty leveraging fictional words<sup>6</sup>.** The ideal compositional generalization requires that the recombination of primitives should be independent of the surface form in primitives. In COFE, we set the target-side primitives as the uppercase of source-side ones (e.g., “cat” → “CAT”). Such *case conversion* is commonly used in semantic parsing tasks. To test whether in-context learning could use fictional words, we replace each target-side word with random characters (e.g., replace “CAT” with “MXR”, detailed in Appendix C.7). Figure 9 shows the huge drops after changing words. Moreover, we investigate the structural accuracy by only keeping the structural terminals (e.g., parentheses and commas) in predictions. Figure 9 shows that the structural accuracy is also affected by fictional words. It indicates that on performing in-context compositional generalization, the prediction of structural sketch is not decoupled with word-level patterns.

## 6 Related Work

**Compositional generalization (CG)** has attracted much attention in NLP field. Most existing benchmarks measured CG under fine-tuning with synthetic semantic parsing tasks, suggesting the limitations of general-purpose neural networks (Lake and Baroni, 2018; Keysers et al., 2019; Kim and Linzen, 2020). Many approaches were proposed to enhance the CG on general-purpose models (Andreas, 2020; Akyürek et al., 2020; Guo et al., 2021; Oren et al., 2021; Shaw et al., 2021; Zhu et al., 2021) or design task-specific methods (Liu et al., 2020; Herzig

<sup>6</sup>The term “fictional words” means that these words are made up by us, so that large language models hardly encounter them during pre-training. Here, we generate fictional words by drawing random characters from the alphabet.



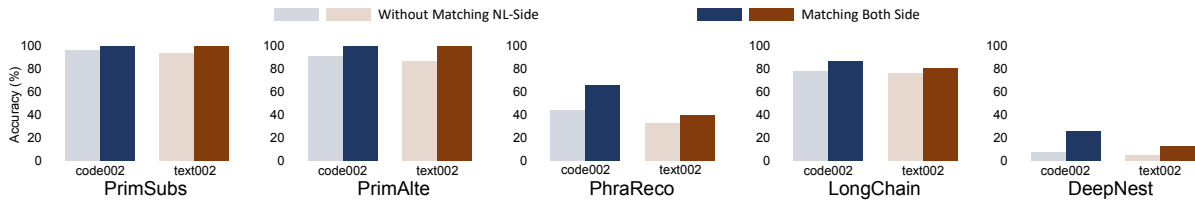


Figure 8: Performance with or without matching linguistic structures in NL expressions.

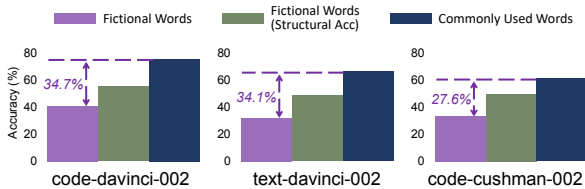


Figure 9: Average exact-match accuracy and structural accuracy with fictional words.

and Berant, 2021; Chen et al., 2020; Liu et al., 2021). Some influential factors that affect CG have been revealed, such as the length bias (Csordás et al., 2021), target-side format (Furrer et al., 2020; Herzig et al., 2021) and local structures (Bogin et al., 2022). Most existing work explored CG under the fine-tuning paradigm, while our work advances the exploration under the in-context learning paradigm.

**In-context learning** (ICL) along with large language models (LLMs) has shown surprising performance in many NLP tasks (Brown et al., 2020; Hendrycks et al., 2020; Patel and Pavlick, 2021; Rae et al., 2021; Zhang et al., 2022a; Hoffmann et al., 2022; Srivastava et al., 2022; Chowdhery et al., 2022; Smith et al., 2022; Wei et al., 2022). Most related to our work, Qiu et al. (2022) and Drozdov et al. (2022) also explored ICL on CG challenges. Qiu et al. (2022) utilized the target-side similarity on structural fragments and reported that LLMs still exhibited much poorer CG than fine-tuned small models on COGS, which is close to our initial observations. Drozdov et al. (2022) designed task-specific inference pipelines for performing CG under a least-to-most manner. Our work provides more general understandings on how to improve CG performance by revealing several factors in selecting in-context examples. In addition, some more recent work has similar observations on the potential of LLMs on CG (Hosseini et al., 2022), gains from diversity (Levy et al., 2022), and challenges under fictional words (Kim et al., 2022)

**Selection of in-context examples** is an essential

#### Ground Truth:

APPRECIATE ( CAT , BESIDE ( GIRL , IN ( BUCKET , BESIDE ( HOUSE , IN ( STAGE , BESIDE ( BOX , BESIDE ( BED , BIKE ) ) ) ) ) ) , NONE )

#### Prediction from code-davinci-002:

APPRECIATE ( CAT , BESIDE ( GIRL , **BESIDE** ( BUCKET , **IN** ( HOUSE , **BESIDE** ( STAGE , **IN** ( BOX , BESIDE ( BED , BESIDE ( BIKE , NONE ) ) ) ) ) ) ) , NONE )

Figure 10: An error case in *DeepNest* (full similarity setting) with **wrong local words** and **redundant parts**.

part for the utilization of ICL. Most existing work considered the similarity as the major metric during selection. Liu et al. (2022) selected  $k$ -nearest neighbors with similar sentence embeddings; Shin et al. (2021) regarded the conditional probability from a pre-trained LLM as the similarity score; Rubin et al. (2021) and Zhang et al. (2022b) separately trained a retriever to score the similarity; Poesia et al. (2021) and Madaan et al. (2022) estimated the target-side similarity. This work demonstrates the necessity of structural similarity in achieving CG, and also reveals the importance of two other factors beyond similarity, i.e., diversity and complexity.

## 7 Conclusion and Future Work

This work investigates how in-context compositional generalization is affected by the selection of examples. The test suite COFE is constructed to study three factors. Experiments show the effects of structural similarity, higher diversity and lower complexity. Two challenges under in-context compositional generalization are further revealed.

To apply our revealed factors outside the COFE test suite, one main challenge for future work is to determine the hidden structures behind expressions without knowing the exact generative grammar. Here, we consider two potential approaches. One is to use a pre-trained parser to generate a parse tree for the input query and then measure tree similarity. The other approach is to pre-train an embedding model with a structure-aware training objective and then compute embedding similarity.

## Limitations

**GPU resources.** This work utilizes extremely large language models and thus has a high cost on GPU resources. Concretely, experiments are conducted on the 8 x NVIDIA A100 GPU station. The maximum inference time on each version of CoFE (containing 4,785 test cases) is  $\sim 8$  hours. The maximum estimation of costed computing resources in this study is  $\sim 500 \times 8$  GPU hours.

**Synthetic data.** As in most previous work on compositional generalization (Lake and Baroni, 2018; Keysers et al., 2019; Kim and Linzen, 2020), the CoFE dataset is constructed using synthetic data rather than natural one. The source-side sentences in CoFE are from COGS, which account for 70–80% of naturally-occurring English sentences (Kim and Linzen, 2020; Roland et al., 2007). Thus, this synthetic test suite could be close to the real-world application scenarios.

