Intrinsic Systematicity Evaluation: Evaluating the intrinsic systematicity of LLMs

Stanford CS224N Custom Project

Ayush Chakravarthy
Symbolic Systems Program
Stanford University
akchak@stanford.edu

Abstract

LLMs have learned a surprising amount of traits that are considered the hallmarks of human cognition - such as the ability to plan, reason, and take-on personas. In this vein, recent work (Zhou et al., 2023; Drozdov et al., 2022) showed that an LLM can be prompted to solve systematic generalization benchmarks such as SCAN (Lake and Baroni, 2018) and COGS (Kim and Linzen, 2020) purely in-context *i.e*. without ever any gradient update propogating task information into the model. The prompting strategy they develop can be broadly characterized as a method to decompose the grammar into simple exemplars, and showing the model both the exemplars and the rules to combine these exemplars. On its face, these findings are remarkable, the fact that LLMs are able to perfectly systematically generalize zero-shot is evidence that these models are a leap toward human cognition. However, through this project, we show preliminary work that these findings cannot be interpreted as suggesting that LLMs can generalize perfectly systematically. We argue that the prompting methods developed by Zhou et al. (2023) and Drozdov et al. (2022) render the systematic generalization benchmarks they evaluate on an invalid measure of the *intrinsic* systematicity possessed by the LLM. Toward this, we present a simple benchmark called **Intrinsic Systematicity Evaluation (ISE)** and show that modern LLMs upto 70-B parameters struggle on ISE.

1 Key Information to include

- Mentor: Shikhar Murty
- External Collaborators (if you have any): Ahmad Jabbar, Jake Russin
- Sharing project: N/A

2 Introduction

Systematicity has been considered one of the key inductive biases that underlie human intelligence. It refers to the ability to recompose previously seen concepts into novel combinations. Paraphrasing from Lake and Baroni (2018), consider the thought experiment where a person knows the meaning and usage of words such as "twice," "and," and "again," once she learns a new verb such as "to dax" she can immediately understand or produce instructions such as "dax twice and then dax again." This human ability to generalize zero-shot to novel combinations allows for incredibly sample efficient learning and information acquisition. During the first AI revolution, Fodor and Pylyshyn (1988) famously argued that Deep Neural Networks could not be viable models of human cognition as they failed at demonstrating this key behavior. Furthermore, multiple recent work suggest that the transformer model itself struggles to learn systematic representations (Chakravarthy et al., 2022; Wu et al., 2024).

Despite these findings, scaling the transformer architecture to predict the next word on an internet-sized corpus of text, has yielded a model that are capable of sophisticated behavior such as producing coherent reasoning chains, writing computer code, and conversing with a human through a web-interface - all behaviors thought to have require systematic representations. Perhaps, the most surprising finding of scaling-up transformer language models is the emergence of in-context learning (Brown et al., 2020). In-Context Learning refers to the observation that an LLM is able to pick up and solve a task, purely by providing more examples in the context window *i.e.* with no gradient information flowing through into the model parameters. These results show the immense capabilities that LLMs possess, and indicate that they are a leap toward building systems equivalent to human cognition.

However, this still leaves out a crucial question which is - is an LLM able to learn as *systematically*? Toward this issue, there have been multiple efforts Zhou et al. (2023); Drozdov et al. (2022), that claim to have solved popular systematicity evaluation benchmarks (Lake and Baroni, 2018; Kim and Linzen, 2020), which suggests that we have an answer to the question! LLMs are indeed perfectly systematic. However, on closer examination, there are multiple confounds that are not addressed in their papers. The key issue, is that the models are *given* the novel combinations in-context, leaving the model little to no *generalizing* left to do. The second key confound, is that since the benchmarks are in English (or contain English words), through the course of pretraining the models have already learned rich embeddings for both the syntactic roles and semantic interpretations of the words that the model is supposed to generalize to.

