A Survey on Compositional Learning of AI Models: Theoretical and Experimental Practices

Sania Sinha *sinhasa3@msu.edu*
Department of Computer Science and Engineering
Michigan State University

Tanawan Premsri *premsrit@msu.edu*
Department of Computer Science and Engineering
Michigan State University

Parisa Kordjamshidi *kordjams@msu.edu*
Department of Computer Science and Engineering
Michigan State University

Reviewed on OpenReview: *https://openreview.net/forum?id=BXDxwItNqQ*

Abstract

Compositional learning, mastering the ability to combine basic concepts and construct more intricate ones, is crucial for human cognition, especially in human language comprehension and visual perception. This notion is tightly connected to generalization over unobserved situations. Despite its integral role in intelligence, there is a lack of systematic theoretical and experimental research methodologies, making it difficult to analyze the compositional learning abilities of computational models. In this paper, we survey the literature on compositional learning of AI models and the connections made to cognitive studies. We identify abstract concepts of compositionality in cognitive and linguistic studies and connect these to the computational challenges faced by language and vision models in compositional reasoning. We overview the formal definitions, tasks, evaluation benchmarks, various computational models, and theoretical findings. Our primary focus is on linguistic benchmarks and combining language and vision, though there is a large amount of research on compositional concept learning in the computer vision community alone. We cover modern studies on large language models to provide a deeper understanding of the cutting-edge compositional capabilities exhibited by state-of-the-art AI models and pinpoint important directions for future research.

1 Introduction

The compositional learning and reasoning of an intelligent agent refers to the ability to understand and manipulate complex structures by decomposing them into simpler parts and composing parts to form new complex concepts with a coherent understanding. This ability is a key factor in generalizing learning to unobserved situations (Hupkes et al., 2023). Compositional learning in intelligent systems is cognitively motivated since humans learn compositionally (Lake et al., 2019). Researchers have examined this phenomenon from cognitive, linguistic, and psychological perspectives (Shepard, 1987; Frankland & Greene, 2020).

The formal notion of compositionality originated from natural language and semantics, with various theories and arguments that elaborate on this concept. The principle of compositionality (Partee, 2004; Janssen & Partee, 1997; Montague, 1974) is defined as *"The meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined"* with three general methods- new meanings, new basic parts, and new constructions. One of the earliest formalizations of compositionality was grounded in grammar trees when cognitive scientists proposed an *information processing* approach to create a model

of the mind (Thagard, 2023). The birth of modern cognitive science happened following the proposal of phrase structure and transformational grammar (Chomsky, 1965). Compositional understanding of linguistic constructs has multiple aspects (Marelli et al., 2014; Li et al., 2021). For example, a nesting description such as "The black tall woman on the left of the car" conveys the intersection of multiple adjectives and spatial relations. Thus, the composition is defined as the intersection of multiple concepts. However, there are cases in which the direct intersection is not applicable, and the meaning should be inferred from the concepts in the global context, such as recognizing the sentiment of the following sentence "The pizza is so good, I hate this place!" Despite the natural language being a prominent manifestation of compositionality, this can be expanded to other areas of human intelligence, such as vision (Saffran et al., 2007). The same notion of intersection, as well as part-whole compositions, is essential for visual intelligence.

Compositional learning is important in complex tasks where high-level goals must be decomposed into smaller subgoals and plans, for example, when instructing an agent to navigate from one point to another (Schmidhuber, 1990). From the computational modeling perspective, traditionally formal grammars have been the means to address the compositional understanding of various modalities primarily in language and extended to vision (Girshick et al., 2011; Hong et al., 2021). While symbolic models inherently address compositional structures, using them alone, that is, parsing raw and noisy data into a structure with manually designed grammars will be brittle in real-world situations. Our study focuses on data-driven approaches based on artificial neural networks and the combined paradigms of neuro-symbolic techniques. Several studies provide both experimental and theoretical analyses, indicating the competitiveness of the neural models in expressing compositional structures such as context-free grammars (Siegelmann & Sontag, 1995; Shi et al., 2021). Their expressive power added to the robustness in dealing with noisy data makes the neural techniques applicable to realistic situations.

In this paper, we examine multiple aspects of compositional learning, including compositional learning facets, datasets, computational models, and evaluation paradigms in both theory and practice. Figure 1 shows the scope and structure of our survey, covering four main topics, that is, Compositional Learning Facets, Datasets, Compositional Learning Models, and Evaluation. **For Compositional Learning Facets**, we overview the different measures of compositionality rooted in cognitive science that define abstract tasks for compositionality and connect them to the other adjacent topics such as continual learning and emergent intelligence. **For Datasets**, we cover existing benchmarks proposed in the AI community. They help bridge the gap between interdisciplinary theoretical definitions and the design of better evaluation benchmarks to pinpoint model capabilities. **For Compositional Learning Models**, we overview the architectures that aim to address compositional learning, divided into categories of basic neural architectures, large language models, and customized architectures, including neuro-symbolic models. These models are mostly evaluated empirically using conventional benchmarks, while fewer studies conduct theoretical analyses. We cover both types of evaluations of various models when available in the existing literature. **For Evaluation**, we overview two evaluation approaches- theoretical and empirical. The theoretical evaluations examine various computational models in a mathematical framework, investigating their expressivity, and capacity for compositional learning, and analyze the generalizations to unobserved situations by computing the error bounds. Empirical evaluations include experimental results on benchmarks set by datasets and tasks created primarily to highlight the core challenges of compositional learning for language and vision understanding. Such results often report performance measures on the tasks designed to test the cognitive aspects of compositionality.

Overall, the cognitive aspects lay the foundation of the concept of compositionality and define the different abstractions associated with it and the tasks that are designed accordingly. The empirical evaluations use these tasks to evaluate compositionality using experimental performance. However, some studies examine the mathematical analysis and functional form of the models independent from the datasets. Finally, both these evaluations are used to develop models that are capable of compositional learning.

2 Compositional Learning Facets

Compositionality is one important aspect of generalization as a whole (Hupkes et al., 2023). Cognitive Science and Linguistics literature have identified broad categorizations of tasks that define compositionality and can be used to evaluate the compositional reasoning of models. The foundations of human natural language lie

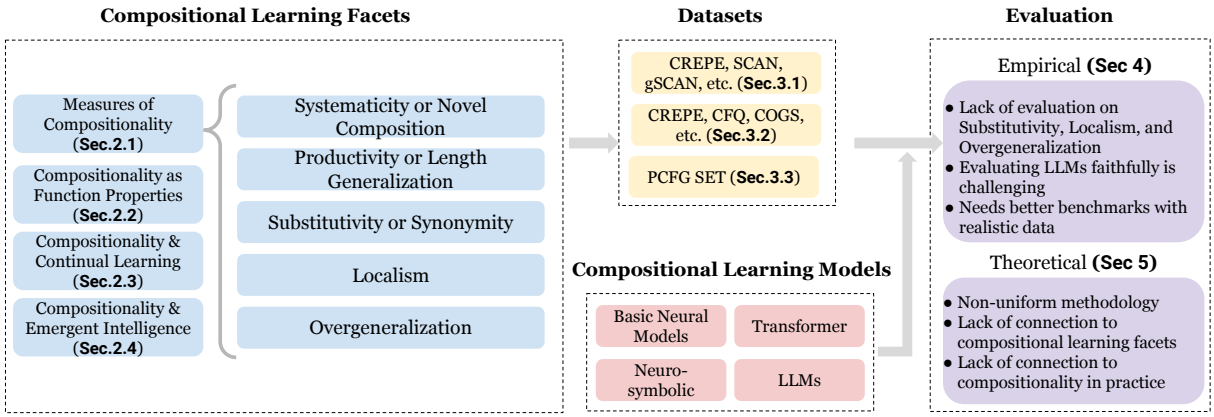

Figure 1: Outline of covered concepts in this survey, related to the structure of the paper. We structure our study of compositional learning by dividing it into four parts of compositional learning facets, datasets, compositional learning models, and evaluation methods from both empirical and theoretical perspectives. The topics have the respective sections associated with them. The main areas of required research and future direction are included in the descriptions in the Evaluation boxes, which are further discussed in Section 6.

in compositionality. A commonly used task categorization, derived from reformulated theoretically grounded tests from Hupkes et al. (2020), defines five main metrics of compositionality: systematicity, productivity, substitutivity, localism, and overgeneralization (Dankers et al., 2021).

2.1 Measures of Compositionality

Hupkes et al. (2020) introduces theoretically grounded tests for compositionality of models based on different interpretations of compositionality (Fodor & Pylyshyn, 1988; Chomsky, 1956). Some of these tasks existed in earlier research under different terms, such as productivity, which builds on a substantial body of prior research on length generalization. These tasks are becoming widely accepted tasks for compositional learning, and there are current datasets that use them for their evaluation splits. We describe these measures in detail below.

Systematicity or Novel Composition

Systematicity is one of the most commonly used notions of compositionality in evaluating the performance of computational models. It has been defined as the ability to systematically recombine known parts and rules (Dankers et al., 2021). It derives directly from the commonly accepted definition of composition, which is the formation of compound expressions as a function of simpler ones (Partee, 2004). Systematicity is a standard concept in cognitive science research on building cognitive architectures that try to formalize the human thinking process (Fodor & Pylyshyn, 1988). The ability to syntactically combine known elements to form new or "unseen" expressions is an integral test when evaluating a model's ability to reason compositionally. It is also called Distribution-Based Compositionality Assessment (Keysers et al., 2020), where two principles are defined: one is to make sure the distribution of atoms is similar in both training and test sets, while another is to ensure the distribution of compounds is different. For example, if "red" and "car" are two separately learned concepts, the model should be able to accurately utilize the unseen concept of "red car."