**Single run.** Due to the high cost on computing resources, we do not take multiple runs with different sets of examples, nor did we take multiple samples with temperature  $> 0$ . Observations under different prompt orders (in Appendix 4.4) imply that with desired factors in selecting in-context examples, there could be low variance in experiments.

## Ethics Statement

Due to the utilization of pre-trained language models, this work could be exposed to some potential risks of ethical issues on general deep learning models (such as social bias and privacy breaches). As explored in this work that the model behavior can be hugely influenced by the provided context, we call for further investigation into how ethical issues can be avoided by controlling the provided context.

## Acknowledgments

We thank all the anonymous reviewers for their valuable comments. Shengnan An and Nanning Zheng were supported in part by NSFC under grant No. 62088102.

## References

Ekin Akyürek, Afra Feyza Akyürek, and Jacob Andreas. 2020. Learning to recombine and re-sample data for compositional generalization. In *International Conference on Learning Representations*.

Shengnan An, Zeqi Lin, Bei Chen, Qiang Fu, Nanning Zheng, and Jian-Guang Lou. 2023. Does deep learning learn to abstract? a systematic probing framework. In *The Eleventh International Conference on Learning Representations*.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Ben Bogin, Shivanshu Gupta, and Jonathan Berant. 2022. Unobserved local structures make compositional generalization hard. *arXiv preprint arXiv:2201.05899*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xinyun Chen, Chen Liang, Adams Wei Yu, Dawn Song, and Denny Zhou. 2020. Compositional generalization via neural-symbolic stack machines. *Advances in Neural Information Processing Systems*, 33:1690–1701.

Noam Chomsky. 1957. Syntactic structures (the Hague: Mouton, 1957). *Review of Verbal Behavior by BF Skinner, Language*, 35:26–58.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. 2021. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 619–634.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun

- Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models. *arXiv preprint arXiv:2209.15003*.
- Jeffrey L. Elman. 1993. Learning and development in neural networks: the importance of starting small. *Cognition*, 48:71–99.
- Jerry A Fodor and Ernest Lepore. 2002. *The compositionality papers*. Oxford University Press.
- Daniel Furrer, Marc van Zee, Nathan Scales, and Nathanael Schärli. 2020. Compositional generalization in semantic parsing: Pre-training vs. specialized architectures. *arXiv preprint arXiv:2007.08970*.
- Yinuo Guo, Hualei Zhu, Zeqi Lin, Bei Chen, Jian-Guang Lou, and Dongmei Zhang. 2021. Revisiting iterative back-translation from the perspective of compositional generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7601–7609.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Jonathan Herzig and Jonathan Berant. 2021. Span-based semantic parsing for compositional generalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 908–921.
- Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, and Yuan Zhang. 2021. Unlocking compositional generalization in pre-trained models using intermediate representations. *arXiv preprint arXiv:2104.07478*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models.
- Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordani, and Aaron Courville. 2022. On the compositional generalization gap of in-context learning. *arXiv preprint arXiv:2211.08473*.
- Rohit J Kate, Yuk Wah Wong, Raymond J Mooney, et al. 2005. Learning to transform natural to formal languages. In *AAAI*, volume 5, pages 1062–1068.
- Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.
- 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)*, pages 9087–9105.
- Najoung Kim, Tal Linzen, and Paul Smolensky. 2022. Uncontrolled lexical exposure leads to overestimation of compositional generalization in pretrained models. *arXiv preprint arXiv:2212.10769*.
- 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.
- Itay Levy, Ben Bogin, and Jonathan Berant. 2022. Diverse demonstrations improve in-context compositional generalization. *arXiv preprint arXiv:2212.06800*.
- Chenyao Liu, Shengnan An, Zeqi Lin, Qian Liu, Bei Chen, Jian-Guang Lou, Lijie Wen, Nanning Zheng, and Dongmei Zhang. 2021. Learning algebraic recombination for compositional generalization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1129–1144.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114.
- Qian Liu, Shengnan An, Jian-Guang Lou, Bei Chen, Zeqi Lin, Yan Gao, Bin Zhou, Nanning

- Zheng, and Dongmei Zhang. 2020. Compositional generalization by learning analytical expressions. *Advances in Neural Information Processing Systems*, 33:11416–11427.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.
- Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *arXiv preprint arXiv:2210.07128*.
- R Montague. 1974. English as a formal language. *Formal Philosophy: Selected Papers of Richard Montague*.
- Inbar Oren, Jonathan Herzig, and Jonathan Berant. 2021. Finding needles in a haystack: Sampling structurally-diverse training sets from synthetic data for compositional generalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10793–10809.
- Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.
- Gabriel Poesia, Alex Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. 2021. Synchromesh: Reliable code generation from pre-trained language models. In *International Conference on Learning Representations*.
- Linlu Qiu, Peter Shaw, Panupong Pasupat, Tianze Shi, Jonathan Herzig, Emily Pitler, Fei Sha, and Kristina Toutanova. 2022. Evaluating the impact of model scale for compositional generalization in semantic parsing. *arXiv preprint arXiv:2205.12253*.
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Douglas Roland, Frederic Dick, and Jeffrey L Elman. 2007. Frequency of basic english grammatical structures: A corpus analysis. *Journal of memory and language*, 57(3):348–379.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.
- Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938.
- Richard Shin, Christopher Lin, Sam Thomson, Charles Chen Jr, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715.
- Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- John M Zelle and Raymond J Mooney. 1996. Learning to parse database queries using inductive

- logic programming. In *Proceedings of the national conference on artificial intelligence*, pages 1050–1055.
- Luke S Zettlemoyer and Michael Collins. 2012. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022a. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Yiming Zhang, Shi Feng, and Chenhao Tan. 2022b. Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.
- Wang Zhu, Peter Shaw, Tal Linzen, and Fei Sha. 2021. Learning to generalize compositionally by transferring across semantic parsing tasks. *arXiv preprint arXiv:2111.05013*.

This is the Appendix of the paper: *How Do In-Context Examples Affect Compositional Generalization?*

## A Grammar

Part of the grammar used in constructing COFE is listed in Table 4. Note that the max recursive times of R-Production Rules is 2 in prompting examples and 12 in test cases. The target-side grammar follows the reconstruction in An et al. (2023). Overall, the original target grammar of COGS is reconstructed to be chain-structured. Concretely, first, the original output tokens in COGS are capitalized; then, the variables (e.g., “ $x_I$ ”) in the original grammar are aligned and replaced with their corresponding terminals; finally, the output clauses are grouped as the function format, in which the function name belongs to “*PRED-FUNC*” and the arguments are ordered as “*AGENT*”, “*THEME*”, and “*RECIPIENT*”. Moreover, if “*PRED-FUNC*” does not contain one or some arguments, the positions of these arguments are filled with “*NONE*” terminal. For the two R-Production rules in Table 4, the first is in chain structure and the second is in nested structure. Moreover, the whole nested “*PP-FUNC*” will be filled into the “*PRED-FUNC*” as an argument, rather than concatenated to the tail of the “*CLAUSE*”.

