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

Anonymous authors
Paper under double-blind review

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, variety of computational models, and theoretical findings. Our primary focus is on linguistic benchmarks or combining language and vision. There are many related research in the vision community that focus on compositional concept learning. 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 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 these 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 these models in expressing compositional structures such as context-free grammars (Siegelmann & Sontag, 1995; Shi et al., 2022). 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 cognitive aspects, computational models, and evaluation paradigms in both theory and practice. Figure 1 shows the scope and structure of our survey, covering three main topics, that is, Cognitive aspects, Evaluation and Models. **For Cognitive Aspects**, we overview the different compositional learning facets that define abstract tasks for compositionality and connect them to the existing datasets 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 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 for testing the cognitive aspects of compositionality. **For 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.

Overall, the cognitive aspects lay the foundation of the concept of compositionality and define the different abstractions associated with it and tasks that are designed accordingly. The empirical evaluations use these tasks to evaluate compositionality using experimental performance. However, the studies look into the mathematical analysis and the functional form of the models independent of 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 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 and some existing notions (Fodor & Pylyshyn, 1988; Chomsky, 1956). Some of these tasks existed in older research by different terms such as productivity inherits a lot of previous research from length generalization. These tasks are becoming widely accepted tasks for compositional learning and there are current datasets that use these for their evaluation splits. We describe these measures in detail below.

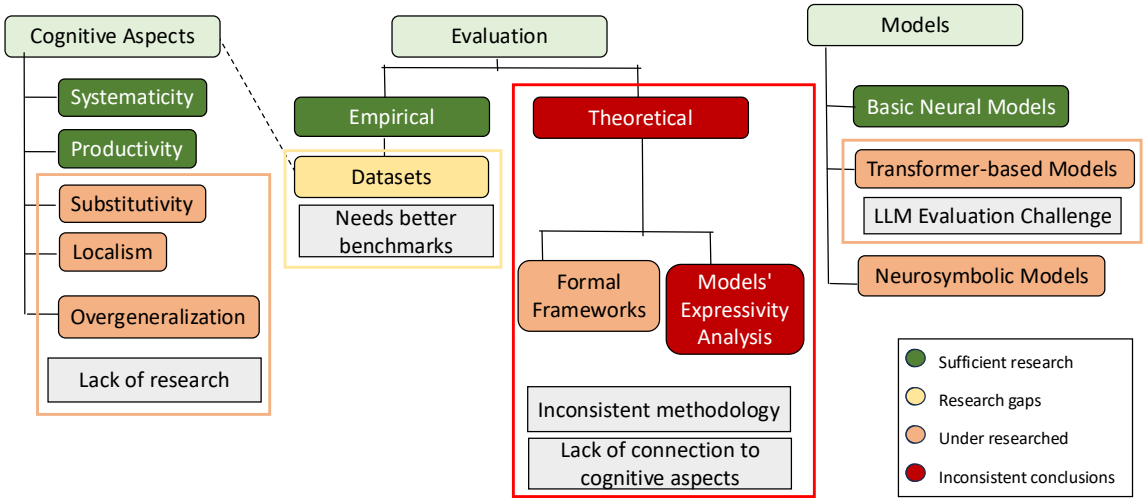

Figure 1: Outline of covered concepts in this survey, color coded to suggest research coverage and work required. We structure our study of compositional learning by dividing it into three parts of cognitive aspects, evaluation methods from both empirical and theoretical perspectives, and the compositional learning models and architectures. Green boxes depict topics that have sufficient research available, yellow boxes depict areas that include research gaps, orange boxes depict under-researched topics, and red boxes depict topics that have inconsistent research findings with non-uniform methodological frameworks. Gray boxes point to the challenge in that topic, which is discussed in the conclusion, Section 6. Systematicity and productivity are the most researched cognitive aspects with clear connections to evaluation benchmarks. Empirical evaluations are more developed than theoretical analyses, which are inconsistent or lacking. The connection between theoretical methods and cognitive aspects is also weak. Other boxes indicate areas of required work such as better benchmarks with datasets, LLM evaluation, and better expressivity for task specific models.

Systematicity or Novel Composition

Systematicity is one of the most commonly used notions of compositionality in evaluating the performance of architectures. 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 tried 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 accurate 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 evaluates model performance on the introduction of synonyms in expressions. For example, testing a model on the translation of the same sentence, switching between synonymous words such as donut/doughnut or aubergine/eggplant. 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" (Hupkes et al., 2020; Carnap, 1947). The local interpretation of compositionality says that these new phrases will have the same truth value 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", making a literal translation does not make any sense as the phrase is an exception and has a specific meaning different from the usual literal meaning. In this scenario, a model can get better performance by considering those 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 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., 2022) and demonstrating multi-hop reasoning skills through Chain-of-Thought prompting (Wei et al., 2023). 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 evaluation and understanding the of emergent abilities of LLMs.