In this preliminary paper, we aim to address the aforementioned issues in previous literature. We introduce a novel systematicity benchmark for LLMs, and evaluate the Llama3 series on our benchmark. We show that LLMs struggle when evaluated on a benchmark that probes their *intrinsic* systematicity, and through multiple variations to our dataset analyze the different failure modes that LLMs still possess, despite their tremendous utility.

3 Related Work

3.1 LLM Evaluation

The development of LLMs can only be measured through the evaluation benchmarks that were developed alonside them. The bechmarks themselves have evolved from general language understanding benchmarks such as GLUE (Wang et al., 2019) and its successor SuperGLUE (Wang et al., 2020), to evaluations on more complex reasoning (Srivastava et al., 2023), language generation (Liang et al., 2023), and knowledge understanding through question-answering (Wang et al., 2024; Yue et al., 2023). However, as outlined by the desideratum presented in Hupkes et al. (2023), there still isn't a clear systematic generalization benchmark for evaluating LLMs.

In order to facilitate apples-to-apples comparison between LLMs that are becoming increasingly closed-off, several leaderboards and evaluation testbeds have been established, such as lms (2024) and Contributors (2023). However, it has been observed that with advancements and updates to the models, these leaderboards become increasingly saturated and hard to use as a fair measure. Perhaps more concerningly, evaluation efforts are further conflated by findings (Mishra et al., 2022) that LLM performance drastically varies by small non-semantic details in the way the evaluation is setup.

3.2 Systematic Generalization

Systematic Generalization has long been studied by cognitive scientists, however, for the sake of brevity, we review recent literature on the study of systematic generalization specific to the transformer architecture [1]. Most approaches aiming to address the problem of systematicity in transformers primarily deal with data augmentation techniques, such as Chakravarthy et al. (2022) induce systematicity through training the model on grammatical roles, Jiang et al. (2022) through mutual-exclusivity training, Patel et al. (2022) through upscaling the number of primitives, and most recently Lake and Baroni (2023) through structuring the data as a meta-learning episode. As for the question of evaluation of systematicity, the picture is still unclear, recent work Wu et al.

[1] See (Russin et al., 2024) for a more rigorous historical treatment of the problem and study of systematic generalization.

```
dax -> YELLOW
fep -> RED
blicket -> PINK
gazzer -> GREEN
blicket kiki fep -> RED PINK
dax kiki blicket -> PINK YELLOW
blicket wif -> PINK PINK PINK
blicket zup fep -> PINK RED PINK
dax wif -> YELLOW YELLOW YELLOW
fep zup dax -> RED YELLOW RED
blicket wif kiki fep -> RED PINK PINK PINK
fep kiki dax zup blicket -> YELLOW PINK YELLOW RED
blicket kiki fep wif -> RED RED RED PINK
fep zup dax kiki blicket -> PINK RED YELLOW RED
fep kiki gazzer ->
```

support

query

Figure 1: Example datapoint containing 14 sample translations and 1 query translation.

dax -> RED
wif -> GREEN
lug -> BLUE
zup -> YELLOW
[u1] fep -> [u1] [u1] [u1]
[u1] blicket [u2] -> [u1] [u2] [u1]
[x1] kiki [x2] -> [x2] [x1]

(a) An example of a SCAN-equivalent grammar

tufa -> YELLOW
wif -> BLUE
lug -> PURPLE
fep -> RED
[x1] gazzer -> [x1] [x1] [x1]
[x1] kiki [u1] -> [x1] [u1] [x1]
[x1] zup -> [x1] [x1]

(b) An example of a randomly generated grammar

Figure 2: Examples of grammars that we sample examples from

(2024) suggest that small implementational details overshadow the conclusions drawn from previous benchmarks (Kim and Linzen, 2020). Additionally, it as yet unclear as to what constitutes a 'fair' systematicity benchmark (Kim and Smolensky, 2024), both for humans and for AI systems.