## B Details of Fine-Tuning

The fine-tuned GPT2-Large contains 762M parameters. For fine-tuning, we take 50,000 training steps with 8 batch size and 1e-5 learning rate (without warm-up strategy). We set weight decay as 1e-2 and label smoothing factor as 1e-1. For inference with GPT2-Large, we set beam size as 5 and set max length as 1,024.

## C Details of Implementation

### C.1 Algorithm

Algorithm 1 shows the greedy searching algorithm for constructing COFE.

### C.2 Key Designs

We give detailed descriptions of some key designs in Algorithm 1.

- $P(\mathbf{T})$ : Return the leaf nodes  $\mathbf{T}_L$  on the tree;
- $S(\mathbf{T})$ : Return the structural combinations on the tree, i.e.,  $\mathbf{T}_S^1 \cup \mathbf{T}_S^{>1}$ ;

---

### Algorithm 1 Greedy-Search Algorithm for Constructing COFE

---

#### Given:

$(\mathbf{X}, \mathbf{Y})$ : Source and target parse trees in one test case;  
 $\mathcal{B}$ : Example bank;  
 $(\mathbf{X}_i, \mathbf{Y}_i) \in \mathcal{B}$ : One candidate case in example bank;  
 $\mathbf{X}_A$  and  $\mathbf{Y}_A$ : Aiming combination;  
 $w_p, w_s, w_c$ : Weights for primitive coverage, structural similarity, and complexity penalty;  
 $P(\cdot)$ : primitives;  
 $S(\cdot)$ : structural combinations;

#### Return:

$C$ : Selected in-context examples;  
1:  $C = \{\}$   
2: **while**  $|C| < n$  **do**  
3:    $\text{max\_score} = 0$   
4:    $\text{candidate} = \text{None}$   
5:   **for**  $(\mathbf{X}_i, \mathbf{Y}_i) \in \mathcal{B}$  **do**  
6:     **Assert**  $\mathbf{X}_A \notin S(\mathbf{X}_i)$   
7:     **Assert**  $\mathbf{Y}_A \notin S(\mathbf{Y}_i)$   
8:      $\text{prim\_score} = 0$   
9:      $\text{stru\_score} = 0$   
10:     **for**  $\text{element} \in P(\mathbf{X}_i) \cup P(\mathbf{Y}_i)$  **do**  
11:       **if**  $\text{element} \in P(\mathbf{X}) \cup P(\mathbf{Y})$  **then**  
12:           $\text{prim\_score} += w_p$   
13:       **end if**  
14:     **end for**  
15:     **for**  $\text{element} \in S(\mathbf{X}_i) \cup S(\mathbf{Y}_i)$  **do**  
16:       **if**  $\text{element} \in S(\mathbf{X}) \cup S(\mathbf{Y})$  and  $\text{element} \notin S(C)$  **then**  
17:           $\text{stru\_score} += w_s$   
18:       **end if**  
19:     **end for**  
20:      $\text{comp\_penalty} = w_c \cdot \text{depth}(\mathbf{X}_i)$   
21:      $\text{score} = \text{prim\_score} + \text{stru\_score} - \text{comp\_penalty}$   
22:     **if**  $\text{score} > \text{max\_score}$  **then**  
23:        $\text{max\_score} = \text{score}$   
24:        $\text{candidate} = (\mathbf{X}_i, \mathbf{Y}_i)$   
25:     **end if**  
26:   **end for**  
27:    $C.\text{add}(\text{candidate})$   
28: **end while**

---

- $w_p, w_s, w_c$ : The initial scores for matching primitives, structural combinations, and complexity penalty, respectively. A higher  $w$  means that the corresponding element is prioritized in greedy search.
- $\text{element} \notin S(C)$ : Already covered elements will not be awarded again, thus encouraging high diversity.
- $\text{depth}(\mathbf{T})$ : return the depth of the tree. Note that  $\text{depth}(\mathbf{X}_i) = \text{depth}(\mathbf{Y}_i)$  in COFE.

### C.3 Prompt Order

We take the structure-closer order, i.e., the examples in  $C$  with a higher  $\text{stru\_score}$  are placed closer to the test case. In Section 4.4, we show the robustness to the other two orders: random order, i.e., all selected in-context examples in  $C$  are randomly

Table 4: Part of the grammar used in constructing COFE.

Formal English Grammar	Semantic Representation	Type
active-verb / passive-verb $\rightarrow S_v$	PRED-FUNC $\rightarrow S_P$	T-Production Rule
subject / direct-object / indirect-object $\rightarrow S_n$	AGENT / THEME / RECIPIENT $\rightarrow S_E$	
pp-mod / pp-s $\rightarrow S_n$	PP-FUNC / PP-S $\rightarrow S_E$	
conj $\rightarrow$ that	CP-CONCAT $\rightarrow$ CCOMP	
prep $\rightarrow$ in / on / beside	PP-CONCAT $\rightarrow$ IN / ON / BESIDE	
sentence $\rightarrow$ subj active-verb	CLAUSE $\rightarrow$ PRED-FUNC ( AGENT, NONE, NONE )	N-Production Rule
sentence $\rightarrow$ subj active-verb direct-obj indirect-obj	CLAUSE $\rightarrow$ PRED-FUNC ( AGENT, THEME, RECIPIENT )	
subject / direct-object / indirect-object $\rightarrow$ pp-mod	AGENT / THEME / RECIPIENT $\rightarrow$ PP-FUNC	
sentence $\rightarrow$ sentence conj sentence	CLAUSE $\rightarrow$ CLAUSE CP-CONCAT CLAUSE	R-Production Rule
pp-mod $\rightarrow$ pp-s prep pp-mod	PP-FUNC $\rightarrow$ PP-CONCAT ( PP-S, PP-FUNC )	

shuffled, and atom-closer order, i.e., the examples in  $C$  with a higher `prim_score` are placed closer to the test case.

#### C.4 Max Depth in $T_S^{>1}$

Since the max repetition times for *LongChain* and *DeepNest* are 2 (as described in Section 2.2), we set the max depth in  $T_S^{>1}$  as 2 in  $S(T)$ .

#### C.5 Similarity Under Diversity and Complexity Settings

Table 5: Statistics of different versions of COFE (*PhraReco* category).

Setting	Average Coverage			
	$T_L$	$T_N$	$T_S^1$	$T_S^{>1}$
Default (Low Diversity, Mid Complexity)	100%	100%	84.4%	19.8%
High Diversity	100%	100%	84.4%	19.8%
Low Complexity	100%	100%	84.4%	19.8%
High Complexity	100%	100%	84.4%	19.8%

While changing diversity and complexity in variants of COFE in Section 4.3, the primitive coverage and structural similarity are still satisfied. Table 5 shows that on *PhraReco*, the statistics of coverage in different diversity and complexity settings are kept identical to the full similarity setting in COFE.