Productivity or Length Generalization

Another commonly explored test for compositionality is length generalization or productivity, as defined in Hupkes et al. (2020). In this evaluation, models are tested on their performance with expressions or sequences that are longer than training data. Longer input sequences may be recursive or nested versions

of seen phrases in the case of natural language inputs (Kim & Linzen, 2020). For example, if the model has seen "A and B," it should be able to understand "A and B and C."

Substitutivity or Synonymity

Substitutivity is another form of evaluation defined in Dankers et al. (2021), which concerns the model performance on the introduction of synonyms in expressions. For example, testing a model on the translation of the same sentence, when switching between synonymous words such as donut to doughnut or aubergine to eggplant the translation would not change. This is one of the less explored axioms of compositionality.

Localism

Another nuance of compositionality is the notion of global versus local composition. According to the principle of compositionality, the locality of the composition operator can vary. The meaning of a complex expression can depend solely on the meaning of its immediate parts (local composition) or the global structure of the context. Localism can be tested by analyzing the meaning a model assigns to a standalone compound versus when that compound is part of a larger expression. For example, sentences X and Y with the same truth value might change when we add a context such as "Peter thinks X" and "Peter thinks Y" this is still not clear, needs to be rephrased: For example, sentences X and Y have the same truth values but if a same context is added to them their composition with the new context might lead to different truth values. For instance, when we add the context "Peter thinks" and obtain "Peter thinks X" and "Peter thinks Y", these two new sentences can have different truth values.(Hupkes et al., 2020; Carnap, 1947). The local interpretation of compositionality says that these new phrases will have the same truth values which might not be the case anymore as Peter might be aware of X and not Y. In other words, considering the phrase that X and Y are a part of, changes the meaning. This is another one of the less explored axioms of compositionality.

Overgeneralization

Overgeneralization, as defined in Dankers et al. (2021), evaluates how much a model prefers an exception versus a rule. The term is originally used in language acquisition literature, also known as overregularization. A well-known example of this is the past-tense debate (Marcus et al., 1992) in language, which is about the rule that English past-tense verbs can be formed by appending -ed to the stem of the verb in most cases while there are some exceptions. This property can be evaluated by testing a model on exceptions of a usual rule in the training data and seeing if the model has over-fitted the training samples (Hupkes et al., 2020). Another example of this task is translating idioms where the meaning of sentences is "exceptions" to usual rules. For example, when translating the idiom "it's raining cats and dogs", a literal translation does not make sense as the phrase is an exception and has a specific meaning different from the usual literal interpretation. In this scenario, a model can achieve better performance by considering sentences in a global sense -that is looking at the bigger picture, such as context from placement in a compound- instead of trying to evaluate the meaning locally, such as by isolating the phrase. This is yet another less evaluated axiom of compositionality.

2.2 Compositionality as Function Properties

In Ram et al. (2024), compositional functions are defined with several components with a *computation directed acyclic graph* (**cDAG**) at the core. Their formal definitions facilitate the evaluation of compositional properties of the learning models (i.e. compositional functions). They relate their defined structure to the learning models' expressivity and sample complexity. In this system, **Systematicity** can be thought of as the expressivity of a compositional function as a low entropy program (for example, decision tree) versus a high entropy program (for example, transformer). **Productivity**, in simple terms, can be interpreted as whether a compositional function is recursive. **Substitutivity** tests whether a compositional function respects important abstractions and can be factored over them. **Localism** measures the stability of a compositional function against local changes, where the structure of the function's elements affects the

importance of the level of locality respected. **Overgeneralization** is the extent of compression of a function, where a function might have a general rule but have exceptions to those in special cases.

2.3 Compositionality and Continual Learning

Compositionality is an important aspect of continual learning, also known as lifelong learning (Mendez & Eaton, 2021). Continual learning (Wang et al., 2024) is the concept of learning new tasks while retaining knowledge from previously learned tasks. In continual learning, compositionality is particularly crucial to prevent catastrophic forgetting, where earlier tasks are forgotten over time due to learning of new tasks (Liao et al., 2023). Since knowledge about novel tasks is compositional, the existing information can be combined in novel ways and used for future tasks. This enables forward transfer of knowledge rather than catastrophic forgetting (Mendez & Eaton, 2023). Both compositional and continual learning share the key challenge of finding reusable knowledge and further connecting those for transfer learning and dealing with complex unobserved situations.

2.4 Compositionality and Emergent Intelligence

The term emergent behavior has been used across various science-related fields, rooted in "More Is Different" by Nobel Prize-winning physicist P.W. Anderson (Anderson, 1972). Its introduction to the language modeling community, specifically in the context of the large language models, begins with Wei et al. (2022). They defined emergent ability as the ability that appears only in large models and is not observed in any smaller models. The emergent abilities demonstrated in their study include performing unseen tasks by following instructions (Ouyang et al., 2024) and demonstrating multi-hop reasoning skills through Chain-of-Thought prompting (Wei et al., 2024). These abilities reflect the model's capacity for generalizing to unobserved situations, which can further extend to the model's compositional learning ability. Therefore, the compositional learning ability of the models can be associated with the emergent intelligence of language models. In the same paper of Wei et al. (2022), creating new compositional learning benchmarks is proposed as one direction for evaluating and understanding LLMs' emergent abilities.

3 Abstract Tasks and Datasets

In this section, we categorize the existing datasets proposed for the evaluation of compositional learning. Our categorization is based on the type of compositionality facet explained in Section 2. Table 1 points to a list of important datasets we have surveyed. In general, there are more common evaluation benchmarks for Systematicity and Productivity. Systematicity focuses on the novel composition of seen atomic concepts and there are several benchmarks established for its evaluation, although some of those works do not explicitly use the term systematicity. The productivity measure was often referred to as length generalization before the term became commonly used. However, despite the abundance of datasets for these tasks, most rely on synthetic data, which poses a risk to their reliability in capturing the variation and complexity of real-world problems. While some datasets in the computer vision community for compositional learning utilize realistic images, they address fewer aspects of compositionality (e.g. object-attribute combination) compared to synthesized linguistic corpora designed for the same purpose. We describe some of the existing datasets and tasks in this section.

3.1 Systematicity or Novel Composition

CREPE. This is a Compositional REPresentation Evaluation benchmark (CREPE) (Ma et al., 2022). The dataset is synthesized and includes multiple splits, one of which relates to Systematicity. The main task setting is that, given an image, the model needs to identify an appropriate text caption describing it among multiple given choices. This systematicity challenge tests whether the model can systematically generate new combinations of seen atomic concepts during training. For example, "Crepe on a skillet" is never observed in the training while Crepe and skillet are observed separately in different contexts.

SCAN. The task is to navigate in a two-dimensional grid world based on natural language instructions. It is the *Simplified* version of the *CommAI Navigation* tasks (SCAN) (Lake & Baroni, 2017; Mikolov et al.,

2018). One of the proposed experiments in SCAN evaluates the model's compositional generalization across primitive commands. Specific compounds are excluded from training where the model has seen the primitives and similar compound structures. Then, these unseen compounds are tested during testing.

gSCAN. The task is to navigate in a two-dimensional grid world based on natural language instructions (Ruis et al., 2020), which is a grounded version of SCAN (gSCAN). It includes 9 splits A-I. Categories B to H present tasks that form tests for systematicity (B,C- Novel Composition of Object Properties, D- Novel Direction, E- Novel Contextual References, F- Novel composition of actions and arguments, G,H- Novel Adverbs). Each split focuses on some form of novel composition of known concepts.

PCFG SET. PCFG SET stands for Probabilistic Context Free Grammar String Edit Task. It is an artificial translation task where sequences produced by probabilistic context-free grammar need to be translated into sequences representing their meaning (Hupkes et al., 2020). The output sequences can be constructed recursively using specified string edit operations applied to the input sequence, e.g., the input "repeat ABC" will be mapped to the sequence "ABCABC." The systematicity test uses a combination of words a and b in the input where the model has seen a but not b and vice versa. However, the combination of a and b is plausible in the corpus.

COGS. This dataset is designed for the Compositional Generalization Challenge (COGS) (Kim & Linzen, 2020). The task is a kind of semantic parsing based on a fragment of English where the models need to determine a formal meaning representation of the input English text. There are 4 categories in COGS that test some form of systematicity, including Novel Combination of Familiar Primitives and Grammatical Roles, Novel Combination of Modified Phrases and Grammatical Roles, Verb Argument Structure Alteration, and Verb Class.

ReaSCAN. This extension of gSCAN overcomes some of its limitations by requiring compositional language interpretation and reasoning about entities and relations (Wu et al., 2021). The challenges for systematicity are Category A, which tests novel object attribute combinations such as novel color modifier, color attribute, and size attribute, which is adapted from gSCAN (Ruis et al., 2020). Category B tests unseen co-occurrences of objects and relations, which is unique to ReaSCAN.

SQOOP. This is a dataset of Spatial Queries On Object Pairs (SQOOP). It is a minimalistic visual question answering, with yes-no answers. The input is an image and a question based on spatial reasoning (Bahdanau et al., 2019). Models are tested for answering questions on all possible object pairs after being trained on only a subset. Questions are of the form X R Y (X and Y are objects while R is a relation) where training sets are generated by controlled sampling of X and Y objects.

CLUTRR. This dataset is on Compositional Language Understanding and Text-based Relational Reasoning (CLUTRR). The task is, given a natural language short story, to answer questions on kinship relations that can be inferred from story (Sinha et al., 2019). Models are tested on combinations of held-out reasoning rules that are unseen during training. Thus, it tests systematic generalization capability or systematicity.

