CME LIVE OPEN ACCESS

Journal of
Advances in Civil and Mechanical Engineering

Research Article

Understanding Natural Language Beyond Surface by LLMs

Yong Yang*

University of Illinois Urbana-Champaign Master in Computer Science, Data Science Track.

Corresponding Author: Yong Yang, University of Illinois Urbana-Champaign Master in Computer Science, Data Science Track.

Received: 2024 Nov 12 **Accepted:** 2024 Nov 29 **Published:** 2024 Dec 07

Abstract

In recent years, transformer-based models like BERT and ChatGPT/GPT-3/4 have shown remarkable performance in various natural language understanding tasks. However, it's crucial to note that while these models exhibit impressive surface-level language understanding, they may not truly understand the intent and meaning beyond the superficial sentences. This paper is a survey of studies of the popular Large Language Models (LLMs) from various research and industry papers and review the abilities in term of comprehending language understanding like what human have, revealing key challenges and limitations associated with popular LLMs including BERTology and GPT alike models.

Keywords

Natural Language, BERTology, Large Language Models (LLM), Semantic Knowledge

1. Introduction

In this paper, I conducted extensive research and strive to understand the capabilities and boundaries of popular Large Language Models (LLM) - BERT, GPT and its sibling variants. The study starts with BERT and its variants (mBERT and RoBERT etc) architecture which is called BERTology. It reveals the knowledge BERT may have: Syntactic Knowledge, Semantic Knowledge and World Knowledge, Commonsense Knowledge, and Reasoning. In order to measure the extent to which the semantic understanding and reasoning capability of the models have reached, we also explore to study the definition of meaning. While NLP gains increasingly significant public exposure nowadays, it is crucial to make it clear on the distinction between the linguistic word form and semantic meaning. Next, we also study the reasoning capability of GPT3 and how to improve the reasoning capability by Chain-of-Thought (CoT) prompting which involves zero-shot and few-shot reasoners as prompting techniques. After all, we summarize the latest studies with existing capabilities and limitations that these popular LLMs have gained today.

2. Background

This is a basically literature review and expected to answer a question: To what extend do LLMs understand the natural language? The focus of the study will be

1 How LLM understand natural language, whether they truly understand the intent and meaning beyond the superficial sentences and the knowledge and capabilities of LLM today?

2 What are the challenges and limitations towards truly understanding natural language today?

There are many LLMs today and they are growing every day. I don't try to enumerate all the models, instead just focus on these popular and well-known models based on Transformer: BERT, GPT and its siblings. Also, understanding natural language with Large Language Models (LLMs) is a broad subject. After consulting the papers, it became apparent that delving deeper into the topic and surpassing the minimum of 4-5 papers is necessary to thoroughly investigate and address the topic.

2.1 BERTology and GPT Overview

BERT BERT is a multi-layer of transformer encoder that comprise multiple self-attention 'head'. It consists of two stages: pre-training and fine-tuning. Pre-training uses Masked Language Modeling (MLM) and Next Sequence Prediction (NSP). It is based on Bidirectional Encoder Representations from Transformers, which alleviates the unidirectional constraints by MLM pre-training objective. mBERT (Multilingual BERT) is a variant of BERT pre-trained to support multilingual natural language processing tasks. RoBERT (Robustly optimized BERT) introduced a modification of BERT by removing Next Sentence Prediction. These variants share the fundamental architecture with language understanding capabilities.

GPT-3 and GPT-4 OpenAI GPT-3 is an autoregressive language model that employs a Transformer model. Be aware GPT-3 is not a single model but a family of models that has different numbers of trainable hyperparameters and fine-tuning settings. Unlike BERT which is open sourced, GPT-3 is closed and black box. As the paper is being written, GPT-4 Turb has just been released which is claimed to be another significant leap on NLP. This review is just based on known information

and studies collected from public papers and experiments. We start reviewing BERT and its variants(a.k.a BERTology) as they are both based on the Transformer model and iterate to newer and larger models of GPT-3. We just focus on text processing only.