3 Abstract Tasks and Datasets

Here, we categorize the datasets based on their type of compositionality mentioned 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. Productivity measure was often referred to as length generalization before the term started being commonly used. However, despite the abundance of datasets for these tasks, most datasets use synthetic data which poses a risk to its reliability in capturing the variation and complexity that exist in 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 on these tasks below.

3.1 Systematicity or Novel Composition

CREPE. This is a Compositional REPresentation Evaluation benchmark (CREPE) (Ma et al., 2023). 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, 2018; Mikolov et al., 2016). 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. The data is designed for the so-called 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 sentences. 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 is an extension of gSCAN that overcomes some of its llimitations by needing 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 on being given an image and 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 is an example of compositionality in vision as the whole tangram image is made out of parts and the model learns the way 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 split 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, and SenTube: Mixture (Liu et al., 2022), 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 novel composition of primitive states and transformations, applied to objects. There are 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., 2023) claims to cover some aspects of compositional evaluation for generative models. The task expects the models to generate text by combining various skills and impose some constraints on the generated text. While the aspect of compositionality is not identified in the paper, we categorize this under systematicity.

3.2 Productivity or Length Generalization

CREPE. The productivity split of CREPE (Ma et al., 2023) 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. There are questions generated at varying levels of complexity. There are various splits 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 seems to test both systematicity and productivity to some extent, though not explicitly, and cannot determine specific areas where a model's compositional behavior may be lacking.

COGS. From previous explanation, one of the splits of this dataset, Category 3 [Deeper Recursion] (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. (2022)
CLUTRR	✓	✗	✓	✗	✗	✗	✗	Gontier et al. (2020), Minervini et al. (2020) Sinha et al. (2019)
SQOOP	✓	✓	✓	✗	✗	✗	✗	D'Amario et al. (2022), 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 & Ilin (2023)
ReaSCAN	✓	✓	✓	✓	✗	✗	✗	Kamali & Kordjamshidi (2023)
CREPE	✓	✓	✓	✓	✗	✗	✗	Lin et al. (2023), 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 understanding (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. Dealing with noisy and uncertain data is hard with pure symbolic AI. However, 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) (Hochreiter & Schmidhuber, 1997), convolutional neural network (CNN) (LeCun et al., 1989), and Transformers for sequence-to-sequence language processing tasks on their proposed PCFG SET tasks- systematicity, productivity, substitutivity, localism, 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, 2018) 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 Transformers, 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. This indicates they make predictions on shallow reasoning and memorization of similar subgraph patterns seen during training as opposed to reasoning holistically based on true compositional reasoning. 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.

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 of Transformers were evaluated- 1) type of position encoding, 2) use of copy decoders, 3) model size, 4) weight sharing, and 5) use of intermediate representations for prediction- on several different datasets and benchmarks. 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 just relative position biases is not enough. 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 were not helpful. 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 specific to a task and were only tested on COGS and CFQ.

Another type of research deviates from evaluating current conventional architectures and instead focuses on designing a new architecture that can compositionally generalize better.In their model, 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. The 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, their proposed evaluation benchmark based on the target identification subtask of ReaSCAN. Another 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 these 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, 2022). 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. There have been several architectures and theories proposed over the years aimed at modeling compositionality in their design. This is connected to the idea that neural architectures are structurally compositional (Lepori et al., 2023), that is, they take advantage of subroutines to break down complex tasks. One of the earlier examples of this type of architecture includes neural modular networks (Andreas et al., 2017). 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, 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), which is a framework for learning to ground and reason across domains with a differentiable, domain-independent, first-order logic–based program executor. It addresses a lack of generalizability across domains in some works like VisProg that we previously discussed, such as a lack of generalization of 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 computing architectures (Smolensky et al., 2022a).

5 Theoretical Findings: Mathematical Formulations of Compositionality

Theoretical analysis is fundamental for further explaining 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 expressivity upper-bounds 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 transformers cannot handle an infinite input length. This conclusion limited the compositionally of the transformer 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 it struggles 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 the older paper (Dehghani et al., 2019). Based on these analyses, they further confirm the lack of productivity aspect of transformer architecture from the limited training length.

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., 2023). 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 publication presents a positive theoretical analysis of reasoning capabilities by studying the chain of thought(CoT) (Wei et al., 2023), which 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.

	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 2022

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

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

Figure 1 shows the research coverage of the main topics identified and discussed in this survey. Green boxes depict topics that have sufficient research available, yellow boxes depict areas that include research gaps, orange boxes depict under-researched topics, and red boxes depict topics that have inconsistent research findings with non-uniform methodological frameworks. Gray boxes point to the challenges in that topic which are discussed further in this section. For Cognitive Aspects, only systematicity and productivity out of the five types mentioned, are well-researched and have clear connections to evaluation benchmarks. Considering that these are the five main metrics of compositionality defined, to expand on our evaluation capabilities we should work towards testing 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, the evaluation of LLMs and the fundamental design decisions

for compositional generalization come with new challenges. We describe some of these challenges in detail below.

