Towards Reasoning in Large Language Models: A Survey

Jie Huang **Kevin Chen-Chuan Chang**
Department of Computer Science, University of Illinois at Urbana-Champaign
{jeffhj, kcchang}@illinois.edu

Abstract

Reasoning is a fundamental aspect of human intelligence that plays a crucial role in activities such as problem solving, decision making, and critical thinking. In recent years, large language models (LLMs) have made significant progress in natural language processing, and there is observation that these models may exhibit reasoning abilities when they are sufficiently large. However, it is not yet clear to what extent LLMs are capable of reasoning. This paper provides a comprehensive overview of the current state of knowledge on reasoning in LLMs, including techniques for improving and eliciting reasoning in these models, methods and benchmarks for evaluating reasoning abilities, findings and implications of previous research in this field, and suggestions on future directions. Our aim is to provide a detailed and up-to-date review of this topic and stimulate meaningful discussion and future work.[1]

1 Introduction

Reasoning is a cognitive process that involves using evidence, arguments, and logic to arrive at conclusions or make judgments. It plays a central role in many intellectual activities, such as problem solving, decision making, and critical thinking. The study of reasoning is important in fields like psychology (Wason and Johnson-Laird, 1972), philosophy (Passmore, 1961), and computer science (Huth and Ryan, 2004), as it helps individuals make decisions, solve problems, and think critically.

Recently, large language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Chung et al., 2022; OpenAI, 2022, *inter alia*) such as ChatGPT have made significant advancements in natural language processing and related fields. It has been shown that these models exhibit emergent behaviors, including the ability to "reason", when they are large enough (Wei et al., 2022a). For example, by providing the models with "*chain of thoughts*", i.e., reasoning exemplars, or a simple prompt "*Let's think step by step*", these models are able to answer questions with explicit reasoning steps (Wei et al., 2022b; Kojima et al., 2022), e.g., "all whales are mammals, all mammals have kidneys; therefore, all whales have kidneys." This has sparked considerable interest in the community since reasoning ability is a hallmark of human intelligence that is frequently considered missed in current artificial intelligence systems (Marcus, 2020; Russin et al., 2020; Mitchell, 2021; Bommasani et al., 2021).

However, despite the strong performance of LLMs on certain reasoning tasks, it remains unclear whether LLMs are actually reasoning and to what extent they are capable of reasoning. For example, Kojima et al. (2022) claim that "LLMs are decent zero-shot reasoners (p. 1)", while Valmeekam et al. (2022) conclude that "LLMs are still far from achieving acceptable performance on common planning/reasoning tasks which pose no issues for humans to do (p. 2)." This limitation is also stated by Wei et al. (2022b):

> "we qualify that although chain of thought emulates the thought processes of human reasoners, this does not answer whether the neural network is actually *reasoning* (p. 9)."

Therefore, in this paper, we aim to provide a comprehensive overview and engage in an insightful discussion on the current state of knowledge on this fast-evolving topic. We initiate our exploration with a clarification of the concept of reasoning (§2). Subsequently, we turn our attention to the techniques for enhancing/eliciting reasoning in LLMs (§3), the methods and benchmarks for evaluating reasoning in LLMs (§4), and the key findings and implications in this field (§5). Finally, we reflect on and discuss the current state of the field (§6).

[1]Paperlist can be found at https://github.com/jeffhj/LM-reasoning.

Findings of the Association for Computational Linguistics: ACL 2023, pages 1049–1065
July 9-14, 2023