3. Knowledge and Reasoning Capabilities

3.1 Syntactic Knowledge

Syntactic representation from BERTology The paper A Primer in BERTology showed that syntactic information can be recovered from BERT token representation, even though it seems that syntactic structure is not directly encoded in self-attention weights [1]. Because BERT is based on a bidirectional encoder it is trained on both left-to-right and right-to-left sequences. Studies showed BERT representations are hierarchical rather than linear, that is, BERT model is akin to a syntactic tree structure in addition to word order information. This provides evidence that BERT "naturally" learns some syntactic information. However, as the study shows BERT couldn't "understand" negation and is insensitive to malformed input. It claimed BERT's predictions were not altered even with shuffled word order. This surprised me because the word order information had been encoded in the embed input indeed, and it must be reflected in the training output. The potential explanation is this could be due to training weights. Per the paper's analysis, this could mean that either BERT's syntactic knowledge is incomplete, or it does not need to rely on it for solving its tasks. There is no concrete answer yet but the latter seems more likely per report [2].

Attention Can Reflect Syntactic Structure The attention mechanism is an innovative part of Transformer architecture and essentially is a mapping in sequence-to-sequence between a query and a set of key-value pairs to an output.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

About syntactic structure, studied that the Transformer model with multiple head attentions mechanism allows it to jointly attend to information from different representations (features) [3]. It has been observed that individual dependency relations were often tracked by specialized heads. In this paper, experiments were conducted with a tree decoding test to show that the attention mechanism was learning to represent the structural objective of the parser. It's surprising that the transformer parameters, K and Q, were only modestly capable of resembling the dependency structure. What is more important is the Value (V) parameters, which play the most faithful representation of the linguistic structure via attention. The experiments in this paper focused on a linguistic structure that the attention-based model can learn and no test tasks were designed to explore semantic-orientation classification. Actually, this is an un-answered question which sets of transformer parameters are suited for learning such semantic information, or not at all? This leads us to study the next paper and the extent to which the transformer-based model, including BERTology, understands the natural language in terms of semantic aspects.

3.2 Semantic Knowledge

BERT's semantic knowledge The paper claims there is evidence that BERT has some knowledge of semantic roles. Eg. "to tip a chef" is better than "to tip a robin", but worse than "to tip a waiter" [1]. But BERT struggles with the representation of numbers because of wordpiece tokenization where similar values can be divided up into substantially different word chunks. BERT encodes information about entity types, relations, and semantic roles since this information can be detected with probing classifiers. However, study shows BERT struggles with representations of numbers. Floating point numbers, i.e. "2.09" can be divided into two chunks of words by the dot, "2" and "09". This breaks up the semantic meaning. Although BERT is "surprisingly" brittle to name entity replacements, it still did not absorb all the relevant entity information during pretraining. So the model couldn't build a generic idea of named entities. So there is no strong or complete evidence for BERT to show the full mastery of semantic knowledge.

The explanation behind this (in this study) is that BERT's self-attention heads do not directly encode any non-trivial linguistic information, basic syntactic information appears earlier in the network and high-level semantic features appear at the higher layers where training in higher layers is very expensive in BERT. Given the fact that BERT is computationally expensive, it is challenging to train high-level semantic understanding capability. GPT-3's semantic knowledge GPT-3 seems doing a better job with linguistic knowledge to identify certain semantic information in most cases, but still fails when there are some types of disturbance happening in the sentence. Per existing studies and experiments, GPT-3 doesn't possess Semantic Knowledge in the same way humans do, but it can generate responses that appear to understand the "meaning" of the input by recognizing patterns and associations in the data it was trained on [4].