Less Explored Facets of Compositionality. As seen in Figure 1, only Systematicity and Productivity have been well researched and have well established connections to evaluation benchmarks. While the other three were introduced as fundamental types of compositionality, they have not received as much attention as they seem to be less 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 the 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). Thus, pointing attention to the other three types of measures can contribute to new formal evaluation benchmarks for more robust systems and dealing with the challenges in the LLMs era.

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., 2023; 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). Though the benchmarks are built for evaluation of certain tasks, the use of mostly synthesized data risks the effective 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 that try to address this issue (Yu et al., 2023). 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 empirical 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 need 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 not included. 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.

References

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via the "edge of stability", 2023.

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, 2017.

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?, 2019.

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://arxiv.org/abs/2305.14428.

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, 2015.

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, 2022.

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

Ta-Chung Chi, Ting-Han Fan, Alexander I. Rudnicky, and Peter J. Ramadge. Transformer working memory enables regular language reasoning and natural language length extrapolation, 2023.

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 Jürgen Schmidhuber. The devil is in the detail: Simple tricks improve systematic generalization of transformers, 2022.

Vanessa D'Amario, Tomotake Sasaki, and Xavier Boix. How modular should neural module networks be for systematic generalization?, 2022.

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 Łukasz Kaiser. Universal transformers, 2019.

Roberto Dessì and Marco Baroni. Cnns found to jump around more skillfully than rnns: Compositional generalization in seq2seq convolutional networks, 2019.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2023.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective, 2023.

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 J. Mooney. Systematic generalization on gscan with language conditioned embedding, 2020.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning, 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 Christopher Pal. Measuring systematic generalization in neural proof generation with transformers, 2020.

Liangke Gui, Yingshan Chang, Qiuyuan Huang, Subhojit Som, Alex Hauptmann, Jianfeng Gao, and Yonatan Bisk. Training vision-language transformers from captions, 2023.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training, 2022.

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, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

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, 2023. URL https://arxiv.org/abs/2310.16035.

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, 2018.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: how do neural networks generalise?, 2020.

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. State-of-the-art generalisation research in nlp: A taxonomy and review, 2023.

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 D. Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes, 2022.

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 Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016.

Danial Kamali and Parisa Kordjamshidi. Syntax-guided transformers: Elevating compositional generalization and grounding in multimodal environments, 2023.

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, 2020.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks, 2023.

Najoung Kim and Tal Linzen. Cogs: A compositional generalization challenge based on semantic interpretation, 2020.

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, 2023.

Kris Korrel, Dieuwke Hupkes, Verna Dankers, and Elia Bruni. Transcoding compositionally: using attention to find more generalizable solutions, 2019.

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, 2018.

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, 2019.

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, 2023.

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, 2023.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation : Learning to solve and explain algebraic word problems, 2017.

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, 2021.

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 2022. 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, 2023.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning?, 2024. URL https://arxiv.org/abs/2309.01809.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally?, 2023.

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, 2021. URL https://arxiv.org/abs/2007.07732.

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, 2023.

Tomas Mikolov, Armand Joulin, and Marco Baroni. A roadmap towards machine intelligence, 2016.

Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, and Tim Rocktäschel. Learning reasoning strategies in end-to-end differentiable proving, 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 D. Manning. Pushdown layers: Encoding recursive structure in transformer language models, 2023.

Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning, 2021. URL https://arxiv.org/abs/2102.01987.

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, 2021.

Maxwell I. Nye, Armando Solar-Lezama, Joshua B. Tenenbaum, and Brenden M. Lake. Learning compositional rules via neural program synthesis, 2020.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task, 2023.

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, 2022. URL https://arxiv.org/abs/2203.02155.

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, 2023.

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, 2023.

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: With supplement, 2024.

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.

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 Ander 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, 2023. URL https://arxiv.org/abs/2310.18018.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage?, 2023.

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, 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, 2022.

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, 2022.

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, 2023.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. Clutrr: A diagnostic benchmark for inductive reasoning from text, 2019.

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, R. Thomas McCoy, Roland Fernandez, Matthew Goldrick, and Jianfeng Gao. Neurocompositional computing: From the central paradox of cognition to a new generation of ai systems, 2022b.

Sam Spilsbury and Alexander Ilin. Improved compositional generalization by generating demonstrations for meta-learning, 2023.

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, 2022. URL https://arxiv.org/abs/2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Zhengxuan Wu, Elisa Kreiss, Desmond C. Ong, and Christopher Potts. Reascan: Compositional reasoning in language grounding, 2021.

Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. Recogs: How incidental details of a logical form overshadow an evaluation of semantic interpretation, 2023.

Guangyue Xu, Joyce Chai, and Parisa Kordjamshidi. Gipcol: Graph-injected soft prompting for compositional zero-shot learning, 2023a. URL https://arxiv.org/abs/2311.05729.

Guangyue Xu, Parisa Kordjamshidi, and Joyce Chai. Metarevision: Meta-learning with retrieval for visually grounded compositional concept acquisition, 2023b. URL https://arxiv.org/abs/2311.01580.

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, 2023. URL https://arxiv.org/abs/2310.17567.

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, 2022.

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, 2024. URL https://arxiv.org/abs/2404.04232.