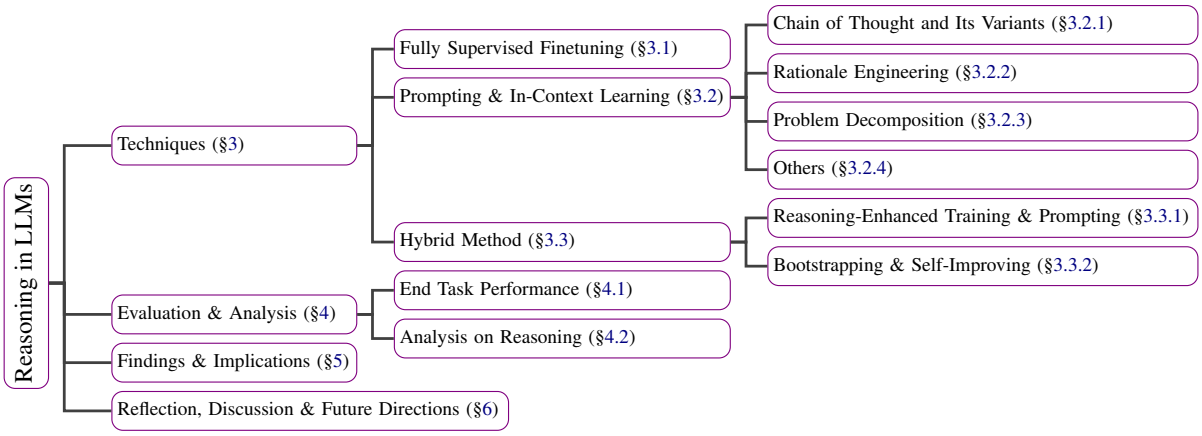

Figure 1: The structure of the paper.

2 What is Reasoning?

Reasoning is the process of thinking about something in a logical and systematic way, using evidence and past experiences to reach a conclusion or make a decision (Wason and Johnson-Laird, 1972; Wason, 1968; Galotti, 1989; Fagin et al., 2004; McHugh and Way, 2018). Reasoning involves making inferences, evaluating arguments, and drawing logical conclusions based on available information. Although "reasoning" is a term that is commonly used in literature and daily life, it is also an abstract concept that can refer to many things. To help the reader better understand this concept, we summarize several main categories of reasoning that are commonly recognized:

Deductive reasoning. Deductive reasoning is a type of reasoning in which a conclusion is drawn based on the truth of the premises. In deductive reasoning, the conclusion must necessarily follow from the premises, meaning that if the premises are true, the conclusion must also be true. For example:

- Premise: All mammals have kidneys.
- Premise: All whales are mammals.
- Conclusion: All whales have kidneys.

Inductive reasoning. Inductive reasoning is a type of reasoning in which a conclusion is drawn based on observations or evidence. The conclusion is likely to be true based on the available evidence, but it is not necessarily certain. For example:

- Observation: Every time we see a creature with wings, it is a bird.
- Observation: We see a creature with wings.
- Conclusion: The creature is likely to be a bird.

Abductive reasoning. Abductive reasoning is a type of reasoning in which a conclusion is drawn based on the best explanation for a given set of observations. The conclusion is the most likely explanation based on the available evidence, but it is not necessarily certain. For example:

- Observation: The car cannot start and there is a puddle of liquid under the engine.
- Conclusion: The most likely explanation is that the car has a leak in the radiator.

Other types of reasoning include *analogical reasoning*, which involves making comparisons between two or more things in order to make inferences or arrive at conclusions; *causal reasoning*, which involves identifying and understanding the causes and effects of events or phenomena; and *probabilistic reasoning*, which involves making decisions or arriving at conclusions based on the likelihood or probability of certain outcomes.

Formal Reasoning vs Informal Reasoning. *Formal reasoning* is a systematic and logical process that follows a set of rules and principles, often used in mathematics and logic. *Informal reasoning* is a less structured approach that relies on intuition, experience, and common sense to draw conclusions and solve problems, and is often used in everyday life. Formal reasoning is more structured and reliable, while informal reasoning is more adaptable and open-ended, but may also be less reliable. We refer the reader to Galotti (1989); Bronkhorst et al. (2020) for a detailed distinction between them.

Reasoning in Language Models. The concept of reasoning in language models has been around for some time, but there is not a clear definition of what it entails. In the literature, the term "reasoning" is often used to refer to informal reasoning, although it is not always explicitly stated that it is informal (Cobbe et al., 2021; Wei et al., 2022b,

inter alia). Different forms of reasoning may be used depending on the task, benchmark, or method being used, e.g., deductive reasoning (Cobbe et al., 2021; Creswell et al., 2022; Han et al., 2022b, *inter alia*), inductive reasoning (Yang et al., 2022; Misra et al., 2022, *inter alia*) or abductive reasoning (Wiegreffe et al., 2022; Lampinen et al., 2022; Jung et al., 2022, *inter alia*). In this paper, we encompass various forms of reasoning, with a particular focus on "informal deductive reasoning" in large language models since it is a widely used form in which the conclusion is guaranteed to be true as long as the premises are true.