4 Approach

Our dataset construction is split into two sections. The first being developing and evaluating on SCAN-equivalent grammars and the next being developing a grammar generator and evaluating on the randomly generated grammars. In both cases, we fix a number of *primitives* (which can be thought of as a simplification of verbs), randomly select one primitive to be 'held-out' and generate 24 (unless specified otherwise) sentences according to the grammar. All sentences that do **not** contain the 'held-out' primitive are populated into the `support` set and the sentences that do contain the 'held-out' primitive into the `query` set. Finally, to ensure productivity within a grammar, all grammars have a hard-coded left-recursion rule.

Our first task developing SCAN-equivalent grammars builds off of the *few-shot instruction learning task* presented in Lake and Baroni (2023). We construct an evaluation dataset using the *completion* condition in which the model receives 14 support samples, then is presented with a query. The query in Figure 1, for example, is `fep kiki gazzer ->` [2]. We populate the samples and queries in such a way that respects the add-prim generalization split. To elucidate what this means, in Figure 1, the primitives `dax, fep, blicket, gazzer` are analogous to the SCAN primitives `walk, look, run, jump`. As the add-prim split holds-out occurences of the `jump` primitive in the full grammar

[2]As described in the Project Proposal, the reason we use random words as compared to merely sampling from the SCAN splits is because we want to challenge the model to learn to generalize in a linguistic context it has not seen in its pretraining data.

Algorithm 1 An algorithm to randomly sample grammars

Require: Number of primitives *nprims*; Number of rules *nrules*; A function to randomly sample and pop the sampled element from the set **1-random-sample**; A function to randomly sample a random number of elements from a set **random-sample**; Cardinality of a set function **length**; Placeholder variable for primitives x; Placeholder variable for function words u.

input-symbols = {`dax, fep, blicket, gazzer, kiki, wif, zup,` $\cdots$}
output-symbols = {`YELLOW, RED, PINK, GREEN,` $\cdots$}
rules = {}
while length(rules) $<$ *nprims* **do**
  rules $\leftarrow$ '**1-random-sample**(input-symbols) $\rightarrow$ **1-random-sample**(output-symbols)$'$
end while
while length(rules) $<$ *nrules* **do**
  LHS = **1-random-sample**($\{`xu', `uu', `uux', `xux', , `xuu', `uuu'\}$)
  Assign u in LHS to **1-random-sample**(input-symbols)
  RHS = **random-sample**(output-symbols)
  rules $\leftarrow$ $`LHS \rightarrow RHS'$
end while
rules $\leftarrow$ $`ux \rightarrow ux'$

from the training set, we similarly hold out the `gazzer` primitive's occurences from the samples. We make similar translations for the target vocabulary, such that the resulting grammar is equivalent to SCAN. By constructing queries of the held-out primitive in varying grammatical constructs, for each set of samples, we create 10 datapoints. Finally, we create an evaluation dataset of 2000 datapoints of the form described in Figure 1 by creating 200 unique sets of samples through varying the grammatical roles of `dax, fep, blicket, gazzer, kiki, wif, zup`. Additionally, we include an example of a SCAN-equivalent grammar in Figure 2a.

For the second task of evaluating on randomly generated grammars, we present our random grammar generation in Algorithm 1 and a random generated grammar in Figure 2b. All details regarding the population of the `support` and `query` set are maintained from the SCAN-equivalent grammars, the only difference here is that the grammar has been generated randomly.

5 Experiments

We evaluate the Llama3 series of models on our systematic generalization evaluation datasets. Following from previous work that evaluates on the SCAN dataset, we use Exact-Match (EM) accuracy as the dependent measure. We ran our experiments vLLM with 2 Nvidia A100 GPUs and set sampling temperature to 0 for determinism and replicability.

Our experiments are centered around 4 central variations to the way the data is presented to the model. These variations are:

1. **Vanilla** Dataset: the sample set is presented in a randomly shuffled ordering.
2. **Sorted** Dataset: the sample set is presented in a increasing-by-length sorted order. As indicated by Lake and Baroni (2023), this condition is expected to do better as it reduces uncertainty about the length of the query translations.
3. **Reverse** Dataset: the set of samples are ordered randomly, however the source and target languages are switched. [3]
4. **Reverse-Sorted** Dataset: the set of samples are ordered in increasing length, however the source and target languages are switched.