3.3 World Knowledge or Commonsense Knowledge

The study shows BERT is lack of World knowledge [5]. It struggles with pragmatic inference, role-based event knowledge, and abstract attributes of objects that are likely to be assumed rather than mentioned. To answer questions like "Does the cake go in the oven?" which looks common sense to humans, BERT does have difficulty answering because of a lack of strong Contextualization.

Commonsense knowledge, an alias of world knowledge, requires context info to learn. In the paper "Cracking the Contextual Commonsense Code: Understanding Commonsense Reasoning Aptitude of Deep Contextual Representations" a method was developed through attribute classification in the semantic datasets and compared the contextual model to traditional word embedding [5]. The result outperforms word type embedding but still lacks some commonsense attributes - visual and perceptual properties. To mitigate this deficiency, a knowledge graph embedding was added in BERT features utilizing CSLB, a semantic norm

dataset [1]. Knowledge Graphs can help encode information that extends beyond BERT's embedding features. A classifier was also introduced to classify if an attribute applies to a candidate object, word, or sentence. It's found the F1 attribute score is much stronger - the median F1 score is nearly double that of GloVe baselines [1]. This means BERT encodes commonsense traits. However, this is not perfect. Some traits exhibit better than others. Specifically, physical traits such as "is made of wood" and "has a top" perform way better than those abstract traits such as "is creepy and is strong".

To answer the question of whether to use a camera flash, it would be thus related to the traits "does have flash", "is dark", and "is light", the model needs fine-tuning on additional data which is manually selected related to attributes that BERT is deficient in. The results (Table 1) show with

System	Accuracy
Human(Golden)	97.4
Random Baseline	48.9
BERT(LARGE)	82.3
with ConceptNet	83.1
with WebChild	82.7
with ATOMIC	82.5
with all KB	83.3
with all KB + RACE(selected)	85.5

Table 1: Test Set Results for Knowledge base Embedding's on MCScript 2.0 [5]

ConceptNet: An open, multilingual knowledge graph (https://conceptnet.io)
WebChild: Fine-grained commonsense knowledge distillation [6].
ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning [7].
RACE: Large-scale reading comprehension dataset from examinations [8].
MCSScript [9].

Explicit knowledge embeddings that each knowledge base improves accuracy, with ConceptNet giving the largest performance boost. ATOMIC gives the smallest boost, likely because the TOMIC edges involve longer phrases, which means fewer matches and the overlap between ATOMIC text and the text present in the task is not as large as either ConceptNet or WebChild [7]. As a result can tell that combining the knowledge base embeddings and the implicit RACE fine-tuning yields the highest accuracy. so fine-tuning is very critical in contextual knowledge learning.

BERTology's capability we have learned so far So given varied studies, BERT does possess a limited amount of syntactic, semantic, and world knowledge although some studies show some. It looks like it has built-in knowledge of syntactic structure due to its nature of encoding and embedding, but lacks strong semantic and world knowledge although some hypes claim to have. Further, BERT has limited reasoning abilities and its performance is heavily attributed to pattern recognition. The awkward situation is there is no single probing method that can reliably tell what extent the knowledge of the model possesses. A given method may favor one over another. This actually leads us to think about the definition of the "meaning" of language since the term "meaning" is so rich and multifaceted.

3.4 What is Meaning vs form

In this paper, defines what is meaning at first [10]. This is important to quantify what extent these LLMs understand natural language. It claims in varied terminologies, reports, and publications of LLMs there have been many misunderstandings of the relationship between linguistic form and meaning. Many claims in both academic and popular publications that claimed to "understand" natural language are ambiguous and misleading, such as "BERT is a system ... to better understand how human beings communicate..." "Here are some examples that ...demonstrate BERT's ability to understand the intent behind your search." It argues that "the language modeling task because it only uses form as training data, cannot in principle lead to learning of meaning". The form is just the observable realization of language, like the mark of page, pixels, or bytes of text binary. Linguistic form is the syntax representation of word sequence, like POS. Then what is the difference between linguistic form and meaning? This paper gave its answer: meaning is the relation between linguistic form and communicative intent.