3 Towards Reasoning in Large Language Models

Reasoning, particularly multi-step reasoning, is often seen as a weakness in language models and other NLP models (Bommasani et al., 2021; Rae et al., 2021; Valmeekam et al., 2022). Recent research has suggested that reasoning ability may emerge in language models at a certain scale, such as models with over 100 billion parameters (Wei et al., 2022a,b; Cobbe et al., 2021). In this paper, we follow Wei et al. (2022a) in considering reasoning as an ability that is rarely present in small-scale models like GPT-2 (Radford et al., 2019) and BERT (Devlin et al., 2019), and therefore focus on techniques applicable to improving or eliciting "reasoning"[2] in LLMs such as GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022).

3.1 Fully Supervised Finetuning

Before discussing reasoning in large language models, it is worth mentioning there is research working on eliciting/improving reasoning in small language models through *fully supervised finetuning* on specific datasets. For example, Rajani et al. (2019) finetune a pretrained GPT model (Radford et al., 2018) to generate rationales that explain model predictions with the built CoS-E dataset, and find that models trained with explanations perform better on commonsense question answering tasks (Talmor et al., 2019). Talmor et al. (2020) train RoBERTa (Liu et al., 2019) to perform reasoning/inference based on both implicit pre-trained knowledge and explicit free-text statements. Hendrycks et al. (2021) finetune pretrained language models to solve competition mathematics problems by generating full step-by-step solutions, though the accuracy is relatively low. Nye et al. (2022) train language models to do multi-step reasoning for program synthesis/execution by generating "scratchpads", i.e., intermediate computations, before producing the final answers. We refer the reader to Helwe et al. (2021); Bhargava and Ng (2022)'s survey for more studies in this line.

There are two major limitations of fully supervised finetuning. First, it requires a dataset containing explicit reasoning, which can be difficult and time-consuming to create. Additionally, the model is only trained on a specific dataset, which limits its application to a specific domain and may result in the model relying on artifacts in the training data rather than actual reasoning to make predictions.

3.2 Prompting & In-Context Learning

Large language models such as GPT-3 (Brown et al., 2020) have demonstrated remarkable few-shot performance across a variety of tasks through in-context learning. These models can be prompted with a question and a few ⟨input, output⟩ exemplars to potentially solve a problem through "reasoning", either implicitly or explicitly. However, research has shown that these models still fall short when it comes to tasks that require multiple steps of reasoning to solve (Bommasani et al., 2021; Rae et al., 2021; Valmeekam et al., 2022). This may be due to a lack of exploration into the full capabilities of these models, as recent studies have suggested.

3.2.1 Chain of Thought and Its Variants

To encourage LLMs to engage in reasoning rather than simply providing answers directly, we may guide LLMs to generate "reasoning" explicitly. One approach for doing this is *chain-of-thought prompting*, proposed by Wei et al. (2022b). This approach involves providing a few examples of "chain of thought" (CoT), which are intermediate natural language reasoning steps, in the prompt to LLMs (Figure 2). Specifically, in CoT prompting, ⟨input, output⟩ demonstrations are replaced with ⟨input, *chain of thought*, output⟩ triples, e.g., "[input] Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? [*chain of thought*] Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. [output] The answer is 11." In this way, given a target question, the model learns to generate explicit ratio-

[2]It is important to note that the term "reasoning" in this paper does not necessarily imply that LLMs are truly capable of reasoning or that they are able to reason in the same way that humans do. We will discuss this issue in more detail in §6.