KiloGram. The task is a reference game task where given a textual description, the model has to select the appropriate image from a set of images. These images are tangrams, and the dataset has rich annotations of these images in two splits, FULL and DENSE, which have varying numbers of annotations, hence the name Kilo Tangram KiloGram) (Ji et al., 2022). There are different variations of this task such as showing parts of the image versus the whole image and making it grayscale versus colored. This dataset is an example of compositionality in vision since the whole tangram image is made out of parts, and the model learns how different parts combine to form different images.

CompMCTG. To evaluate the compositional learning of generative language models, Compositional multi-aspect controllable text generation (CompMCTG) benchmark(Zhong et al., 2024) is proposed. The task is to generate sentences given a set of concepts, including aspects of sentiment, topic, tense, person, and stuff. The benchmark tests systematicity by evaluating the model's capability to generate sentences with novel or unseen combinations of such concepts. For example, if the model has seen sentences with concepts of (red, car) and (blue, hat), it should be able to generate sentences for (blue, car). To perform these evaluations, they split the dataset into in-distribution (I.D.) and compositional, which have no intersection, although recombining elements in the I.D. set can form elements in the compositional set. Their evaluation protocol

includes three test splits of Hold-Out, Attribute Compound Divergence (ACD), and Few-Shot evaluation. It is based on four popular textual corpora- Amazon Reviews (He & McAuley, 2016), a combination of IMDB, OpeNER, SenTube: Mixture (Liu et al., 2022a), YELP (Yelp, 2014), and Fyelp (Lample et al., 2019).

MIT-States. In this evaluation benchmark (Isola et al., 2015), the general task is to explain a collection of images in terms of the novel composition of primitive states and transformations applied to objects. It includes three different tasks: discovering relevant transformations (such as slicing, wilting), parsing states (such as sliced, wilted), and finding smooth transitions. It tests systematicity in image concepts such as being able to identify "sliced tomato" versus "whole tomato" and generalize it to unobserved compositions such as "sliced apple" versus "whole apple." There exists other similar works (Yu & Grauman, 2014; 2017; Tokmakov et al., 2019; Xu et al., 2023b; Bao et al., 2024) in the vision community that evaluate such compositional concept learning.

Skill-Mix. There is no commonly used benchmark for the evaluation of compositional learning abilities of LLMs. A very recent benchmark, called Skill-Mix (Yu et al., 2024) covers some aspects of compositional evaluation for generative models. The task expects the models to generate text by combining various skills and imposing some constraints on the generated text. While the notion of compositionality is not highlighted in the paper, we categorize their tasks under systematicity.

3.2 Productivity or Length Generalization

CREPE. The productivity split of CREPE (Ma et al., 2022) evaluates if a model can perform the trained task of longer sets of expressions. In this task, there are variations of complexity for n-subgraphs with $n \in \{4, 5, ...12\}$ and variations in the type of hard negatives used in the generation of text options. There are three types of hard negatives used- atomic hard negatives, swap hard negatives, and negation hard negatives.

PCFG SET. For the Productivity test (Hupkes et al., 2020), the data is split based on sequence lengths. The model is tested on sequences longer than the ones observed during training. For example, in a grammar context, if the model has been trained on the syntax entity-relation-entity, it will be tested on a longer, nested version of this concept, entity-relation-(entity-relation-entity).

CFQ. This is a dataset of Compositional Freebase Questions (CFQ). It is a natural language question-answering task (Keysers et al., 2020), focusing on semantic parsing, with a SPARQL query against the Freebase knowledge base. Questions are generated at varying levels of complexity. Various splits are available based on input length, output length, input pattern, or output pattern. All these splits aim to maximize compound divergence while minimizing atom divergence. This task appears to test both systematicity and productivity to some extent, although not explicitly. However, it cannot identify specific areas where a model's compositional behavior may be deficient.

COGS. While some splits of this dataset are explained in Section 3.1, it also contains a split for length generalization. Category 3 (i.e. Deeper Recursion), one of the splits of this dataset, (Kim & Linzen, 2020) is a test for length generalization by increasing the length of the input sequence recursively during testing. Input sequences are generated using nesting of phrases that are longer than those seen during training.

3.3 Other Generalization Criteria

To the best of our knowledge, PCFG SET (Hupkes et al., 2020) is the only benchmark that evaluates the other three additional criteria. The **Substitutivity** or synonymity test uses an input sequence with an atomic unit replaced by a synonymous atomic unit to evaluate how the model prediction changes. **Localism** is tested by using input sequences composed of smaller sequences A and B. The model is used to translate the full sequences first and then forced to process A and B separately. The outputs of these two experiments are compared to evaluate how local versus global the model is in its compositional reasoning. **Overgeneralization** test evaluates the model's results on input sequences that do not conform to the general rules of the dataset, that is, input sequences that are exceptions to the dataset rules. For example, the acquisition of past-tense forms such as the common "ed" ending (open-opened) versus more uncommon forms such as break to broke.

Name	Text	MM.	Sys.	Prod.	Subst.	Loc.	Overgen.	References
PCFG SET	✓	✗	✓	✓	✓	✓	✓	Csordás et al. (2021)
CLUTRR	✓	✗	✓	✗	✗	✗	✗	Gontier et al. (2020), Minervini et al. (2020) Sinha et al. (2019)
SQOOP	✓	✓	✓	✗	✗	✗	✗	D'Amario et al. (2021), Bahdanau et al. (2019)
CFQ	✓	✗	✗	✓	✗	✗	✗	Furrer et al. (2021), Herzig et al. (2021), Liu et al. (2021), Cao et al. (2022)
SCAN	✓	✗	✓	✓	✗	✗	✗	Korrel et al. (2019), Nye et al. (2020), Dessì & Baroni (2019)
COGS	✓	✗	✓	✓	✗	✗	✗	Haurilet et al. (2019), Wu et al. (2023), Klinger et al. (2024)
gSCAN	✓	✓	✓	✓	✗	✗	✗	Gao et al. (2020), Spilsbury et al. (2024)
ReaSCAN	✓	✓	✓	✓	✗	✗	✗	Kamali & Kordjamshidi (2023)
CREPE	✓	✓	✓	✓	✗	✗	✗	Lin et al. (2024), Singh et al. (2023)
KiloGram	✓	✓	✓	✗	✗	✗	✗	Kojima et al. (2023), Gui et al. (2023) Ji et al. (2022)
CLEVR	✓	✓	✓	✗	✗	✗	✗	Bahdanau et al. (2020), Johnson et al. (2016), Niemeyer & Geiger (2021)
MIT-States	✗	✓	✓	✗	✗	✗	✗	Xu et al. (2023a), Naeem et al. (2021)

Table 1: A summary of datasets and the compositional aspects they address with references of relevant papers on compositionality using these datasets. MM: multi-modal, sys: systematicity, prod: productivity, subst: substitutivity, loc: localism, overgen: overgeneralization

4 Empirical Findings: Compositional Learning Models

Traditional symbolic AI models naturally support compositional reasoning using classical logic applied to formal language (Szabó, 2022), formal verification (Giannakopoulou et al., 2018), and more. First-order logic can express objects, their properties, and complex compositional relations. Logical operations like conjunction, disjunction, and implication can express compositional structures on which inference rules are applied, supporting complex compositional reasoning (Porto, 2002). Another classical symbolic formalism includes grammars (Chomsky, 1965), which can express and generate complex compositional structures. However, dealing with noisy and uncertain data is hard with pure symbolic AI. Moreover, probabilistic augmentations and structured output prediction models have been able to explicitly model structural dependencies and support compositional reasoning based on their learned complex patterns from the data (Pearl, 1988). Nevertheless, scalability becomes a challenge for training and inference as the structural dependencies and the number of correlated variables increase. Given these long-lasting challenges of traditional models of compositionality, current neural models have shown success in both scalability and dealing with noisy and sensory data (OpenAI, 2024). Especially in modern large language models, complex compositional patterns can be memorized and resemble compositional reasoning. In the rest of this section, we overview the research focused on the development, design, and empirical evaluation of different types of neural models for compositional reasoning. We relate the type of empirical evaluations to the cognitive aspects of compositionality that they are testing. While some of these models utilize tasks and datasets covered in Section 3, others have their own tasks, specific to the problem of their choice.

4.1 Basic Neural Models

In Hupkes et al. (2020), different neural models are tested on a set of compositional learning tasks. They evaluate Long short-term memory (LSTM) Networks (Hochreiter & Schmidhuber, 1997), Convolutional Neural Networks (CNN) (LeCun et al., 1989), and Transformers for sequence-to-sequence language processing tasks on their proposed PCFG SET tasks including systematicity, productivity, substitutivity, localism and overgeneralization. On average, Transformer outperformed the other two models, but within the two classic neural models, the convolutional model performs better than the LSTM counterpart. In the reported results, two specific architectures, called LSTMS2S and ConvS2S were used. LSTMS2S is a recurrent, bidirectional encoder-decoder model with attention where the encoder and decoder are LSTMs, from the OpenMT-py framework. ConvS2S is a convolution-based sequence-to-sequence model as used in Gehring et al. (2017).