#### C.6 Excluding NL-Side Matching

For excluding source-side matching in Section 5, besides removing the first term in Equation 2, we also limit the matching of  $X_S^1$ . Concretely, we require that the sentence rule in test case should not be covered by in-context examples. The sentence rule is an N-Production rule that contains the non-terminal “*sentence*” as the left hand. To achieve this, we filter out test cases that can not meet this constraint. Finally, 1,037 out of 4,785 test cases are kept in this variant of COFE.

#### C.7 Fictional Words

For each target-side word that contain  $l$  characters, we sequentially and randomly sample  $l$  characters from alphabet as a fictional word to replace the original word. In addition, for the experiments on fictional words, we take the atom-closer prompt order, since the model with this order performs better the default structure-closer order.

#### D Excluding Target-Side Matching

In Section 5, we show that the performance drops with excluding the source-side matching. Here, we examine the effect of target-side matching. For constructing data, we directly remove the second term in Equation 2. As shown in Table 6, the performances with or without target-side matching are nearly identical. Such an observation is similar to the comparison between oracle and non-oracle settings in Qiu et al. (2022) that also utilized COGS benchmark, but different from Poesia et al. (2021) which suggested the importance of target-side similarity in code generation tasks. We suppose there are mainly two reasons that could cause this difference. On the one hand, different from general code generation tasks, the test suite for compositional generalization requires the exclusion of certain aiming combinations. Therefore, the performance bottleneck in compositional generalization benchmarks mainly lies in the lacked aiming combinations. On the other hand, in most compositional generalization benchmarks, the source-side matching could largely take over the target-side matching, since the terminals and rules in source grammar in these benchmarks are mapped many-to-one to the target grammar. Therefore, when seeking for the source-side matching, the target-side matching is also improved.

Table 6: Performances under only matching source side.

Model	Setting	PrimSubs	PrimAlte	PhraReco	LongChain	DeepNest	Average
code-davinci-002	matching both side	99.8	99.7	65.3	87.0	26.0	75.6
	only matching source side	99.3	99.7	63.2	88.9	25.8	75.4
text-davinci-002	matching both side	99.7	99.4	39.4	80.2	12.7	66.3
	only matching source side	98.8	99.6	35.6	81.1	12.5	65.5
code-cushman-002	matching both side	98.9	99.0	28.5	64.0	15.1	61.1
	only matching source side	98.6	99.4	26.7	66.8	16.3	61.6
code-cushman-001	matching both side	99.1	98.4	20.7	11.1	8.9	47.6
	only matching source side	99.2	99.6	17.4	13.1	8.6	47.6
davinci	matching both side	97.5	95.4	12.3	13.4	1.4	44.0
	only matching source side	97.7	94.7	7.2	14.7	2.1	43.3

## E Illustration of Defined Notations

Figure 11 illustrates the notations defined in Section 3.2 based on a concrete expression “*Jackson in a room observed a baby*”.

Note that for all sub-structures in  $\mathbf{T}_S^1 \cup \mathbf{T}_S^{>1}$ , we require them to be complete sub-structures.

**Definition: Complete sub-structure (CSS).** A CSS is a subgraph in a tree  $\mathbf{T}$ , satisfying that if an internal node in  $\mathbf{T}$  and one of its child nodes are covered in this CSS, all other child nodes must be also covered in this CSS.

## F Case Study

We provide case study to further understanding the performance of compositional generalization observed in the main text. For ease of reading, we include the following contents in the caption of figures.

### F.1 Two Types of Errors in *DeepNest*

Figure 12 shows two error cases in *DeepNest* with code-davinci-002 model and full similarity setting. The overall structure of predictions are close to the ground truth, but the model makes mistakes on some local parts. Concretely, some local semantics are incorrect (in red), and some words are redundant (in gray).

Moreover, we also calculate the word-level coverage in predictions. Besides the instance-level accuracy, we further investigate a word-level error rate on *DeepNest*. We find that in *DeepNest*, 96.8% of the words in the ground truth are contained by the predictions from code-davinci-002 (while only 48.8% for GPT2-Large). It indicates that the low instance-level accuracy is mainly caused by the wrong positions of words and redundant words.

### F.2 Structural Errors with Fictional Words

Figure 13 shows the comparison of performance between fictional words (left) and commonly used words (right). For the provided contexts on the left and right, the only difference is that the target-side words on the left are randomly selected characters while on the right they are uppercase of the source-side words. It shows that by changing only the target-side words, the model not only makes word-level errors (i.e., missing two words “*ES*” and “*NVCWI*” in prediction), it also generates the wrong parentheses structure (i.e., generate a 2-depth structure while in ground truth it is 3-depth).

### F.3 Fail to Recognize Semantic Equivalence

Figure 14 shows the comparison of performances between excluding NL-side matching (left) and containing NL-side matching (right). For the test input “*Matthew shipped the professor a chair .*”, it contains the sentence structure “*subject verb object\_1 object\_2*” behind the NL expression. Context on the left does not explicitly contain this sentence structure, but it contains a semantically equivalent structure (i.e., “*subject verb object\_2 to object\_1*”). However, the model generates the correct prediction on the right while fails on the left. Concretely, according to the wrong prediction on the left, the model perhaps considers that the semantics of “*subject verb object\_1 object\_2*” is equivalent with “*subject verb object\_1 to object\_2*”.

### F.4 Low Diversity Block Generalization

Figure 15 shows the comparison of performances on *PhraReco* under high diversity (left) and low diversity (right). For the test input “*A girl in the house slept*”, “*subject slept*” is one element contained in  $\mathbf{T}_S^{>1}$ . This element is repeatedly covered in the context on the right (low diversity) while only covered once on the left (high diversity). However,



under high repetitiveness, the model fails on the test case, but succeed when there is low repetitiveness.

### **F.5 High Complexity Block Generalization**

Figure 16 shows the comparison of performance on *PhraReco* under low complexity (left) and high diversity (right). With low complexity, the test case is covered by simple and short in-context examples, and the model succeeds on the test case. With high complexity, the test case is covered by more complex and longer examples, and the model fails on the test case.

### **G Full Results**

Due to the page limitation for main text, here we list our full results in Section 4. The results in *Assembling* are the best performance under each category among all combinations of factors.