Models	**Vanilla**	**Sorted**	**Reverse**	**Reverse-Sorted**
Llama3-8B	42.65	32.00	24.35	22.70
Llama3-8B-Instruct	47.05	41.50	26.30	25.45
Llama3-70B	45.45	49.05	28.30	27.15
Llama3-70B-Instruct	50.20	50.55	27.55	26.05

Table 1: EM accuracy for linguistic variations on the dataset.

Models	**Vanilla**	primed-**Vanilla**
Llama3-8B	42.65	45.65
Llama3-8B-Instruct	47.05	46.30
Llama3-70B	45.45	49.30
Llama3-70B-Instruct	50.20	52.45

(a) EM accuracy with priming.

Models	**Random Source**	**Random Target**
Llama3-8B	2.35	3.05
Llama3-8B-Instruct	5.30	5.50
Llama3-70B	3.75	3.60
Llama3-70B-Instruct	3.45	4.45

(b) EM accuracy when input-label bindings are broken.

Table 2: Effects of confounds in prompting the models.

5.1 SCAN-equivalent Grammars

5.1.1 Linguistic evaluations

The first axes of variation that we evaluate our models across are linguistically-motivated. The first comparative result that we present is the between the **Vanilla** and **Reverse** columns of Table 1. As the target language is constructed with common colors, it can be expected that the model have learned rich embeddings about these words. However, since the source language consists of words that are random, they most likely are represented as groups of embeddings that the model has not previously encountered. As we observe, the model shows an impressive ability to show systematic generalization when the embeddings are *information-poor*, as compared to the converse [4].

The next interesting finding, is in contrast to Lake and Baroni (2023), we don't see consistent improvements when giving the model a set of samples that are sorted in increasing length. For the Llama3-8B variants, both models suffer on the **Sorted** variant of the dataset. And this phenomenon is further observed when comparing the **Reverse** and **Reverse-Sorted** columns in Table 1 suggesting that this heuristic observed in previous work does not hold for all variants of the same systematic generalization evaluation setup.

5.1.2 Prompt-sensitivity evaluations

Two potential weaknesses that we try to address in this section is the model's sensitivity to confounds within the prompt that affect the dependent measure.

The first, is the effect of 'priming' the model. Priming the model, in our context, amounts to simply prepending the prompt given to the model with - "You are a subject in a psycholinguistics experiment.". The effect of this change are presented in Table 2a. As observed in previous literature, these kinds of changes have slight impacts on the dependent measure with the largest change observed with the Llama3-70B model. The second, is a sanity check based on recent work Min et al. (2022); Weber et al. (2023) suggesting that the input-label bindings in in-context examples don't hurt performance. But as we observe by the performance collapse in Table 2b where we shuffle either the source or the target datapoints in a set of samples drawn from the **Vanilla** dataset, we can safely conclude that our task is sufficiently different from Min et al. (2022) that we can proceed with this task design.

5.2 Random-generated Grammars

For this section we only report results with Llama3-70B and Llama3-70B-Instruct. The reason for this decision, is that these models were the highest performing models. We present our first set of results in Table 3a. It is evident from the complete lack of traction that these models get on this novel

[3] Since the grammars we sample from are relatively simple, two source sequences will always lead to two distinct target sequences, which makes this evaluation reasonable.

[4] This result suggests all models are able to learn bindings from random sets of embeddings to information-rich embeddings a lot quicker than the other way around.

Models	**Vanilla**	**Sorted**
Llama3-70B	0.47	0.97
Llama3-70B-Instruct	0.72	0.97

(a) EM accuracy for episodes produced by random generated grammars.

Models	**Vanilla**	**Sorted**
Llama3-70B	0.55	0.85
Llama3-70B-Instruct	0.59	0.80

(b) EM accuracy for episodes produced by the simplified random generated grammars.

Table 3: Evaluations on randomly generated grammars.