Several other works (Hupkes et al., 2018; Zheng & Lapata, 2021; 2022; Lake & Baroni, 2017) have used similar models to conclude that neural sequence models can exploit recursive compositional structure (Bowman et al., 2015; Irsoy & Cardie, 2014) in solving tasks. The related work in compositionality in computer vision indicates similar results. In Klinger et al. (2020), MLP, CNN, ResNet (He et al., 2015), and relational networks such as WReN (Barrett et al., 2018) and PrediNet (Shanahan et al., 2020) are evaluated on PCFG SET substitutivity and productivity tests. Their results indicate that compositional reasoning is challenging for the evaluated models and calls for more sophisticated architectures.

4.2 Transformer-based Architectures

The compositional capability of large language models is currently a controversial topic. They have been evaluated on general tasks such as arithmetic, logic, and dynamic programming that are compositional by nature. Some of these evaluation efforts conducted in Dziri et al. (2023) concluded that GPT (OpenAI, 2024) family models, solve these tasks by reducing them to linearized subgraph matching, without developing true compositional reasoning skills. Moreover, it is shown that, asymptotically, they have architectural limitations in solving highly complex compositional tasks with novel patterns due to error propagation of the composition of erroneous building block functions. There is a substantial gap in the performance of Transformers on in-domain and low-complexity compositional examples versus out-of-domain instances. The tested tasks were 1) multi-digit multiplication (Hiebert, 2013), 2) Einstein's puzzle, which is a constraint satisfaction problem (Prosser, 1993), and 3) NP-complete maximum weighted independent set problem (Kleinberg & Tardos, 2005). These tasks are mostly aimed at testing systematicity and productivity. Their results indicate that transformers make predictions on shallow reasoning and memorization of similar subgraph patterns seen during training as opposed to reasoning holistically based on true compositional reasoning. This hypothesis also aligns with the findings presented in Chang & Bisk (2024). They conduct experiments with counting problems, a basic form of length generalization tasks. While transformers can count in observed cases, they dramatically fail to perform out-of-domain for the same task, indicating that transformers rely on memorizing observations. The results on transformers often extend to transformer-based language models. As shown in this work (Anil et al., 2022), fine-tuned transformer-based large language models lack generalization capabilities irrespective of the model size, when tested on other length generalization tasks such as parity and boolean variable assignment. This is due to the transformer's tendency to learn non-sequential patterns that do not apply to longer sequences. However, combining a pretrained LLM's in-context learning ability with scratchpad prompting significantly improves performance on longer sequences. This also implies that despite having access to an infinite data pool, LLMs can potentially learn some tasks better through in-context learning than finetuning, confirmed by theoretical work on LLMs explained in Section 5.

While a series of research papers focused on evaluating the compositional generalization of Transformers (Dehghani et al., 2019; Hahn, 2020; Feng et al., 2023), some recent research investigated specific architectural factors that can impact the performance of Transformers on compositional tasks, following the claim that Transformers cannot reason compositionally (Dziri et al., 2023). In Ontanon et al. (2022), five configurations were evaluated on several different datasets and benchmarks, by varying five properties of Transformers including, 1) type of positional encoding, 2) use of copy decoders, 3) model size, 4) weight sharing, and 5) use of intermediate representations for prediction. The employed tasks were Addition, AdditionNegatives, Reversing, Duplication, Cartesian, Intersection, SCAN-length and SCAN-add-jump, PCFG productivity and systematicity, COGS, and CFQ-mcd1. This work concluded that relative positional encodings usually help, but using embeddings is necessary, and merely relative position biases are not sufficient. Tasks like SCAN and CFQ were not affected by positional embeddings. Tasks like Duplication or PCFG benefit from a copy decoder because it can learn a type of symmetry like learning a certain position of the input. As for model size, it was found that for algorithmic tasks, large models did not make a difference. However, for PCFG, large models seemed to outperform their smaller variants. Weight sharing across transformer layers seems to improve accuracy in most tasks. Intermediate representations also improved performance by creating new levels of abstraction that make reasoning easier for solving the end task. Specifically, using intermediate representations achieved state-of-the-art performance on COGS by converting the task from seq2seq to sequence tagging. Using intermediate representation on CFQ, eliminating the need to perform Cartesian products by using triple representations, also showed a significant performance improvement. However, intermediate representations need to be crafted specifically for a task and were only tested on COGS and CFQ. Other

works (Ruoss et al., 2023; Kazemnejad et al., 2023; Chang & Bisk, 2024) also investigate the effect of different positional encodings on the performance of Transformers for specifically length generalization. Kazemnejad et al. (2023) concludes that commonly used positional encodings such as ALiBi and APE are not well suited for downstream tasks but transformers without any positional encoding (NoPE) outperform other explicit positional encoding methods when testing decoder-only transformers on tasks such as SCAN and copying and reversing. Despite this, Chang & Bisk (2024) presents compelling evidence suggesting that each type of positional encoding demonstrates robustness on unique length generalization tasks. Consequently, the performance of Transformers on compositional tasks varies depending on the selected positional encoding. The authors even suggest integrating multiple positional encodings into Transformers for better compositionality by combining different strengths of those encodings.

Another type of research deviates from evaluating current conventional architectures and instead focuses on designing a new architecture that can compositionally generalize better. One of such models, a multi-modal transformer called GroCoT (Grounded Compositional Transformer) (Sikarwar et al., 2022), was designed and achieved state-of-the-art results on GSRR and ReaSCAN. Another multi-modal transformer from Qiu et al. (2021) is used as a backbone model with changes to Encoder, Decoder, modified spatial representation, interleaving self-attention, and a modified world state encoding. This work also showed that a single-layer transformer with a single head can ground and compose novel combinations of visual object attributes. They tested the generalizability on RefEx, that is on grounding referring expressions and, their proposed evaluation benchmark based on the target identification subtask of ReaSCAN. Another customized architecture example is adding Pushdown Layers to transformer architecture (Murty et al., 2023) which was recently presented as a replacement for standard self-attention. The recursive structure of natural language is challenging for self-attention layers to capture due to the lack of an explicit recursive-state tracking mechanism, which Pushdown Layers try to overcome. Pushdown layers have a stack tape that helps them model the recursive state of language. This helps Transformer-based language models softly modulate attention over tokens when predicting new ones. Other architectures derive inspiration from cognitive sciences. In the same line of work, RegularGPT (Chi et al., 2023) takes inspiration from working memory. It modifies the Transformer architecture to use weight sharing, adaptive depth, and sliding-dilated attention for better length generalization. When tested on the task of natural language extrapolation, it was found that it captures the local windowed attention patterns, which previous work identified as essential for the task. Additionally, it can efficiently model regular languages such as PARITY.

4.3 Neuro-Symbolic Architectures

A rising trend in cutting-edge research on modeling intelligent systems is neurosymbolic modeling. As the need for general-purpose AI models grows, there is a need for highly compositional models that can reason based on previously trained simpler tasks to do novel and complex ones. Although not explicitly mentioned in this research, they mainly address systematicity and productivity. One approach in this direction is to use natural language explanations to generate formal specifications that explicitly lay out a compositional task in terms of required simpler steps. The formal specifications then are passed to appropriate *engine*s to solve the problem. A prominent vision understanding model that follows this approach is VisProg (Gupta & Kembhavi, 2023). VisProg is a modular neurosymbolic model that can solve various compositional visual reasoning tasks given natural language instruction relying merely on the in-context learning of large language models. It produces modular programs in Python to obtain the solution. This approach provides an interpretable reasoning for how the model derives the solution. These modular programs use built-in modules supported by VisProg such as off-the-shelf neural computer vision models, image preprocessing modules, or Python subroutines, and solve complex tasks without any task-specific training. Another example in this line of work is to generate a formal logical specification of the problem from natural language explanations and pass the logical form to a logical reasoner engine (Poesia et al., 2023). This work uses large language models such as GPT-3 or GPT-3.5 Turbo, for producing "guides" to solving complex compositional tasks by breaking those down into smaller steps based on a reasoning chain. Similar to this, many recent works focused on different prompting strategies that can be used to solve complex compositional tasks with a modular approach. Examples include Decomposed Prompting (Khot et al., 2023), which uses a modular approach to decompose a complex task into simpler sub-tasks via prompting and pass on these sub-tasks to LLMs that are capable of solving them. This method allows for the optimization of a prompt for a specific sub-task, which can

be further decomposed, or replaced with more effective prompts, trained models, or symbolic functions as necessary. A similar approach for mapping to probabilistic logical reasoning is proposed in Nafar et al. (2024). Neuro-symbolic modeling has also been used previously in generative models for concept learning (Hofer et al., 2021) such as in the context of auditory signals for learning evolved combinatorial structure in language.

Another branch of Neuro-symbolic modeling explores leveraging the ability of large language models to do reasoning. This includes casual reasoning with an introduced benchmark, CLadder (Jin* et al., 2023). Their approach is to provide step-by-step structured prompts as a form of a chain-of-thought strategy called CausalCoT. The chain of thought (COT) conveys the formal symbolic representation of the causal reasoning problem. The CoT prompting is influenced by natural language rationales or reasoning processes (Qiao et al., 2023), which is similar to the use of answer rationales in inducing arithmetic problems (Ling et al., 2017).

Compositional Neural Architectures. Over the years, several architectures and theories have been proposed to model compositionality in their design. This aligns with the idea that neural architectures are structurally compositional (Lepori et al., 2023), meaning they leverage subroutines to break down complex tasks. One of the earlier examples of this type of architecture includes neural modular networks (Andreas et al., 2016). Neural modular networks were designed to model the inherent compositionality that exists in linguistic structures. The conceptual modules are built in the neural architecture based on the problem specification. For example, in a visual scene understanding problem, we can place modules for detecting objects, their compositional properties, and relations, which are the main building blocks for abstract reasoning needed for complex scene understanding.