$M = E \times I$

Which contains pairs (e, i) of natural language expression e and the communicative intents i they can evoke. Communicative intents are about something out of language. For example, when a teacher says "It is cold in the room", the intent behind the utterance is that "we should close the window" or "increase the heater temperature to make the room warmer." It claims that LLMs trained purely on form will not learn meaning because there is no sufficient signal to learn the relationship between the form and non-linguistic intent of human language users.

Octopus test Why meaning can't be learned from linguistic form alone? Because it lacks the ability to connect its utterances to the world. The Octopus test described in is designed to run experiments based on two isolated

Octopus, A, and B on two stranded islands, they can only communicate by a wire in the sea [10]. There is the third one O who can learn the communication between A and B. O is very good at detecting statistical patterns and learning and can predict with great accuracy how B will respond to each of A's utterances. However, this is working well until someday a new situation beyond the existing utterances happen. Dealing with new situations or new tasks requires the ability to map accurately between words and realworld entities as well as reasoning and creative thinking, which cannot be learned from statistics summary. When a run into an emergency on confronted with a bear never seen before and ask for help from B, the middle Octopus O who never had such experience has no idea how to deal with and respond.

Hype One hype for believing LM might be learning meaning is the claim that human children can acquire language just by listening to it. This is not true based on some studies. Actually, kids won't pick up a language from passive exposure such as TV or radio. The critical part of language learning is not just plain attention but also joint attention where interaction is important to boost the meaning of understanding. So the conclusion is that learning a linguistic system is like human learning. Communication relies on joint attention and intersubjectivity: the ability to be aware of what another human is attending to and guess/interact with what they intend to communicate. It cannot be learned by purely passive "learning", the key point is "interaction" between learner and teacher. Does BERTopogy learn meaning In conclusion of this paper, BERTopogy doesn't learn "meaning", it just learns some reflection of meaning in linguistic form.

Category	Poor scoring attributes(fit score <1.0)	Perfect scoring attributes (fit score = 1.0)
Visual	is triangular, is long	has a back, has a top
Perceptual	is wet,is rough, creepy and strong	does drive, does bend, live in river
Taxonomic	is a home, is a garden tool	is cat, is a body part

Table 2: Fine-Grained Comparison Across Categories between Attributes by BERT Representation

Per the above fine-grained comparison (Table 2) between attributes using BERT representations, overall BERT is strong enough to fit many features that would easily be represented in text such as *"does bend"*, "*does drive*", or *"does live in river"*, but still seems to have difficulty to fit those that most pertain to abstract common-sense, such as "*is hardy*" and *"has a strong smell"*. So this paper tells BERT shows a strong ability to encode various commonsense features in its embedding space, particularly those that are easily represented in text while facing challenges with abstract commonsense attributes [5].

3.5 Understanding Source Code

In the paper "The Larger they are, the Harder they Fail: Language Models do not Recognize Identifier Swaps in Python" studies have been conducted to probe the "memorization" hypothesis via counterfactual tasks [11]. The idea is to take a reasoning task that an LLM knows well and create a reversed or fake version of that task that requires abstract reasoning ability but is very less likely to show in the training dataset. As an example, we exchanged the Python built-in functions: len and print. And we ask LLM to generate a function to print out the length (Figure 1). The LLM gave the wrong answer in BERT and GPT-3 first-generation. (GPT-4 may be doing better, but is believed to have no fundamental change because of fundamentally the same model architecture.) All tested models always prefer the incorrect the output resulting in zero classification accuracy, the log-likelihood of the incorrect output is always significantly higher than the uniform baseline, but it varies with the model.