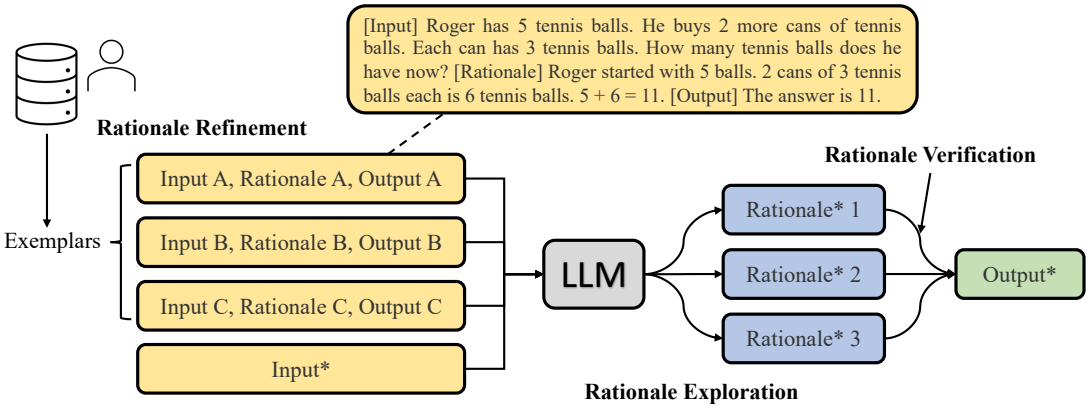

Figure 2: An illustration of *Chain-of-Thought Prompting* and *Rationale Engineering*, where asterisk (*) denotes the target problem to be solved.

nale before producing the final answer. Experimental results show that this simple idea can improve LLMs' few-shot performance on arithmetic, symbolic, and commonsense reasoning tasks, sometimes to a striking degree.

There are several variants of chain-of-thought prompting that have been proposed in the literature, in a different form or to solve a specific problem.

Different Form: Kojima et al. (2022) introduce *Zero-shot-CoT*, in which LLMs are simply prompted with the phrase "Let's think step by step" after the input, in order to elicit reasoning without the need for few-shot demonstrations. Madaan et al. (2022); Gao et al. (2022); Chen et al. (2022) find that LLMs trained with code, e.g., Codex (Chen et al., 2021), can achieve better performance on reasoning tasks by framing reasoning as code generation. Wang et al. (2022a) propose to iteratively prompt chain of thought. He et al. (2023) attempt to retrieve external knowledge in CoT to improve faithfulness of reasoning.

Specific Problem/Setting: Before chain of thought, Nye et al. (2022) also try to use intermediate computations, named "scratchpads", to improve language models' reasoning performance in both finetuning and few-shot regimes, with a particular focus on programs. Shi et al. (2022) attempt to solve multilingual reasoning tasks with CoT in the native language, CoT in English (regardless of the problem language), and CoT in English (with the problem translated to English). Chen (2022) apply CoT to table-based reasoning, finding that LLMs can achieve strong performance on table tasks with only one exemplar. Prystawski et al. (2022) demonstrate that CoT can improve LLMs' performance on paraphrase selection for metaphors. Lu et al. (2022) apply chain of thought to solve multimodal science questions.

3.2.2 Rationale Engineering

The original version of chain-of-thought prompting, proposed by Wei et al. (2022b), relies on manually crafted examples of intermediate reasoning steps and applies greedy decoding in the generation. *Rationale engineering* aims to more effectively elicit or utilize reasoning in LLMs. This can be achieved through *rationale refinement*, which involves creating more effective examples of reasoning steps, or through *rationale exploration* and *rationale verification*, which involve exploring and verifying the rationales produced by LLMs. A summary of raltionale engineering is illustrated in Figure 2.

Rationale refinement. The choice of exemplars can significantly affect the few-shot performance of LLMs, as demonstrated in research such as Liu et al. (2022b), which also appears in chain-of-thought prompting. *Rationale refinement* aims to create and refine rationale examples that are better able to elicit reasoning in LLMs. Fu et al. (2022b) propose *complexity-based prompting* to create rationales with more reasoning steps. Their experiments show that the performance of LLMs improves with the increased rationale complexity. Similarly, Zhou et al. (2022c) propose *algorithmic prompting*, which suggests that providing more thorough examples of solutions can help improve reasoning performance on some simple math calculations. Zhang et al. (2022b) design *Auto-CoT* to automatically construct exemplars by partitioning questions from a given dataset into clusters and then using Zero-Shot-CoT (Kojima et al., 2022) to generate the rationale for a representative question from each

cluster. The analysis shows that making exemplars diverse is important in prompting LLMs to produce better rationales.