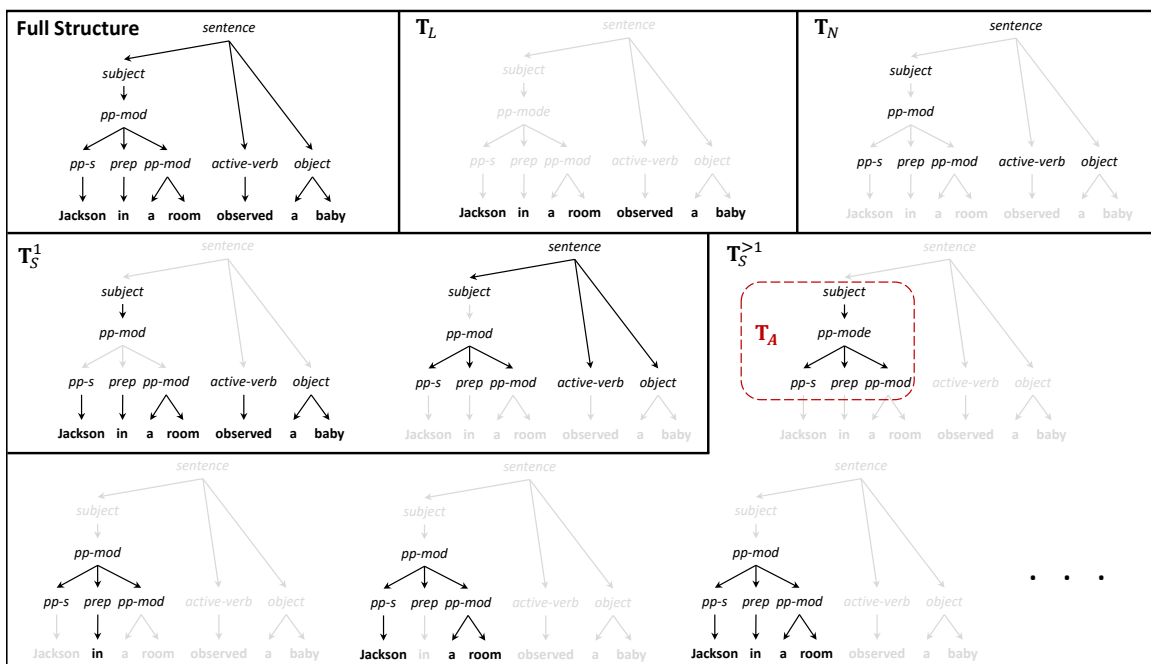


Figure 11: Illustration of defined notations. This test case belongs to *PhraReco* category.  $T_A$  means the aiming combination in this test case.

Model: code-davinci-002  
Category: DeepNest  
Setting: full similarity setting

**context:**

context:

input: A host was sold the donut on the bench beside a table .  
output: SELL ( NONE , ON ( DONUT , BESIDE ( BENCH , TABLE ) ) , HOST )

input: Liam gave the monkey a chalk in the container .  
output: GIVE ( LIAM , IN ( CHALK , CONTAINER ) , MONKEY )

input: A dog gave Emma the drink in the house .  
output: GIVE ( DOG , IN ( DRINK , HOUSE ) , EMMA )

input: The dog gave the spoon beside the table to Olivia .  
output: GIVE ( DOG , BESIDE ( SPOON , TABLE ) , OLIVIA )

input: A child was lended a cake on a stool in a garden by the dog .  
output: LEND ( DOG , ON ( CAKE , IN ( STOOL , GARDEN ) ) , CHILD )

input: Emma mailed a boy a cake on a paper in a room .  
output: MAIL ( EMMA , ON ( CAKE , IN ( PAPER , ROOM ) ) , BOY )

input: Emma gave the cake on the table on the bed to Olivia .  
output: GIVE ( EMMA , ON ( CAKE , ON ( TABLE , BED ) ) , OLIVIA )

input: A girl gave Olivia a cake on a table .  
output: GIVE ( GIRL , ON ( CAKE , TABLE ) , OLIVIA )

input: A dog gave a girl the cake beside the bed on a chair .  
output: GIVE ( DOG , BESIDE ( CAKE , ON ( BED , CHAIR ) ) , GIRL )

input: Charlotte offered the boy a cake on the tree on a plate .  
output: OFFER ( CHARLOTTE , ON ( CAKE , ON ( TREE , PLATE ) ) , BOY )

input: A dog gave Olivia a cake on the table on a stool in the container on the bench on a plate .

**label:**

GIVE ( DOG , ON ( CAKE , ON ( TABLE , IN ( STOOL , ON ( CONTAINER , ON ( BENCH , PLATE ) ) ) ) ) , OLIVIA )

**pred:**

GIVE ( DOG , ON ( CAKE , ON ( TABLE , ON ( STOOL , IN ( CONTAINER , ON ( BENCH , PLATE ) ) ) ) ) , OLIVIA )

Model: code-davinci-002  
Category: DeepNest  
Setting: full similarity setting

**context:**

input: The girl posted a teacher the cake in a condo .  
output: POST ( GIRL , IN ( CAKE , CONDO ) , TEACHER )

input: Olivia lended a politician a game beside a cup in a room .  
output: LEND ( OLIVIA , BESIDE ( GAME , IN ( CUP , ROOM ) ) , POLITICIAN )

input: Emma sold the girl a cake in the trailer beside a warrior .  
output: SELL ( EMMA , IN ( CAKE , BESIDE ( TRAILER , WARRIOR ) ) , GIRL )

input: Oliver was given a melon in the car beside a bed .  
output: GIVE ( NONE , IN ( MELON , BESIDE ( CAR , BED ) ) , OLIVER )

input: A politician packed the cake in a house .  
output: PACK ( POLITICIAN , IN ( CAKE , HOUSE ) , NONE )

input: The girl ate the cup on a glacier on the bed .  
output: EAT ( GIRL , ON ( CUP , ON ( GLACIER , BED ) ) , NONE )

input: A cat was served a cake in the pod beside a table by Liam .  
output: SERVE ( LIAM , IN ( CAKE , BESIDE ( POD , TABLE ) ) , CAT )

input: Oliver ate the jacket in a bag on a table .  
output: EAT ( OLIVER , IN ( JACKET , ON ( BAG , TABLE ) ) , NONE )

input: Lucas was given the cake in a house in a garden by a goose .  
output: GIVE ( GOOSE , IN ( CAKE , IN ( HOUSE , GARDEN ) ) , LUCAS )

input: Emma ate the cake in the drawer on a tree .  
output: EAT ( EMMA , IN ( CAKE , ON ( DRAWER , TREE ) ) , NONE )

input: A politician ate the cake in a cup in the garden beside a bed in a house in a trailer in the car on the tree in the pod in a bag on a glacier in a condo .

**label:**

EAT ( POLITICIAN , IN ( CAKE , IN ( CUP , BESIDE ( GARDEN , IN ( BED , IN ( HOUSE , IN ( TRAILER , ON ( CAR , IN ( TREE , IN ( POD , ON ( BAG , IN ( GLACIER , CONDO ) ) ) ) ) ) ) ) ) ) , NONE )

**pred:**

EAT ( POLITICIAN , IN ( CAKE , IN ( CUP , IN ( GARDEN , BESIDE ( BED , IN ( HOUSE , IN ( TRAILER , IN ( CAR , ON ( TREE , IN ( POD , IN ( BAG , ON ( GLACIER , IN ( CONDO , NONE ) ) ) ) ) ) ) ) ) ) ) , NONE )

Figure 12: Two error cases in DeepNest with code-davinci-002 model and full similarity setting. The overall structures of predictions in error cases are close to the ground truth, but the model makes mistakes on some local parts. Concretely, some local semantics are incorrect (in red), and some words are redundant (in gray).