Models	10p-**Vanilla**	10p-**Sorted**	20p-**Vanilla**	20p-**Sorted**
Llama3-70B	0.26	0.26	0.36	0.34
Llama3-70B-Instruct	0.26	0.26	0.29	0.27

Table 4: EM accuracy for episodes produced by increasing context by increasing number of primitives.

task, that these models seem to be struggling on either grammar acquisition and being systematic in those grammars or the generalization gap being too far for the models to reasonably generalize over.

In order to pry out the failure modes observed in Table 3a, we attempt to reduce the effects of grammar acquisition by following the results presented in Patel et al. (2022). Their central result was showing that benchmarks such as SCAN are under-specified, and increasing the number of primitives, yields a monotonic increase in the dependent measure on the generalization split. We hypothesize that this result stems from the model accruing a larger set of evidence over the same grammar rather than the intrinsic systematicity of representations learned by the model. Following this hypothesis, we create longer contexts, not by naively sampling more examples from the randomly generated grammar, but by holding the characteristics of the randomly generated grammars, increasing the number of primitives, and maintaining the proportion of examples between the `support` set and `query` set (only changing *nprims* in Algorithm 1, and scaling the number examples while respecting the proportion between the support and query set).

We present the results for evaluation on data upscaled through this technique in Table 4. Each column's prefix indicates the number of primitives that are present in the grammar. And, on average the datasets containing 10 primitives, roughly contained 45 support samples and 30 query samples, while the datasets containing 25 primitives, roughly contained 90 support samples and 60 query samples. Naturally, these datasets ended up being larger than the previously discussed evaluation dataset, with 6000 and 12000 evaluation episodes each. Even with these modifications, the models struggle with our generalization split, presumably because the generalization gap was too large.

The final confound which we aim to mitigate for is the challenge of reasoning over infrequently seen embeddings. One plausible explanation for the complete collapse of these models is that the there is a slight non-zero chance that the random sequences of embeddings that some of the words that are present in the generated grammars can cause an additional layer of difficulty which is not really a measure of systematicity. In order to evaluate the effects of this hypothesis, we construct *simplified* variants of the grammars, by changing the set of input-symbols in Algorithm 1, to the set of capital English letters, that is, modify $\{\texttt{dax, fep, blicket, gazzer, kiki, wif, zup, } \cdots\}$ to $\{\texttt{A, B, C, D, E, F, G, } \cdots\}$ and use this set to produce grammars, and thereby evaluation episodes. We present results for this evaluation in Table 3b. And, following the trend, we still find that the models struggle to gain traction on this variation of the dataset.

6 Analysis

By comparing the results between Sections 5.1 and 5.2, we come to see the difference in between the in-context systematicity across grammars. Since the SCAN grammar we experimented with in Section 5.1, is a plausible grammar which could be generated by the grammar generator that we presented in Algorithm 1. As a result, it is plausible to expect that there exist certain grammatical structures that LLMs find *easier* to acquire and be systematic over and other grammatical structures that they find much harder to acquire and be systematic over. Comparing the results also indicates these models learn a *prior* over linguistic structures. Our results when viewed in conjunction with the results from Akyürek et al. (2024) which suggest that LLMs acquire language through specialized "n-gram heads", also supports this hypothesis.

Another explanation for this difference in results, is the possibility of data contamination. It is not unlikely that a model pretrained on internet scale data, and in particular, arXiv papers, could have been trained on literature either referencing or describing the SCAN grammar. As a result, the performance on the SCAN equivalent grammars could be a mirage and the collapse in performance on random grammars perhaps indicates a complete lack of systematicity within the learned representations even after the massive data and compute that were invested into these models.

Unfortunately, it was difficult for us to perform a clean and detailed failure-mode analysis of the evaluation and try to come up with an explanation for the collapse of performance in Section 5.2 owing to compute restrictions and the difficulty of disentangling failures on certain linguistic structures as nested in other linguistic structures. One example of this, is the empirical observation that all the models we evaluated on struggled on acquiring the 'reversal' rule. In most examples we found of this rule, the models exhibited the 'iconic-concatenation' [5] failure mode identified also in humans by Lake and Baroni (2023). An example of this rule is the last rule from Figure 2a [6]. However, due to the left recursion and the phrase rules, it is harder to construct an accurate measure of the accuracy on just this rule to thereby concretely argue about the acquisition of this rule.