Rationale exploration. In addition to providing better exemplars, we can allow LLMs to fully explore various ways of reasoning to improve their performance on reasoning tasks, named *rationale exploration*. Based on the idea that complex problems often admit multiple ways of thinking that can lead to their unique correct answer, Wang et al. (2022c) present a decoding strategy called *self-consistency* to improve upon the traditional greedy decoding used in chain-of-thought prompting. This strategy involves sampling a diverse set of rationales, rather than just the greedy one, and selecting the most consistent answer by marginalizing out the sampled rationales. The idea is also used in Fu et al. (2022b) to vote over the top complex rationales. To further improve performance, Li et al. (2022b) suggest providing different demonstrations for each question by sampling exemplars from an exemplar base, in order to increase the diversity of the sampled rationales.

Rationale verification. Ensuring that the rationales produced by LLMs are valid is critical, as incorrect rationales can lead to incorrect final predictions (Ye and Durrett, 2022). To address this issue, the process of *rationale verification* aims to verify whether the rationales produced by LLMs lead to the correct final answers. Cobbe et al. (2021) propose augmenting LLMs with a trained verifier that assigns a score to each rationale and solution generated by the LLM, selecting the highest-ranked solution as the final answer when solving math word problems. Li et al. (2022b) also use this technique to guide rationale selection, in conjunction with the process of rationale exploration. Different from the above methods that train an external verifier to verify the rationales, Weng et al. (2022) suggest using LLMs themselves as the verifiers.

3.2.3 Problem Decomposition

Chain-of-thought prompting, while effective for eliciting reasoning in LLMs, can struggle with complex tasks, e.g., tasks that require compositional generalization (Lake and Baroni, 2018; Keysers et al., 2020). To solve a complex problem, it is helpful to first break it down into smaller, more manageable subproblems. By solving each of these subproblems, we can effectively solve the complex problem. This technique is called *problem decomposition* or *divide and conquer* (Talmor and Berant, 2018; Min et al., 2019; Perez et al., 2020).

Based on this idea, Zhou et al. (2022a) propose *least-to-most prompting*, which consists of two steps: decomposing the complex problem into subproblems and solving these subproblems in a specific order, with each subproblem being facilitated by the answers obtained from previously solved subproblems. As follow-up work, Drozdov et al. (2022) introduce *dynamic least-to-most prompting*, which is designed to solve more realistic semantic parsing problems by decomposing the problems with prompting-based syntactic parsing and dynamically selecting exemplars based on the decomposition. In addition, Khot et al. (2022) design *decomposed prompting*, which breaks down a complex problem into subproblems that can be handled by a shared library of prompting-based LLMs, each specialized in a particular subproblem. Furthermore, Dua et al. (2022) develop *successive prompting*, which iteratively decomposes a complex problem into a simple problem, with the next subproblem prediction having access to the answers to the previous subproblems. While the above methods decompose or solve compositional questions with multiple forward passes, Press et al. (2022) suggest decomposing and solving the input question in one forward pass using CoT prompting. Overall, these techniques show promise for helping LLMs to solve complex tasks by decomposing the problem into more manageable subproblems.

3.2.4 Others

There are other techniques that have been developed to facilitate reasoning in LLMs for specific tasks or settings. For instance, Creswell et al. (2022); Creswell and Shanahan (2022) introduce a *selection-inference* framework that uses LLMs as modules to select and infer reasoning steps from a set of facts that culminate in the final answer. Kazemi et al. (2022) suggest using backward chaining, i.e., from goal to the set of facts that support it, instead of forward chaining like Creswell et al. (2022); Creswell and Shanahan (2022). In addition, Jung et al. (2022) propose a method for solving binary questions by prompting LLMs abductively and recursively to rationalize each option. Zhou et al. (2022b) design a technique for performing numerical reasoning on complex numbers by replacing the complex numbers with simple numbers to produce simpler expressions, and then using these expressions to perform calculations on the

complex numbers. There are also efforts to distill reasoning from LLMs into smaller models, such as the work by Li et al. (2022a); Shridhar et al. (2022); Magister et al. (2022). Finally, we refer the reader to Dohan et al. (2022)'s position paper on *language model cascade*, which presents a unifying framework for understanding chain-of-thought prompting and research in this line.