Model: code-davinci-002  
 Category: *PhraReco*  
 Setting: fictional words (full similarity setting)

**context:**  
 input: William shortened the cookie on a tray .  
 output: UPVIENL ( ZKCJBMB , ZY ( TXHMRM , UECQ ) , NONE )

input: The girl proved that Olivia ate the baby on a stage .  
 output: QXFTO ( LAWW , NONE , NONE ) CCOMP ULX ( ESTEVR , ZY ( VYVA , PEDQY ) , NONE )

input: Elizabeth screamed .  
 output: KXSKRS ( OWKPPDAJM , NONE , NONE )

input: A hero screamed .  
 output: KXSKRS ( ZUNS , NONE , NONE )

input: The crocodile screamed .  
 output: KXSKRS ( QPAKTMJAF , NONE , NONE )

input: Olivia respected that Asher returned a ball on the table in a house to a baby .  
 output: GAFPLFB ( ESTEVR , NONE , NONE ) CCOMP IEFKEF ( AFQWF , ZY ( HFJZ , ES ( RGIJS , NVCWI ) ) , VYVA )

input: A kitty gave a scientist a cake in a can in the house .  
 output: FMBI ( CYQUL , ES ( QXGW , ES ( VSP , NVCWI ) ) , GFJNRFBWW )

input: The cat was given the mandarin in the room in the house .  
 output: FMBI ( NONE , ES ( PLNLARRT , ES ( ETIO , NVCWI ) ) , MXR )

input: The baby floated a cake beside the cabinet in the house .  
 output: PSEGL ( VYVA , CIDPYO ( QXGW , ES ( DBZKXEC , NVCWI ) ) , NONE )

input: A boy confessed that Olivia returned the girl the donut on the tripod in the house .  
 output: RKKPVRU ( CRD , NONE , NONE ) CCOMP IEFKEF ( ESTEVR , ZY ( QXKYR , ES ( OEHBXJ , NVCWI ) ) , LAWW )

input: The baby on a tray in the house screamed .

**label:**  
 KXSKRS ( ZY ( VYVA , ES ( UECQ , NVCWI ) ) , NONE , NONE )

**pred:**  
 KXSKRS ( ZY ( VYVA , UECQ ) , NONE , NONE )

Model: code-davinci-002  
 Category: *PhraReco*  
 Setting: commonly used words (full similarity setting)

**context:**  
 input: William shortened the cookie on a tray .  
 output: SHORTEN ( WILLIAM , ON ( COOKIE , TRAY ) , NONE )

input: The girl proved that Olivia ate the baby on a stage .  
 output: PROVE ( GIRL , NONE , NONE ) CCOMP EAT ( OLIVIA , ON ( BABY , STAGE ) , NONE )

input: Elizabeth screamed .  
 output: SCREAM ( ELIZABETH , NONE , NONE )

input: A hero screamed .  
 output: SCREAM ( HERO , NONE , NONE )

input: The crocodile screamed .  
 output: SCREAM ( CROCODILE , NONE , NONE )

input: Olivia respected that Asher returned a ball on the table in a house to a baby .  
 output: RESPECT ( OLIVIA , NONE , NONE ) CCOMP RETURN ( ASHER , ON ( BALL , IN ( TABLE , HOUSE ) ) , BABY )

input: A kitty gave a scientist a cake in a can in the house .  
 output: GIVE ( KITTY , IN ( CAKE , IN ( CAN , HOUSE ) ) , SCIENTIST )

input: The cat was given the mandarin in the room in the house .  
 output: GIVE ( NONE , IN ( MANDARIN , IN ( ROOM , HOUSE ) ) , CAT )

input: The baby floated a cake beside the cabinet in the house .  
 output: FLOAT ( BABY , BESIDE ( CAKE , IN ( CABINET , HOUSE ) ) , NONE )

input: A boy confessed that Olivia returned the girl the donut on the tripod in the house .  
 output: CONFESS ( BOY , NONE , NONE ) CCOMP RETURN ( OLIVIA , ON ( DONUT , IN ( TRIPOD , HOUSE ) ) , GIRL )

input: The baby on a tray in the house screamed .

**label:**  
 SCREAM ( ON ( BABY , IN ( TRAY , HOUSE ) ) , NONE , NONE )

**pred:**  
 SCREAM ( ON ( BABY , IN ( TRAY , HOUSE ) ) , NONE , NONE )

Figure 13: Comparison of performance between fictional words (left) and commonly used words (right). For the provided contexts on the left and right, the only difference is that the target-side words on the left are randomly selected characters while on the right they are uppercase of the source-side words. It shows that by changing only the target-side words, the model not only makes word-level errors (i.e., missing two words “ES” and “NVCWI” in prediction), it also generates the wrong parentheses structure (i.e., generate a 2-depth structure while in ground truth it is 3-depth).

Model: code-davinci-002  
 Category: *PrimAlte*  
 Setting: exclude NL-side matching

**context:**  
 input: A governor was passed a chalk beside the computer beside a stage by Matthew .  
 output: PASS ( MATTHEW , BESIDE ( CHALK , BESIDE ( COMPUTER , STAGE ) ) , GOVERNOR )

input: A cat offered the rose to a professor .  
 output: OFFER ( CAT , ROSE , PROFESSOR )

input: A chair was rented to Emma by a cat .  
 output: RENT ( CAT , CHAIR , EMMA )

input: A cookie was given to Hazel by Matthew .  
 output: GIVE ( MATTHEW , COOKIE , HAZEL )

input: Benjamin lendend the brain to a professor .  
 output: LEND ( BENJAMIN , BRAIN , PROFESSOR )

input: Grace was lendend the chair by a boy .  
 output: LEND ( BOY , CHAIR , GRACE )

input: A chair was sold to James by Ava .  
 output: SELL ( AVA , CHAIR , JAMES )

input: The girl shipped a cookie in a container to Scarlett .  
 output: SHIP ( GIRL , IN ( COOKIE , CONTAINER ) , SCARLETT )

input: A banana was fed to the professor by Emma .  
 output: FEED ( EMMA , BANANA , PROFESSOR )

input: Matthew lendend a cake beside the chair to the doctor .  
 output: LEND ( MATTHEW , BESIDE ( CAKE , CHAIR ) , DOCTOR )

input: Matthew shipped the professor a chair .