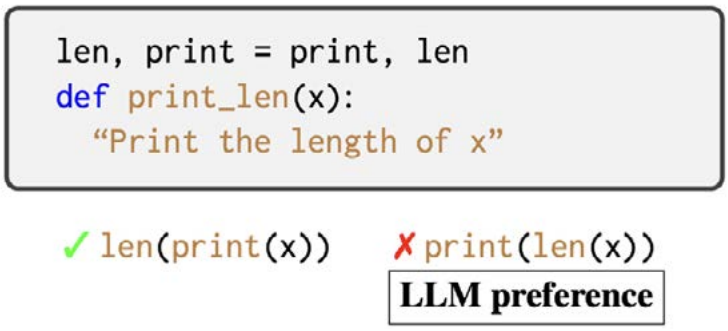

Figure 1: Given a Python Prompt (on top) which Swaps of two Built-in Functions, Large Language Models Prefer the Incorrect but Statistically Common Continuation (right) to the Correct but unusual one (left) [11]

3.6 Reasoning Capability

Few-shot reasoner While in-context learning with LLMs provides some degree of capacity for deep understanding and reasoning, there is always some limitation and LLMs are not very good at reasoning. However recent studies and experiments have shown the ability for reasoning can be substantially increased by making them produce step-by-step reasoning by few-shot prompting. Notably, a recent technology so-called chain of thought (CoT) prompting for eliciting complex multi-step reasoning through step-by-

step answer examples achieved significant performance boosts in multi-step arithmetic and logical reasoning [12]. The paper ("Few-shot learning") studies "Chain of thoughts" (CoT) prompting can attribute LLM to semantic and reasoning learning capability and empower LLMs to perform complex reasoning over text. As shown in Figure 2, in the experiments of web tables with CoT prompting, GPT-3 with CoT prompting was doing a very good job in reasoning and also provided high-quality explanations to justify their decision-making [12]. As shown in the GPT-3 experiments using various Table-based datasets (Davincitext-002), GPT-3 outperforms T5 and pipeline models, it is even closed human thought. By the "few-shot reasoning", we as humans provide the model with several exemplars of reasoning chains, which guide LLM toward the right track, so LLM can learn to follow the template to solve difficult unseen tasks [13].

This is more like teaching a kid to solve a complex problem when he/she is stuck and the teacher just gives the kid some hint and the kid figures out with some clue. A real-life example would be, let's ask an 8-year-old kid what the next number of the sequence 1,1,2,3,5,8... The kid may be stuck and have no idea. Once the teacher gave some hint, "Hey, can you find some pattern of the sum of each adjacent number pair?". Then the kid would suddenly realize this is just a Fibonacci sequence and the next number must be 13=5+8. CoT is the same thinking process with step-by-step guidance. A few-shot reasoner typically refers to a type of learning to perform reasoning tasks with only a few examples or shots of data. In GPT-3, few-shot reasoning involves providing the model with a prompt or a few examples of the desired behavior, and the model then generalizes from those examples to perform tasks or answer questions. In the paper the LLM is fed with several promptings to build more context as instructed so the LLM can iterate to answer long and complex questions [13]. The experiments were run in Table format of questions, which is a kind of semi-structured dataset but still needs context reasoning