In a similar line of work (Kuo et al., 2020), a network architecture was built that is compositional in nature and makes it possible to interpret what each part of the network learns. It solves tasks in gSCAN, which has agent navigation tasks in a 2D environment following a natural language command. The neural architecture built in this work assembles a *command-specific* network from previously trained modules, modeling the compositional nature of the command (task). Later research showed that in Neural Module Networks, it is hard to make the designed modules faithful to expressing the concepts that they are designed for. This is despite the overall network achieving high accuracy for the target task (Subramanian et al., 2020).

Given the importance of compositional learning in lifelong and continual learning, neurosymbolic approaches have been a suitable framework to address continuality through combining program synthesis and neural modeling. One such example is HOUDINI (Valkov et al., 2018), which is a neurosymbolic framework for lifelong learning of tasks combining perception and procedural reasoning such as counting, summing, and shortest-path computation. They use program synthesis to search over networks described as typed functional programs for the given task, whose parameters are then tuned end-to-end based on stochastic gradient descent. Another example is the Logic-Enhanced Foundation Model (LEFT) (Hsu et al., 2023), a framework designed to learn grounding and reasoning across domains using a differentiable, domain-independent, first-order logic-based program executor. It addresses the lack of generalizability across domains seen in related work, such as VisProg, which we previously discussed, including limitations in generalizing concepts from 2D to 3D images.

Neurocompositional Computing. Neuro-symbolic modeling has been motivated by its connection to *neurocompositional computing*. The term "neurocompositional computing" was coined in Smolensky et al. (2022b). It defines a type of computing that underlies human cognition as argued in contemporary cognitive science theories in Smolensky & Legendre (2006) and incorporates principles of Continuity and Compositionality. The Continuity Principle states that the encoding and processing of information should be continuous, that is, represented by real numbers that vary continuously and can be changed by arbitrarily small amounts. The Compositionality Principle states that larger, more complex structures can be decoded on the basis of smaller, simpler, and familiar building blocks. According to the *Central Paradox of Cognition*, the human brain follows both a continuous neural computing structure and a discrete compositional-structure computer. Following this theory, neuro-symbolic models that are both continuous and discrete in architecture seem like the ideal approach to modeling compositional behavior in computational models (Smolensky et al., 2022a).

5 Theoretical Findings: Mathematical Formulations of Compositionality

Theoretical analysis is fundamental for deepening our understanding of the compositionality of learning models. It can reveal intriguing and previously uncovered information that experimental analysis may overlook. Many research works have proposed diverse approaches for investigating the compositionality of learning models. We highlight three different approaches, including a mathematical framework for defining compositionality (Ram et al., 2023), exploring the upper-bounds of expressivity that relate to compositionality (Merrill & Sabharwal, 2023), and analyzing error-bounds to demonstrate the model's limitations in solving compositional learning problems (Dziri et al., 2023). In the rest of this section, we provide a detailed overview of these cases and explain the theoretical results on compositional generalization of classical neural networks, transformers, and modern language models. We also relate the mentioned techniques to aspects of compositionality when applicable.

5.1 Classical Neural Network

Ram et al. (2023) provides a mathematical definition of compositionality for learning models and connects their expressively to computational complexity. They frame the existing well-known models, such as variations of RNN and CovNets, with the provided formal definition to explain properties related to their compositional generalization. Hewitt et al. (2020) further investigates the RNN's ability to generate natural language with a certain nesting depth. They claim that RNNs with optimal memory and $O(m \log k)$ hidden units can generate a natural language of well-nested brackets of k types and m bounded nesting depth. With the rise of LLMs, compositional generalization has recently become more critical. Due to their large-scale parameters and training data, LLMs perform empirically well on many tasks. However, the empirical performance measures are now less reliable, as the high performance on test data can not be interpreted as compositional generalization anymore. This issue is due to the nature of internet-scale training of LLMs and data contamination. Consequently, there is more urgency for theoretical studies to understand their limitations and measure their reliability in unobserved situations. However, Ahn et al. (2023) argued that studying the smaller models at the single neuron level potentially leads to a better understanding of the large/deep models' learning behavior, which is related to explaining the Systematicity of the model. They also establish a connection between the *Edge of Stability* identified by the learning rate of the gradient descent approach for non-convex optimization and the emergent abilities in learning. This result remains limited to the scope of a single neuron and has not yet been extended to large models.

5.2 Transformers

To define the limitations of LLMs, it is essential to investigate the limitations of transformers and their underlying architectural component. In this work (Merrill & Sabharwal, 2023), the authors assume a specific transformer type, suggesting that their arithmetic precision is logarithmic in the number of input tokens. Based on this assumption, they demonstrate that transformers cannot accurately solve linear equalities or check membership in an arbitrary context-free grammar with empty productions. The studies of transformer precision have been explored before in Dehghani et al. (2019). They claim that standard transformers have limited precision, implying that they cannot handle an infinite input length. This conclusion indicates the limitations in of the compositionally of the transformers in terms of the Productivity aspect. Another notable theoretical investigations focus on the activation functions to explain the limitation of the transformer (Dehghani et al., 2019; Hahn, 2020). Hahn (2020) analyzes both hard-attention and soft-attention transformers. For hard attention, they prove that the transformer ignores most of the input information diagnosed by the specific modifications applied to the input. According to their analysis, transformers with hard attention will be unable to solve problems that require processing the entire input, such as PARITY and logical formula problems. However, this conclusion contradicts older papers that state transformers are Turing complete (Pérez et al., 2021). They utilize the strong assumption that all input information is accessible using hard attention to prove Turing completeness. This leads to a different conclusion, stating that the transformer can compute and access the entire internal dense representation. Hahn (2020) also investigate the model's behavior with soft attention. They illustrate that Transformers struggle with solving long input by demonstrating the influence of input on output substantially drops as the input gets longer. This is a

similar conclusion, using a different approach analyzed in an older study of (Dehghani et al., 2019). Based on these analyses, they further confirm the lack of productivity aspect of transformer architecture caused by limited training data lengths.

Despite proving weaknesses of the transformer, Hahn (2020) claims that the transformer has the potential to solve small input tasks completely. Two recent works also support this claim. The first work provides proof utilizing computation graphs and a theoretical study of error propagation in transformers. They claim that the auto-regressive transformer's error reduces as the size of the input decreases (Dziri et al., 2023). Moreover, they show that transformers reduce problems into multi-step compositional problems to solve larger tasks, which is strongly related to the Novel Composition of the compositional aspects. The second work supports the mentioned claim based on the study of sub-sequential finite-state transducers (SFSTs) (Valvoda et al., 2022). They generate a set of random SFSTs following Montague's Compositionality theorem to discover the coverage limitation. This limitation is inversely related to the size of the dataset and significantly impacts the probability of a model's successful generalization.

5.3 Large Language Models

In addition to inconclusive theoretical studies on transformer limitations, there are controversial results on large language models. The most noteworthy study is on the emerging abilities and capabilities claimed to be unique in the large models. The emerging abilities relate to the generalization to new and complex tasks in LLMs. This kind of ability is also a feature of models' compositional learning ability, allowing them to perform in novel compositional situations (Yu et al., 2024). Multiple works have shown the existence of emergent abilities of LLMs (Wei et al., 2022). The recent work of Arora & Goyal (2023) provides a mathematical framework for identifying complex skills in language models. They use the LLM Scaling Rule to argue that emergent skills are the results of reducing excessive loss. This excessive loss enables the model to learn how to utilize and combine skills from downstream tasks during training. Their claims are based on the assumption that language inherently contains a random mix of complex skills. Although several experiments reveal these emerging capabilities, at least two papers disclaim their existence. The first group provides a theoretical proof based on a mathematical framework. They illustrate that the emerging ability appears due to the selected evaluation metrics that are nonlinear and discontinuous (Schaeffer et al., 2023). They show as an artifact of the evaluation metrics, even simple models such as CNNs can show emerging abilities. Therefore, they conclude that emerging abilities may not be a fundamental property of the large models. Moreover, Lu et al. (2024) provides an extensive empirical study with 1000 experiments on 22 tasks with different LLMs. However, given the inconsistency in some results and the unpredictability of emerging abilities, they do not find any strong evidence of how they emerge. They associate the performance with in-context learning techniques, memorization, and data contamination. However, a recent study presents a positive theoretical analysis of reasoning capabilities by investigating the chain of thought (CoT) (Wei et al., 2024) and draws a different conclusion. They argue that the log-precision transformer can perform fundamental operations such as multiplication and a look-up table. Consequently, it can solve linear equations and other reasoning problems if it stores all the input information. However, the architecture alone struggles with storing the entire input, as observed in Dehghani et al. (2019); Merrill & Sabharwal (2023). They show that the model addresses this limitation by repeatedly referring to the input by enabling CoT (Feng et al., 2023). Therefore, with the right number of CoT examples, LLMs can overcome the transformer's weakness in solving mathematical reasoning. This is aligned with previous empirical results of in-context learning discussed in Section 4.2.