7 Conclusion

In this preliminary paper, we presented a novel evaluation benchmark called ISE. We showed that our benchmark is a significantly challenging benchmark for reasonably-sized LLMs, and that previous work aiming to address the issue of systematicity in LLMs dramatically overestimate the abilities of LLMs. However, there are a few key limitations that our evaluations still lack. The first, is that benchmarks such as SCAN and COGS have only be claimed to solved through prompting strategies (Drozdov et al., 2022; Zhou et al., 2023). It is, therefore, important for us to replicate these prompting strategies into our evaluation strategy and study whether prompting can solve our task. The second key limitation is the issue of explainability. We have results suggesting a failure mode of LLMs, however, we still have no understanding as to why they fail, or the algorithm they try to employ in the cases they do succeed. Addressing these limitations will be the focus of future work on this project.

8 Ethics Statement

Since our work develops a new evaluation benchmark, we do not forsee any intrinsic ethical or societal impacts. However, since our evaluation highlighted a previously unknown and understudied failure mode of LLMs, it does highlight certain societal impacts which arise from the training of LLMs. Despite the limited size of the models which we evaluated on, it is not implausible to believe that a similar failure mode could be exposed on larger models such as GPT-4. This raises the question of the necessity for the immense resources that are being pushed into developing frontier foundation models. This claim would only hold worth, if we run experiments with humans and decisively show that humans can solve these tasks while LLMs cannot.

The second ethical concern is that around data privacy. For this discussion, let us assume that the Llama3 model has been trained on some information around the SCAN benchmark. If the gap between an equivalent generalization gap on trained data versus unseen data is around 50%, these models are much better at memorizing, than reasoning when brought truly out-of-distribution. The reason this result can be cast as an ethical issue, is around tasks which require creative expression. Imagine the counterfactual where an LLM is trained on all of the writings of a particular author, the model would be able to perfectly mimic the behavior of the author and write in their style. Such a model, can be very easily used to put multiple such authors out of a job.

A potential mitigation strategy (which is quite impractical) is the manual/automated vetting of the pretraining corpus. Through this, it could be estimated how much these models have learned and what are other such aspects the models have memorized, that we do not want them to.

[5] The bias of producing random target vocabulary tokens to match the length of the source sequence.

[6] $[x1]$ kiki $[x2] \rightarrow [x2][x1]$

References

2024. Lmsys arena: A platform for evaluating language models. `https://arena.lmsys.org/`. Accessed: 2024-06-06.

Ekin Akyürek, Bailin Wang, Yoon Kim, and Jacob Andreas. 2024. In-context language learning: Architectures and algorithms.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Ayush K Chakravarthy, Jacob Labe Russin, and Randall O'Reilly. 2022. Systematicity emerges in transformers when abstract grammatical roles guide attention. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 1–8, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models.

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

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. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Yichen Jiang, Xiang Zhou, and Mohit Bansal. 2022. Mutual exclusivity training and primitive augmentation to induce compositionality.

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

Najoung Kim and Paul Smolensky. 2024. Structural generalization of modification in adult learners of an artificial language.

Brenden Lake and Marco Baroni. 2023. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions.

Arkil Patel, Satwik Bhattamishra, Phil Blunsom, and Navin Goyal. 2022. Revisiting the compositional generalization abilities of neural sequence models.

Jacob Russin, Sam Whitman McGrath, Danielle J. Williams, and Lotem Elber-Dorozko. 2024. From frege to chatgpt: Compositionality in language, cognition, and deep neural networks.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark.

Lucas Weber, Elsa M. Bruni Bruni, and Dieuwke Hupkes. 2023. The icl consistency test. *ArXiv*, abs/2312.04945.

Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. 2024. Recogs: How incidental details of a logical form overshadow an evaluation of semantic interpretation.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models.