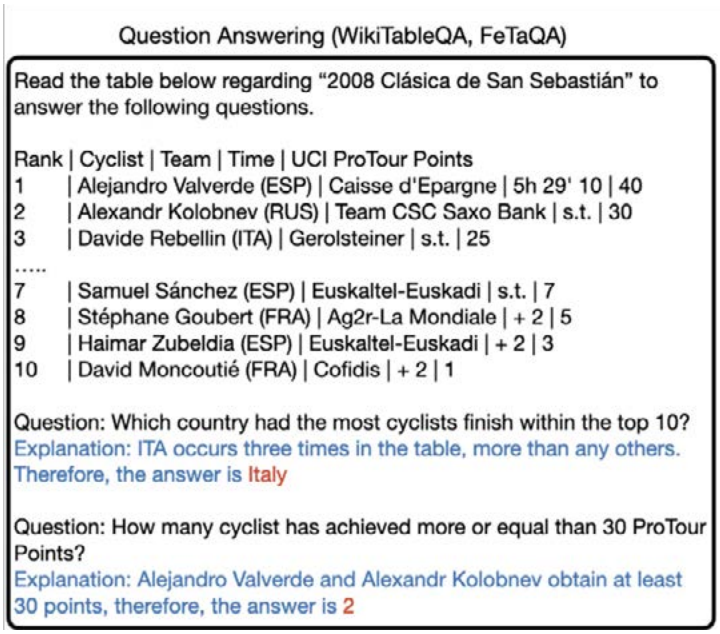

Figure 2: Question Answer by Few-Show Reasoner [13]

And deduction in natural language processing. The Chain of Thoughts (CoT) Reasoning is the key point to prompt the model step by step and so empower LLM to discover more context it may already have with more self-conciseness. For instance, you could provide a few examples of how you want the model to answer questions about a specific topic, and the few-shot reasoner would use that information to generate responses to new, similar queries.

Advance further: Zero-shot To advance further, this still needs one or more shots in terms of prompting examples. Here raise a question, can we do better? Even without shots but still empowering the model to reason itself upfront or internally without user intervention?

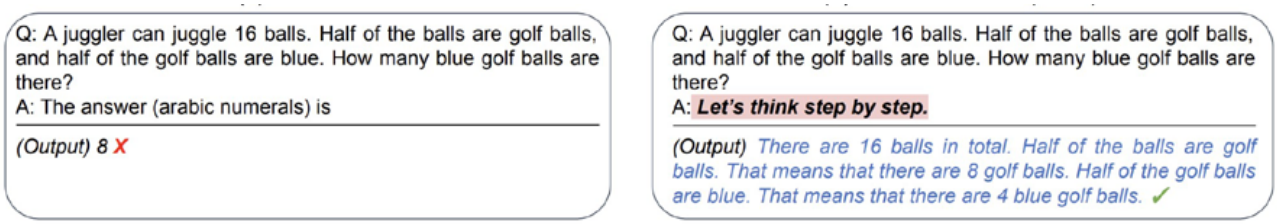

Figure 3: Left is Standard Zero-Shot and Right Zero-Shot-CoT [14]

Another paper "Large Language Models are zero-shot reasoners" shows zero-shot-CoT prompt examples that demonstrate good reasoning capability [14]. Zero-shot reasoning refers to the ability of LLMs to perform multi-step reasoning tasks on unseen domains without any hand-crafted examples. It enables them to generalize knowledge from their training data and apply it to new, unseen situations. The idea behind this is to trigger LLMs by simply adding a "Let's think step by step" prompt to generate a reasoning path in the LLM's background processing that decomposes a complex problem into two or more "simpler" and breaks it down into sub-problems. This looks very simple, and actually, I think the key point behind this is we teach the model to explore a reasoning path that decomposes the complex reasoning into multiple simpler steps. This style of "Chain of thought prompting" demonstrated good performance in arithmetic and logical reasoning.

4. Limitation and Generalization

Even with zero-shot and few-shot reasoning, which are really prompting techniques, to help to unleash the potentials of LLM including GPT-3, it is still a question what is the boundary and limit of LLM in reasoning. The paper was trying to answer this question [15]. It measured the limitation of transformers in compositional tasks with 3 representative compositional tasks: long-form multiplication, logic grid puzzle, and a classic dynamic programming problem. These experiments suggest that if an output element heavily relies on a single or small set of input features, transformers are likely to recognize such correlation during training and directly map these input features to predicate the output element in testing without going through rigorous multi-hop reasoning. The paper hypothesizes that beyond simple memorization, transformers largely rely on pattern matching for solving these tasks. This is contradictive to the prior paper claiming "that our results cannot be explained solely by direct memorization."