6 Discussion and Future Direction

There has been a large amount of research on the compositional learning ability of humans from a cognitive perspective (Fodor & Pylyshyn, 1988; Ito et al., 2022). Researchers in linguists and formal languages have formalized the notion of compositionality since languages have inherent compositional structure (Chomsky, 1965; 2002). However, from the AI and machine learning perspective, ideas are borrowed from both cognitive and linguistics, and computational tasks and models are designed focusing on narrow aspects of compositionality (Hupkes et al., 2020). Our investigation of AI models indicates several challenges regarding the

	Model Type	Theoretical Analysis	Empirical
Basic Neural Models	RNN	Hewitt et al. 2020	Bowman et al. 2015
	CNN	✗	Hupkes et al. 2020
	LSTM	Siegelmann & Sontag 1995	
Transformer-based	(Customized) Transformers	Hahn 2020, Pérez et al. 2019, Dehghani et al. 2019	Ontanon et al. 2022
	LLM	Dziri et al. 2023	Schaeffer et al. 2023
Neuro-symbolic	Neural Modular Network	✗	Kuo et al. 2020
	Other Models	✗	Gupta & Kembhavi 2023

Table 2: Summary of computational models with compositional learning ability from the theoretical perspective and an example from the experimental perspective.

designed tasks, benchmarks, and theoretical frameworks that make the evaluation of computational models problematic.

Figure 1 shows the connections between the main topics identified and discussed in this survey. Among the five identified types of compositional learning facets, only systematicity and productivity are well-researched and have clear connections to evaluation benchmarks. Given that these are the five main metrics of compositionality, we should aim to expand our evaluation capabilities by developing tests for the other three as well. Empirical evaluations are comparatively more well-studied compared to theoretical analyses. Theoretical evaluations are either lacking or do not follow a consistent methodology. There is a lack of connection between the theoretical methods and the cognitive aspects, making these results hard to use to guide better architectural design. For Models, different types of architectures have been designed. However, evaluating LLMs and making fundamental design decisions for compositional generalization present new challenges. We outline some of these challenges in detail below.

Less Explored Facets of Compositionality. Only Systematicity and Productivity have been well-researched and have established connections to evaluation benchmarks. While the other three were introduced as fundamental types of compositionality, they have received less attention, as they appear to be less commonly occurring aspects of compositionality. However, in the era of LLMs and the emergent in-context learning, Substitutivity and Localism are potential bottlenecks in the performance of LLMs for attending the appropriate context for solving problems. Moreover, Overgeneralization can be associated with hallucination and generating unfounded and incorrect information by making up new unrealistic abstractions. While hallucination is a broader concept than overgeneralization, this compositional learning facet can highlight an important aspect to be addressed to prevent hallucination (Huang et al., 2023). Therefore, directing attention to the other three types of measures can help establish new formal evaluation benchmarks, contributing to the development of more robust systems and addressing the challenges of the LLMs.

Synthetic and Unrealistic Evaluations. One issue in current evaluations is that controlled and clean tests of compositonality are mostly synthesized (Wu et al., 2021; Ruis et al., 2020). This is evidenced by our examples in Section 3. Even in rare cases that claim to work with realistic data (Keysers et al., 2020), synthesized questions are used to query knowledge graphs. However, more recent studies on language models' evaluation of compositionality focus on more challenging problems such as multi-hop question answering (Press et al., 2023; Liu et al., 2022b; Okawa et al., 2023; Mirzaee et al., 2021) as well as complex puzzles with combinatorial search solutions or compositional mathematical reasoning (Dziri et al., 2023). Although the benchmarks are designed for evaluating specific tasks, the reliance on mostly synthesized data risks the effectiveness of generalization to real-world data, which is often more complex.

Misalignment of Performance and Compositional Learning (LLM Evaluation Challenge). The second challenge that mostly applies to LLMs is data contamination. Though the recent research compares language models to the specialized architectures and indicates their outperformance in compositional tasks (Furrer et al., 2021), this result does not necessarily mean these models have better generalizations in recognizing unobserved compositions (Press et al., 2023). A major issue with these evaluations on realistic data is the difficulty in disentangling the compositional reasoning from the data contamination (Sainz et al., 2023) and memorization. However, very little recent work has been done to address this issue (Yu et al.,

2024). Based on our investigation of theoretical studies in Section 5, we identify that the generalization abilities can be an artifact of observing more complex data in larger contexts as well as evaluation metrics as has been pointed out in Schaeffer et al. (2023) and is a pertinent issue that needs to be addressed in new evaluation benchmarks.

Inconsistent Theoretical Methodology. Given the difficulty in obtaining conclusive empirical studies, theoretical understandings become even more important, nowadays. However, the lack of a well-established and practically informative theoretical framework for investigating the limitations and capabilities of LLMs has been a challenge to a deep understanding of their generalizability. According to our studies on the theoretical explanation of transformers in Section 5, their compositionality is still under discussion. Some results illustrate that transformers possess compositional generalizability based on their ability to solve complex tasks based on smaller subtasks (Feng et al., 2023; Dziri et al., 2023). However, other results based on a different evaluation methodology suggest that the emergence of such abilities, including compositional learning, is associated with the user's choice of evaluation metrics (Schaeffer et al., 2023). Similar lines of study deny the emergence of intelligence and relate the new abilities to in-context learning methods, models memory, and linguistic knowledge (Lu et al., 2024). Many research results confirm the limitations in the generalizability of LLMs. For example, the building blocks of these models, i.e. transformers, still have severe limitations in comprehending large inputs (Hahn, 2020; Dehghani et al., 2019). Despite these controversial discussions, there are only relatively few studies on the theoretical analysis of transformer-based models. The methodological frameworks for examining the LLMs' capabilities are not standardized. Therefore, the compositional capabilities of the current SOTA models require more attention from the research community. This research direction will help towards more consistent and conclusive results on the limitations of the model's generalizability, specifically the compositional generalization.

Cognitive Motivation. The fundamental capabilities of current AI models have been debated and criticized by scientists in cognitive science and psychology (Bender & Koller, 2020; Marcus, 2018). Despite giant leaps of performance progress in modern AI, there are distinct differences between these machines and human intelligence. From our exploration of cognitive aspects of compositional learning in Sections 1 and 2, we observe that cognitive foundations do not have strong ties to model building yet. Evaluating different models reveals that they often rely simply on pattern recognition (Geirhos et al., 2020; Dziri et al., 2023), instead of a holistic understanding of a problem grounded in reality and situation, as was seen in Section 5. Understanding human intelligence from Cognitive Science literature suggests that we must move beyond current engineering trends to build causal models of the world that support knowledge and understanding. The key ingredients of such human-like rich and efficient learning are compositionality and learning-to-learn (Lake et al., 2017).

7 Limitations

Despite the comprehensive nature of the survey and our efforts to cover and connect most research relating to compositional learning, we would like to acknowledge some limitations. The scope of this survey covers a broad spectrum of topics and tries to capture both theoretical and experimental frameworks, but there might be some relevant papers that are missed. Compositional learning is an interdisciplinary topic across Computer Science, Linguistics, Cognitive Science, etc. Although we have included insights and connections from across these fields, our work has a more in-depth focus on Computer Science literature, especially Natural language processing. While we tried to provide the overall picture of the related research and build a coherent story, we might not capture the detailed nuances of each definition and application.

Acknowledgements

This project is partially supported by the Office of Naval Research (ONR) grant N00014-23-1-2417. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research. We thank Dr. Tim Klinger and the anonymous reviewers for their constructive feedback, which greatly helped us improve this manuscript.

References

Kwangjun Ahn, Sebastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=9cQ6kToLnJ.

P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972. doi: 10.1126/science.177.4047.393. URL https://www.science.org/doi/abs/10.1126/science.177.4047.393.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39–48, 2016. doi: 10.1109/CVPR.2016.12.

Cem Anil, Yuhuai Wu, Anders Johan Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Venkatesh Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=zSkYVeX7bC4.

Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models, 2023. URL https://arxiv.org/abs/2307.15936.

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: What is required and can it be learned? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkezXnA9YX.

Dzmitry Bahdanau, Harm de Vries, Timothy J. O'Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron Courville. Closure: Assessing systematic generalization of clevr models, 2020.

Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning, 2024. URL https://openreview.net/forum?id=LAEd3kHao9.

David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 511–520. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/barrett18a.html.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463.

Samuel R. Bowman, Christopher D. Manning, and Christopher Potts. Tree-structured composition in neural networks without tree-structured architectures. In *Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches - Volume 1583*, COCO'15, pp. 37–42, Aachen, DEU, 2015. CEUR-WS.org.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6101–6119, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.422. URL https://aclanthology.org/2022.acl-long.422.

Rudolf Carnap. *Meaning and necessity: A study in semantics and modal logic*. University of Chicago Press, 1947.

Yingshan Chang and Yonatan Bisk. Language models need inductive biases to count inductively, 2024. URL https://arxiv.org/abs/2405.20131.

Ta-Chung Chi, Ting-Han Fan, Alexander Rudnicky, and Peter Ramadge. Transformer working memory enables regular language reasoning and natural language length extrapolation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5972–5984, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.397. URL https://aclanthology.org/2023.findings-emnlp.397.

N. Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3): 113–124, 1956. doi: 10.1109/TIT.1956.1056813.

Noam Chomsky. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge, 1965. URL http://www.amazon.com/Aspects-Theory-Syntax-Noam-Chomsky/dp/0262530074.

Noam Chomsky. *Backmatter*, pp. 115–118. De Gruyter Mouton, Berlin, New York, 2002. ISBN 9783110218329. doi: doi:10.1515/9783110218329.bm. URL https://doi.org/10.1515/9783110218329.bm.

Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 619–634, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.49. URL https://aclanthology.org/2021.emnlp-main.49.

Vanessa D'Amario, Tomotake Sasaki, and Xavier Boix. How modular should neural module networks be for systematic generalization? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=dwY40cSK-dt.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. The paradox of the compositionality of natural language: a neural machine translation case study, 2021. URL https://arxiv.org/abs/2108.05885.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Roberto Dessì and Marco Baroni. CNNs found to jump around more skillfully than RNNs: Compositional generalization in seq2seq convolutional networks. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3919–3923, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1381. URL https://aclanthology.org/P19-1381.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 70293–70332. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/deb3c28192f979302c157cb653c15e90-Paper-Conference.pdf.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=qHrADgAdYu.

Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71, 1988. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(88)90031-5. URL https://www.sciencedirect.com/science/article/pii/0010027788900315.

Steven M. Frankland and Joshua D. Greene. Concepts and compositionality: In search of the brain's language of thought. *Annual Review of Psychology*, 71(Volume 71, 2020):273–303, 2020. ISSN 1545-2085. doi: https://doi.org/10.1146/annurev-psych-122216-011829. URL https://www.annualreviews.org/content/journals/10.1146/annurev-psych-122216-011829.

Daniel Furrer, Marc van Zee, Nathan Scales, and Nathanael Schärli. Compositional generalization in semantic parsing: Pre-training vs. specialized architectures, 2021.

Tong Gao, Qi Huang, and Raymond Mooney. Systematic generalization on gSCAN with language conditioned embedding. In Kam-Fai Wong, Kevin Knight, and Hua Wu (eds.), *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 491–503, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-main.49.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1243–1252. JMLR.org, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://dx.doi.org/10.1038/s42256-020-00257-z.

Dimitra Giannakopoulou, Kedar S. Namjoshi, and Corina S. Păsăreanu. *Compositional Reasoning*, pp. 345–383. Springer International Publishing, Cham, 2018. ISBN 978-3-319-10575-8. doi: 10.1007/978-3-319-10575-8_12. URL https://doi.org/10.1007/978-3-319-10575-8_12.

Ross Girshick, Pedro Felzenszwalb, and David McAllester. Object detection with grammar models. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/6faa8040da20ef399b63a72d0e4ab575-Paper.pdf.

Nicolas Gontier, Koustuv Sinha, Siva Reddy, and Chris Pal. Measuring systematic generalization in neural proof generation with transformers. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 22231–22242. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fc84ad56f9f547eb89c72b9bac209312-Paper.pdf.

Liangke Gui, Yingshan Chang, Qiuyuan Huang, Subhojit Som, Alexander G Hauptmann, Jianfeng Gao, and Yonatan Bisk. Training vision-language transformers from captions. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=xLnbSpozWS.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, June 2023.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, December 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00306. URL http://dx.doi.org/10.1162/tacl_a_00306.

Monica Haurilet, Alina Roitberg, and Rainer Stiefelhagen. It's not about the journey; it's about the destination: Following soft paths under question-guidance for visual reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1930–1939, 2019. doi: 10.1109/CVPR.2019.00203.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pp. 507–517, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883037. URL https://doi.org/10.1145/2872427.2883037.

Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, and Yuan Zhang. Unlocking compositional generalization in pre-trained models using intermediate representations, 2021.

John Hewitt, Michael Hahn, Surya Ganguli, Percy Liang, and Christopher D. Manning. RNNs can generate bounded hierarchical languages with optimal memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1978–2010, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.156. URL https://aclanthology.org/2020.emnlp-main.156.

J. Hiebert. *Conceptual and procedural knowledge: The case of mathematics*. Lawrence Erlbaum Associates, Inc., 08 2013. ISBN 9781136559761. doi: 10.4324/9780203063538.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Matthias Hofer, Tuan Anh Le, Roger Levy, and Josh Tenenbaum. Learning evolved combinatorial symbols with a neuro-symbolic generative model, 2021.

Yining Hong, Qing Li, Song-Chun Zhu, and Siyuan Huang. Vlgrammar: Grounded grammar induction of vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1665–1674, October 2021.

Joy Hsu, Jiayuan Mao, Joshua B. Tenenbaum, and Jiajun Wu. What's left? concept grounding with logic-enhanced foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sq4o3tjWaj.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. URL https://arxiv.org/abs/2311.05232.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *J. Artif. Int. Res.*, 61(1):907–926, January 2018. ISSN 1076-9757.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? (extended abstract). In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 5065–5069. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/708. URL https://doi.org/10.24963/ijcai.2020/708. Journal track.

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. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174, 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00729-y. URL https://doi.org/10.1038/s42256-023-00729-y.

Ozan Irsoy and Claire Cardie. Deep recursive neural networks for compositionality in language. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/2cfd4560539f887a5e420412b370b361-Paper.pdf.

Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015.

Takuya Ito, Tim Klinger, Doug Schultz, John Murray, Michael Cole, and Mattia Rigotti. Compositional generalization through abstract representations in human and artificial neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 32225–32239. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d0241a0fb1fc9be477bdfde5e0da276a-Paper-Conference.pdf.

Theo M.V. Janssen and Barbara H. Partee. Chapter 7 - compositionality. In Johan van Benthem and Alice ter Meulen (eds.), *Handbook of Logic and Language*, pp. 417–473. North-Holland, Amsterdam, 1997. ISBN 978-0-444-81714-3. doi: https://doi.org/10.1016/B978-044481714-3/50011-4. URL https://www.sciencedirect.com/science/article/pii/B9780444817143500114.

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 582–601, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.38. URL https://aclanthology.org/2022.emnlp-main.38.

Zhijing Jin*, Yuen Chen*, Felix Leeb*, Luigi Gresele*, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models. In *NeurIPS*, 2023. URL https://zhijing-jin.com/files/papers/CLadder_2023.pdf.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997, 2016. URL https://api.semanticscholar.org/CorpusID:15458100.

Danial Kamali and Parisa Kordjamshidi. Syntax-guided transformers: Elevating compositional generalization and grounding in multimodal environments. In Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Koustuv Sinha, Amirhossein Kazemnejad, Christos Christodoulopoulos, Ryan Cotterell, and Elia Bruni (eds.), *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pp. 130–142, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.genbench-1.10. URL https://aclanthology.org/2023.genbench-1.10.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Drrl2gcjzl.

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. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygcCnNKwr.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_nGgzQjzaRy.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.731. URL https://aclanthology.org/2020.emnlp-main.731.

Jon Kleinberg and Eva Tardos. *Algorithm Design*. Addison-Wesley Longman Publishing Co., Inc., USA, 2005. ISBN 0321295358.

Tim Klinger, Dhaval Adjodah, Vincent Marois, Josh Joseph, Matthew Riemer, Alex 'Sandy' Pentland, and Murray Campbell. A study of compositional generalization in neural models, 2020.

Tim Klinger, Luke Liu, Soham Dan, Maxwell Crouse, Parikshit Ram, and Alexander Gray. Compositional program generation for few-shot systematic generalization, 2024.

Noriyuki Kojima, Hadar Averbuch-Elor, and Yoav Artzi. A joint study of phrase grounding and task performance in vision and language models. *Trans. Mach. Learn. Res.*, 2024, 2023. URL https://api.semanticscholar.org/CorpusID:261556878.

Kris Korrel, Dieuwke Hupkes, Verna Dankers, and Elia Bruni. Transcoding compositionally: Using attention to find more generalizable solutions. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 1–11, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4801. URL https://aclanthology.org/W19-4801.

Yen-Ling Kuo, Boris Katz, and Andrei Barbu. Compositional networks enable systematic generalization for grounded language understanding, 2020. URL https://arxiv.org/abs/2008.02742.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:46761158.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017. doi: 10.1017/s0140525x16001837.

Brenden M. Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. In *Annual Meeting of the Cognitive Science Society*, 2019. URL https://api.semanticscholar.org/CorpusID:58006558.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text rewriting. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1g2NhC5KQ.

Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, R. Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky (ed.), *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989. URL https://proceedings.neurips.cc/paper_files/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf.

Michael A. Lepori, Thomas Serre, and Ellie Pavlick. Break it down: Evidence for structural compositionality in neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=rwbzMiuFQl.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. On compositional generalization of neural machine translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 4767–4780, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.368. URL https://aclanthology.org/2021.acl-long.368.

Weiduo Liao, Ying Wei, Mingchen Jiang, Qingfu Zhang, and Hisao Ishibuchi. Does continual learning meet compositionality? new benchmarks and an evaluation framework. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 33499–33513. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6a42b45af2b72e6e5b5e3a6fe695809f-Paper-Datasets_and_Benchmarks.pdf.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *ICML*, 2024. URL https://openreview.net/forum?id=J5VB1h3Aed.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015.

Chenyao Liu, Shengnan An, Zeqi Lin, Qian Liu, Bei Chen, Jian-Guang Lou, Lijie Wen, Nanning Zheng, and Dongmei Zhang. Learning algebraic recombination for compositional generalization. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1129–1144, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.97. URL https://aclanthology.org/2021.findings-acl.97.

Guisheng Liu, Yi Li, Yanqing Guo, Xiangyang Luo, and Bo Wang. Multi-attribute controlled text generation with contrastive-generator and external-discriminator. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5904–5913, Gyeongju, Republic of Korea, October 2022a. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.516.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pp. 423–439, Berlin, Heidelberg, 2022b. Springer-Verlag. ISBN 978-3-031-19789-5. doi: 10.1007/978-3-031-19790-1_26. URL https://doi.org/10.1007/978-3-031-19790-1_26.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5098–5139, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.279. URL https://aclanthology.org/2024.acl-long.279.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. @ crepe: Can vision-language foundation models reason compositionally? *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10910–10921, 2022. URL https://api.semanticscholar.org/CorpusID:254685851.

Gary Marcus. Deep learning: A critical appraisal, 2018.

Gary F. Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, Fei Xu, and Harald Clahsen. Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57(4):i–178, 1992. ISSN 0037976X, 15405834. URL http://www.jstor.org/stable/1166115.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

Jorge A Mendez and Eric Eaton. Lifelong learning of compositional structures. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ADWd4TJO13G.