**label:**  
 SHIP ( MATTHEW , CHAIR , PROFESSOR )

**pred:**  
 SHIP ( MATTHEW , PROFESSOR , CHAIR )

Model: code-davinci-002  
 Category: *PrimAlte*  
 Setting: full similarity setting

**context:**  
 input: A chair was sold to James by Ava .  
 output: SELL ( AVA , CHAIR , JAMES )

input: Grace was lendend the chair by a boy .  
 output: LEND ( BOY , CHAIR , GRACE )

input: A girl awarded a professor the box .  
 output: AWARD ( GIRL , BOX , PROFESSOR )

input: The girl shipped a cookie in a container to Scarlett .  
 output: SHIP ( GIRL , IN ( COOKIE , CONTAINER ) , SCARLETT )

input: Matthew lendend a cake beside the chair to the doctor .  
 output: LEND ( MATTHEW , BESIDE ( CAKE , CHAIR ) , DOCTOR )

input: Matthew lendend Emma a cake .  
 output: LEND ( MATTHEW , CAKE , EMMA )

input: Harper offered the professor the pickle .  
 output: OFFER ( HARPER , PICKLE , PROFESSOR )

input: Matthew mailed Emma a cake .  
 output: MAIL ( MATTHEW , CAKE , EMMA )

input: Matthew handed Emma a strawberry in the house .  
 output: HAND ( MATTHEW , IN ( STRAWBERRY , HOUSE ) , EMMA )

input: A dog fed the professor the cake .  
 output: FEED ( DOG , CAKE , PROFESSOR )

input: Matthew shipped the professor a chair .

**label:**  
 SHIP ( MATTHEW , CHAIR , PROFESSOR )

**pred:**  
 SHIP ( MATTHEW , CHAIR , PROFESSOR )

Figure 14: Comparison of performances between excluding NL-side matching (left) and containing NL-side matching (right). For the test input “Matthew shipped the professor a chair .”, it contains the sentence structure “subject verb object\_1 object\_2” behind the NL expression. Context on the left does not explicitly contain this sentence structure, but it contains a semantically equivalent structure (i.e., “subject verb object\_2 to object\_1”). However, the model generates the correct prediction on the right while fails on the left. Concretely, according to the wrong prediction on the left, the model perhaps considers that the semantics of “subject verb object\_1 object\_2” is equivalent with “subject verb object\_1 to object\_2”.

Model: code-davinci-002  
Category: *PhraReco*  
Setting: high diversity

**context:**

input: A cake was forwarded to Levi by Charlotte .  
output: FORWARD ( CHARLOTTE , CAKE , LEVI )

input: A cake rolled .  
output: ROLL ( NONE , CAKE , NONE )

input: Emma rolled a teacher .  
output: ROLL ( EMMA , TEACHER , NONE )

input: A rose was helped by a dog .  
output: HELP ( DOG , ROSE , NONE )

input: The sailor dusted a boy .  
output: DUST ( SAILOR , BOY , NONE )

input: Evelyn rolled the girl .  
output: ROLL ( EVELYN , GIRL , NONE )

input: The girl needed to cook .  
output: NEED ( GIRL , NONE , NONE ) XCOMP COOK ( GIRL , NONE , NONE )

input: The captain ate .  
output: EAT ( CAPTAIN , NONE , NONE )

input: Emma broke a girl in the house .  
output: BREAK ( EMMA , IN ( GIRL , HOUSE ) , NONE )

input: The monster slept .  
output: SLEEP ( MONSTER , NONE , NONE )

input: A girl in the house slept .

**label:**  
SLEEP ( IN ( GIRL , HOUSE ) , NONE , NONE )

**pred:**  
SLEEP ( IN ( GIRL , HOUSE ) , NONE , NONE )

Model: code-davinci-002  
Category: *PhraReco*  
Setting: low diversity (full similarity setting)

**context:**

input: Emma broke a girl in the house .  
output: BREAK ( EMMA , IN ( GIRL , HOUSE ) , NONE )

input: Liam respected that Noah slept .  
output: RESPECT ( LIAM , NONE , NONE ) CCOMP SLEEP ( NOAH , NONE , NONE )

input: Emma liked that Jack slept .  
output: LIKE ( EMMA , NONE , NONE ) CCOMP SLEEP ( JACK , NONE , NONE )

input: Luke slept .  
output: SLEEP ( LUKE , NONE , NONE )

input: Elizabeth slept .  
output: SLEEP ( ELIZABETH , NONE , NONE )

input: Amelia slept .  
output: SLEEP ( AMELIA , NONE , NONE )

input: The fish slept .  
output: SLEEP ( FISH , NONE , NONE )

input: The monster slept .  
output: SLEEP ( MONSTER , NONE , NONE )

input: A girl tolerated that Amelia slept .  
output: TOLERATE ( GIRL , NONE , NONE ) CCOMP SLEEP ( AMELIA , NONE , NONE )

input: The girl wished that Ava slept .  
output: WISH ( GIRL , NONE , NONE ) CCOMP SLEEP ( AVA , NONE , NONE )

input: A girl in the house slept .

**label:**  
SLEEP ( IN ( GIRL , HOUSE ) , NONE , NONE )

**pred:**  
SLEEP ( GIRL , IN ( NONE , HOUSE ) , NONE )

Figure 15: Comparison of performances on *PhraReco* under high diversity (left) and low diversity (right). For the test input “A girl in the house slept”, “subject slept” is one element contained in  $T_S^{>1}$ . This element is repeatedly covered in the context on the right (low diversity) while only covered once on the left (high diversity). However, under high repetitiveness, the model fails on the test case, but succeed when there is low repetitiveness.

Model: code-davinci-002  
Category: *PhraReco*  
Setting: low complexity

**context:**

input: Liam appreciated a mouse on the gravel .  
output: APPRECIATE ( LIAM , ON ( MOUSE , GRAVEL ) , NONE )

input: A cat ate a basket beside the table .  
output: EAT ( CAT , BESIDE ( BASKET , TABLE ) , NONE )

input: Luna ate .  
output: EAT ( LUNA , NONE , NONE )

input: A dog ate .  
output: EAT ( DOG , NONE , NONE )

input: The prince ate .  
output: EAT ( PRINCE , NONE , NONE )

input: The coach ate .  
output: EAT ( COACH , NONE , NONE )

input: A monster ate .  
output: EAT ( MONSTER , NONE , NONE )

input: The priest ate .  
output: EAT ( PRIEST , NONE , NONE )

input: The captain ate .  
output: EAT ( CAPTAIN , NONE , NONE )

input: A mouse ate .  
output: EAT ( MOUSE , NONE , NONE )

input: A mouse beside the table ate .