This poses an open question of how the LLM reasoning works. The difference in experiment observations may come from varied datasets or task domain settings. The result heavily relies on the dataset size and may scale up and down as the dataset scales. LLM reasoning exhibits unpredictable randomness and cannot generalize to large or varied categories of datasets. How to evaluate the quality of LLM models specifically for reasoning impacts the test accuracy and reliability. Many studies have been done to investigate the generalization capabilities. The paper demonstrates how pattern matching can even hinder generalization [15]. We are still hyperthesis that those popular LLMs based on transformers, BERT/GPT-3, still have challenges to fully master the semantics and reasoning for these complex tasks even with various "zero- or few-shots" prompting techniques. This is still an open and challenging area to be conquered in the iteration of LLMs in the future.

5. Discussion

GPT-4 GPT-4 self-corrected itself in the middle of writing his answer if you told it's wrong. This could be prompted by human feedback to guide the model to choose another path or choose a secondary good answer as a backup. Considering the earlier section few-shot and zero-shot reasoning, it is a topic to empower the model itself to do self-reasoning and fact-check before replying. Harmful information LLM may generate instructions for dangerous or potentially harmful or illegal activities. The LLM may not tell the difference between bad and good. Actually, it is even arguable for humans to reliable to distinguish without full knowledge. This is still an open big question of how to improve the robustness and safety of language models.

6. Conclusion

We started from BERTology and GPT-3 3, studied the capability from syntax knowledge 4.1, world knowledge 4.3, Semantic Knowledge 4.2, and contextual information. The surface knowledge including syntax could be easier to retrieve from statistical patterns and attention mechanisms of transformer-based models. We also learned the difference between linguistic form and semantic meaning 4.4. It is not that obvious and even challenging to learn Semantic knowledge and reasoning capability. Although some recent studies are showing that few-shot and zero-shot reasoning by Chain-of-Thought prompting can empower LLMs with stronger reasoning capability 4.6 to break down complex problems, there is an open question of how to fully understand natural language like Human does.

References

1. Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics,* 8, 842-866.
2. Glavaš, G., & Vulić, I. (2020). Is supervised syntactic parsing beneficial for language understanding? an empirical investigation. *arXiv preprint arXiv*:2008.06788.
3. Ravishankar, V., Kulmizev, A., Abdou, M., Søgaard, A., & Nivre, J. (2021). Attention can reflect syntactic structure (if you let it). *arXiv preprint arXiv:*2101.10927.
4. Zhang, L., Wang, M., Chen, L., & Zhang, W. (2022, December). Probing GPT-3's linguistic knowledge on semantic tasks. *In Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 297-304).
5. Da, J., & Kasai, J. (2019). Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. *arXiv preprint arXiv*:1910.01157.
6. Tandon, N., De Melo, G., & Weikum, G. (2017, July). Webchild 2.0: Fine-grained commonsense knowledge distillation. *In Proceedings of ACL 2017, System Demonstrations* (pp. 115-120).
7. Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., ... & Choi, Y. (2019, July). Atomic: An atlas of machine commonsense for if-then reasoning. *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 3027-3035).
8. Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv*:1704.04683.
9. Ostermann, S., Modi, A., Roth, M., Thater, S., & Pinkal, M.

(2018). Mcscript: A novel dataset for assessing machine comprehension using script knowledge. *arXiv preprint arXiv*:1803.05223.

10. Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. *In Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185-5198).
11. Miceli-Barone, A. V., Barez, F., Konstas, I., & Cohen, S. B. (2023). The larger they are, the harder they fail: Language models do not recognize identifier swaps in python. *arXiv preprint arXiv*:2305.15507.
12. Wu, D., Zhang, J., & Huang, X. (2023). Chain of thought prompting elicits knowledge augmentation. *arXiv preprint arXiv*:2307.01640.
13. Chen, W. (2022). Large language models are few (1)-shot table reasoners. *arXiv preprint arXiv*:2210.06710.
14. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems, 35,* 22199-22213.
15. Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., ... & Choi, Y. (2024). Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems, 36.*