Jorge A. Mendez and Eric Eaton. How to reuse and compose knowledge for a lifetime of tasks: A survey on continual learning and functional composition, 2023. URL https://arxiv.org/abs/2207.07730.

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023. doi: 10.1162/tacl_a_00562. URL https://aclanthology.org/2023.tacl-1.31.

Tomas Mikolov, Armand Joulin, and Marco Baroni. A roadmap towards machine intelligence. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, pp. 29–61, Cham, 2018. Springer International Publishing. ISBN 978-3-319-75477-2.

Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, and Tim Rocktäschel. Learning reasoning strategies in end-to-end differentiable proving. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4582–4598, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.364. URL https://aclanthology.org/2021.naacl-main.364.

Richard Montague. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven,, 1974.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. Pushdown layers: Encoding recursive structure in transformer language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3233–3247, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.195. URL https://aclanthology.org/2023.emnlp-main.195.

Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 953–962, 2021. URL https://api.semanticscholar.org/CorpusID:231786709.

Aliakbar Nafar, Kristen Brent Venable, and Parisa Kordjamshidi. Probabilistic reasoning in generative large language models, 2024.

Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11448–11459, Los Alamitos, CA, USA, June 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.01129. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01129.

Maxwell I. Nye, Armando Solar-Lezama, Joshua B. Tenenbaum, and Brenden M. Lake. Learning compositional rules via neural program synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=frVo9MzRuU.

Santiago Ontanon, Joshua Ainslie, Zachary Fisher, and Vaclav Cvicek. Making transformers solve compositional tasks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings*

of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3591–3607, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.251. URL https://aclanthology.org/2022.acl-long.251.

OpenAI. GPT-4 Technical Report, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Barbara Hall Partee. *Compositionality in Formal Semantics: Selected Papers of Barbara H. Partee*. Blackwell, Malden, MA, 2004.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan-Kaufmann, 1988.

Gabriel Poesia, Kanishk Gandhi, Eric Zelikman, and Noah D. Goodman. Certified reasoning with language models, 2023.

António Porto. Structural abstraction and application in logic programming. In Zhenjiang Hu and Mario Rodríguez-Artalejo (eds.), *Functional and Logic Programming*, pp. 275–289, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-45788-6.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=feiAVaSXdb.

Patrick Prosser. Hybrid algorithms for the constraint satisfaction problem. *Computational Intelligence*, 9(3): 268–299, 1993. doi: https://doi.org/10.1111/j.1467-8640.1993.tb00310.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.1993.tb00310.x.

Jorge Pérez, Javier Marinković, and Pablo Barceló. On the turing completeness of modern neural network architectures. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyGBdo0qFm.

Jorge Pérez, Pablo Barceló, and Javier Marinkovic. Attention is turing-complete. *Journal of Machine Learning Research*, 22(75):1–35, 2021. URL http://jmlr.org/papers/v22/20-302.html.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5368–5393, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.294. URL https://aclanthology.org/2023.acl-long.294.

Linlu Qiu, Hexiang Hu, Bowen Zhang, Peter Shaw, and Fei Sha. Systematic generalization on gSCAN: What is nearly solved and what is next? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2180–2188, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.166. URL https://aclanthology.org/2021.emnlp-main.166.

Parikshit Ram, Tim Klinger, and Alexander G. Gray. How compositional is a model? In *International Joint Conference on Artificial Intelligence 2023 Workshop on Knowledge-Based Compositional Generalization*, 2023. URL https://openreview.net/forum?id=OImyRhNLv3.

Parikshit Ram, Tim Klinger, and Alexander G. Gray. What makes models compositional? a theoretical view. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 4824–4832. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/533. URL https://doi.org/10.24963/ijcai.2024/533. Main Track.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. A benchmark for systematic generalization in grounded language understanding, 2020. URL https://arxiv.org/abs/2003.05161.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1889–1903, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.161. URL https://aclanthology.org/2023.acl-short.161.

Jenny Saffran, Seth Pollak, Rebecca Seibel, and Anna Shkolnik. Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, 105:669–80, 12 2007. doi: 10.1016/j.cognition.2006.11.004.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

Jurgen Schmidhuber. Towards compositional learning in dynamic networkstechnical report, 1990. URL https://api.semanticscholar.org/CorpusID:17783975.

Murray Shanahan, Kyriacos Nikiforou, Antonia Creswell, Christos Kaplanis, David Barrett, and Marta Garnelo. An explicitly relational neural network architecture. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Roger N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(4820): 1317–1323, 1987. doi: 10.1126/science.3629243. URL https://www.science.org/doi/abs/10.1126/science.3629243.

Hui Shi, Sicun Gao, Yuandong Tian, Xinyun Chen, and Jishen Zhao. Learning bounded context-free-grammar via lstm and the transformer: Difference and explanations. In *AAAI Conference on Artificial Intelligence*, 2021. URL https://api.semanticscholar.org/CorpusID:245329622.

H.T. Siegelmann and E.D. Sontag. On the computational power of neural nets. *Journal of Computer and System Sciences*, 50(1):132–150, 1995. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1995.1013. URL https://www.sciencedirect.com/science/article/pii/S0022000085710136.

Ankur Sikarwar, Arkil Patel, and Navin Goyal. When can transformers ground and compose: Insights from compositional generalization benchmarks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 648–669, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.41. URL https://aclanthology.org/2022.emnlp-main.41.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pp. 869–893, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.56. URL https://aclanthology.org/2023.emnlp-main.56.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4506–4515, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1458. URL https://aclanthology.org/D19-1458.

Paul Smolensky and Géraldine Legendre. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic GrammarVolume I: Cognitive Architecture (Bradford Books)*. The MIT Press, 2006. ISBN 0262195267.

Paul Smolensky, R. Thomas McCoy, Roland Fernandez, Matthew Goldrick, and Jianfeng Gao. Neurocompositional computing in human and machine intelligence: A tutorial. Technical Report MSR-TR-2022-5, Microsoft, May 2022a. URL https://www.microsoft.com/en-us/research/publication/neurocompositional-computing-in-human-and-machine-intelligence-a-tutorial/. 52 pages main text, 78 pages total, 11 figures, 2 Appendices, 239 references. For a short presentation of some of this material, see https://arxiv.org/abs/2205.01128 (to appear in AI Magazine).

Paul Smolensky, Richard McCoy, Roland Fernandez, Matthew Goldrick, and Jianfeng Gao. Neurocompositional computing: From the central paradox of cognition to a new generation of ai systems. *AI Magazine*, 43(3):308–322, Sep. 2022b. doi: 10.1002/aaai.12065. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/18599.

Sam Spilsbury, Pekka Marttinen, and Alexander Ilin. Generating demonstrations for in-context compositional generalization in grounded language learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, FL, USA, November 2024. Association for Computational Linguistics.

Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. Obtaining faithful interpretations from compositional neural networks, 2020. URL https://arxiv.org/abs/2005.00724.

Zoltán Gendler Szabó. Compositionality. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

Paul Thagard. Cognitive Science. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. Learning compositional representations for few-shot recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6371–6380, 2019. doi: 10.1109/ICCV.2019.00647.

Lazar Valkov, Dipak Chaudhari, Akash Srivastava, Charles Sutton, and Swarat Chaudhuri. Houdini: lifelong learning as program synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 8701–8712, Red Hook, NY, USA, 2018. Curran Associates Inc.

Josef Valvoda, Naomi Saphra, Jonathan Rawski, Adina Williams, and Ryan Cotterell. Benchmarking compositionality with formal languages. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6007–6018, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.525.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Zhengxuan Wu, Elisa Kreiss, Desmond C. Ong, and Christopher Potts. ReaSCAN: Compositional reasoning in language grounding. *NeurIPS 2021 Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=Rtquf4Jk0jN.

Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. ReCOGS: How incidental details of a logical form overshadow an evaluation of semantic interpretation. *Transactions of the Association for Computational Linguistics*, 11:1719–1733, 2023. doi: 10.1162/tacl_a_00623. URL https://aclanthology.org/2023.tacl-1.96.

Guangyue Xu, Joyce Chai, and Parisa Kordjamshidi. Gipcol: Graph-injected soft prompting for compositional zero-shot learning. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5762–5771, 2023a. URL https://api.semanticscholar.org/CorpusID:265128547.

Guangyue Xu, Parisa Kordjamshidi, and Joyce Chai. MetaReVision: Meta-learning with retrieval for visually grounded compositional concept acquisition. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12224–12236, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.818. URL https://aclanthology.org/2023.findings-emnlp.818.

Yelp. Yelp dataset, 2014. URL https://www.yelp.com/dataset/.

Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 192–199, 2014. doi: 10.1109/CVPR.2014.32.

Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. doi: 10.1109/iccv.2017.594. URL http://dx.doi.org/10.1109/ICCV.2017.594.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. SKILL-MIX: a flexible and expandable family of evaluations for AI models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Jf5gplvglq.

Hao Zheng and Mirella Lapata. Compositional generalization via semantic tagging. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1022–1032, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.88. URL https://aclanthology.org/2021.findings-emnlp.88.

Hao Zheng and Mirella Lapata. Disentangled sequence to sequence learning for compositional generalization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4256–4268, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.293. URL https://aclanthology.org/2022.acl-long.293.

Tianqi Zhong, Zhaoyi Li, Quan Wang, Linqi Song, Ying Wei, Defu Lian, and Zhendong Mao. Benchmarking and improving compositional generalization of multi-aspect controllable text generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6486–6517, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.351. URL `https://aclanthology.org/2024.acl-long.351`.