**label:**  
SLEEP ( IN ( GIRL , HOUSE ) , NONE , NONE )

**pred:**  
SLEEP ( IN ( GIRL , HOUSE ) , NONE , NONE )

Model: code-davinci-002  
Category: *PhraReco*  
Setting: high complexity

**context:**

input: Emma liked a mouse on a table beside the machine .  
output: LIKE ( EMMA , ON ( MOUSE , BESIDE ( TABLE , MACHINE ) ) , NONE )

input: The boy ate the drink in a house beside the table .  
output: EAT ( BOY , IN ( DRINK , BESIDE ( HOUSE , TABLE ) ) , NONE )

input: A girl liked that Emma ate .  
output: LIKE ( GIRL , NONE , NONE ) CCOMP EAT ( EMMA , NONE , NONE )

input: Emma said that a dog ate .  
output: SAY ( EMMA , NONE , NONE ) CCOMP EAT ( DOG , NONE , NONE )

input: The father liked that the boy ate .  
output: LIKE ( FATHER , NONE , NONE ) CCOMP EAT ( BOY , NONE , NONE )

input: A princess proved that Emma ate .  
output: PROVE ( PRINCESS , NONE , NONE ) CCOMP EAT ( EMMA , NONE , NONE )

input: Emma liked that the teacher ate .  
output: LIKE ( EMMA , NONE , NONE ) CCOMP EAT ( TEACHER , NONE , NONE )

input: A horse said that Emma respected that William ate .  
output: SAY ( HORSE , NONE , NONE ) CCOMP RESPECT ( EMMA , NONE , NONE ) CCOMP EAT ( WILLIAM , NONE , NONE )

input: The spokesman hoped that a girl liked that a turtle ate .  
output: HOPE ( SPOKESMAN , NONE , NONE ) CCOMP LIKE ( GIRL , NONE , NONE ) CCOMP EAT ( TURTLE , NONE , NONE )

input: A mouse hoped that a girl hoped that Olivia ate .  
output: HOPE ( MOUSE , NONE , NONE ) CCOMP HOPE ( GIRL , NONE , NONE ) CCOMP EAT ( OLIVIA , NONE , NONE )

input: A mouse beside the table ate .

**label:**  
EAT ( BESIDE ( MOUSE , TABLE ) , NONE , NONE )

**pred:**  
EAT ( MOUSE , BESIDE ( NONE , TABLE ) , NONE )

Figure 16: Comparison of performance on *PhraReco* under low complexity (left) and high diversity (right). With low complexity, the test case is covered by simple and short in-context examples, and the model succeeds on the test case. With high complexity, the test case is covered by more complex and longer examples, and the model fails on the test case.

Table 7: Full results.

Model	Primitive	Similarity		Diversity		Complexity			PrimSubs	PrimAlte	PhraReco	LongChain	DeepNest	Average
		Rough	Precise	Low	High	Low	Mid	High						
code-davinci-002	✓								92.2	77.1	60.8	62.1	12.3	60.9
	✓	✓	✓	✓		✓			99.8	99.7	65.3	87.0	26.0	75.6
	✓	✓		✓		✓			97.7	92.1	77.6	80.4	18.3	73.2
	✓	✓	✓		✓	✓			-	-	80.0	87.6	26.2	64.6
	✓	✓	✓	✓	✓		✓		-	-	67.6	87.3	25.6	60.2
	✓	✓	✓	✓				✓	-	-	56.9	87.6	26.0	56.8
	Assembling Desired Factors									99.8	99.7	80.0	87.6	26.2
text-chat-davinci-002	✓								92.2	75.4	47.0	65.0	6.3	57.2
	✓	✓	✓	✓		✓			99.5	99.3	53.4	87.7	18.9	71.8
	✓	✓		✓		✓			96.1	89.7	62.9	80.1	11.7	68.1
	✓	✓	✓		✓	✓			-	-	69.2	87.6	18.2	58.3
	✓	✓	✓	✓	✓		✓		-	-	55.1	87.6	19.0	53.9
	✓	✓	✓	✓				✓	-	-	45.1	88.2	19.2	50.8
	Assembling Desired Factors									99.5	99.3	69.2	88.2	19.2
text-davinci-002	✓								88.5	66.4	38.7	46.5	2.9	48.6
	✓	✓	✓	✓		✓			99.7	99.4	39.4	80.2	12.7	66.3
	✓	✓		✓		✓			94.9	86.7	55.9	66.3	8.1	62.4
	✓	✓	✓		✓	✓			-	-	60.6	78.7	12.3	50.5
	✓	✓	✓	✓	✓	✓			-	-	43.2	79.9	12.9	45.3
	✓	✓	✓	✓				✓	-	-	33.5	80.2	12.8	42.2
	Assembling Desired Factors									99.7	99.4	60.6	80.2	12.9
code-cushman-002	✓								82.6	55.6	21.3	29.3	5.0	38.8
	✓	✓	✓	✓		✓			98.9	99.0	28.5	64.0	15.1	61.1
	✓	✓		✓		✓			94.0	77.7	31.4	44.7	10.3	51.6
	✓	✓	✓		✓	✓			-	-	40.8	62.4	14.9	39.4
	✓	✓	✓	✓	✓		✓		-	-	31.9	64.3	15.8	37.3
	✓	✓	✓	✓				✓	-	-	22.6	64.5	14.6	33.9
	Assembling Desired Factors									98.9	99.0	40.8	64.5	15.8
code-cushman-001	✓								76.6	60.7	16.9	5.0	1.0	32.0
	✓	✓	✓	✓		✓			99.1	98.4	20.7	11.1	8.9	47.6
	✓	✓		✓		✓			92.5	86.0	24.7	8.0	3.5	42.9
	✓	✓	✓		✓	✓			-	-	31.4	12.8	8.4	17.5
	✓	✓	✓	✓	✓		✓		-	-	23.2	12.7	8.9	14.9
	✓	✓	✓	✓				✓	-	-	18.6	11.5	8.7	12.9
	Assembling Desired Factors									99.1	98.4	31.4	12.8	8.9
code-cushman-001	✓								69.4	52.3	9.4	2.3	0.2	26.7
	✓	✓	✓	✓		✓			97.5	95.4	12.3	13.4	1.4	44.0
	✓	✓		✓		✓			79.4	66.6	18.8	4.3	1.3	34.1
	✓	✓	✓		✓	✓			-	-	20.0	10.2	1.3	10.5
	✓	✓	✓	✓	✓	✓			-	-	14.7	13.8	1.4	10.0
	✓	✓	✓	✓				✓	-	-	7.8	13.5	1.3	7.5
	Assembling Desired Factors									97.5	95.4	20.0	13.8	1.4
Fine-Tuned GPT2-Large									93.6	97.9	14.0	5.4	0.0	42.2



## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
*Limitations*
- A2. Did you discuss any potential risks of your work?  
*Ethics Statement*
- A3. Do the abstract and introduction summarize the paper’s main claims?  
*Abstract and Section 1*
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

*Section 3 & Section 4.1*

- B1. Did you cite the creators of artifacts you used?  
*Section 1*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
*Not applicable. Left blank.*
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
*Not applicable. Left blank.*
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
*Not applicable. Left blank.*
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
*Section 2.2*
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
*Appendix D.5*

### C Did you run computational experiments?

*Section 4 & Section 5*

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
*Section 2.3 & Limitations*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

*Section 2.3 & Section 4.1*

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

*Appendix 1*

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

*Not applicable. Left blank.*

**D  Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

*No response.*

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

*No response.*

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

*No response.*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*No response.*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

*No response.*