3.3 Hybrid Method

While "prompting" techniques can help elicit or better utilize reasoning in large language models to solve reasoning tasks, they do not actually improve the reasoning capabilities of the LLMs themselves, as the parameters of the models remain unchanged. In contrast, the "hybrid approach" aims to simultaneously improve the reasoning capabilities of LLMs and make better use of these models in order to solve complex problems. This approach involves both enhancing the reasoning capabilities of the LLMs and using techniques such as prompting to effectively utilize these capabilities.

3.3.1 Reasoning-Enhanced Training and Prompting

One approach to improving the reasoning capabilities of LLMs is to pretrain or finetune the models on datasets that include "reasoning". Lewkowycz et al. (2022); Taylor et al. (2022) find that LLMs trained on datasets containing scientific and mathematical data can achieve better performance on reasoning tasks like quantitative reasoning problems when using CoT prompting[3]. Pi et al. (2022) show that continually pretraining with SQL data can boost the performance of language models, e.g., T5 (Raffel et al., 2020), on natural language reasoning such as numerical reasoning and logical reasoning. Furthermore, Chung et al. (2022) develop Flan models by finetuning PaLM (Chowdhery et al., 2022) and T5 (Raffel et al., 2020) with 1.8k finetuning tasks, including CoT data, and find that CoT data are critical to keeping reasoning abilities. Similarly, Yu et al. (2022) finetune OPT (Zhang et al., 2022a) on 10 reasoning datasets and observe that it can improve some reasoning capabilities of LLMs. Anil et al. (2022) study the length generalization abilities of LLMs, i.e., whether LLMs learned with short problem instances can generalize to long ones. They discover that the combination of few-shot scratchpad (or chain of thought) finetuning and scratchpad prompting results in a significant improvement in LLMs' ability to generalize to longer problems, while this phenomenon is not observed in the standard fully supervised finetuning paradigm.

[3]This may also be true for models trained with code (Chen et al., 2021; Fu et al., 2022a).

3.3.2 Bootstrapping & Self-Improving

Instead of finetuning LLMs on pre-built datasets that include reasoning, there are studies that have explored the idea of using LLMs to self-improve their reasoning abilities through a process known as bootstrapping. One example of this is the *Self-Taught Reasoner (STaR)* introduced by Zelikman et al. (2022), in which a LLM is trained and refined on its own output iteratively. Specifically, with CoT prompting, the model first generates initial rationales. And then, the model is finetuned on rationales that lead to correct answers. This process can be repeated, with each iteration resulting in an improved model that can generate better training data, which in turn leads to further improvements. As a follow-up to this work, Huang et al. (2022a) show that LLMs are able to self-improve their reasoning abilities without the need for supervised data by leveraging the self-consistency of reasoning (Wang et al., 2022c).

4 Measuring Reasoning in Large Language Models

We summarize methods and benchmarks for evaluating reasoning abilities of LLMs in this section.

4.1 End Task Performance

One way to measure reasoning abilities of LLMs is to report their performance, e.g., accuracy, on end tasks that require reasoning. We list some common benchmarks as follows.

Arithmetic Reasoning. *Arithmetic reasoning* is the ability to understand and apply mathematical concepts and principles in order to solve problems involving arithmetic operations. This involves using logical thinking and mathematical principles to determine the correct course of action when solving mathematical problems. Representative benchmarks for arithmetic reasoning include GSM8K (Cobbe et al., 2021), Math (Hendrycks et al., 2021), MathQA (Amini et al., 2019), SVAMP (Patel et al., 2021), ASDiv (Miao et al., 2020), AQuA (Ling et al., 2017), and MAWPS (Roy and Roth, 2015). It is worth

mentioning that Anil et al. (2022) generate the *Parity Datasets* and the *Boolean Variable Assignment Dataset* for analyzing the length generalization capabilities of LLMs (§3.3.1).

Commonsense Reasoning. *Commonsense Reasoning* is the use of everyday knowledge and understanding to make judgments and predictions about new situations. It is a fundamental aspect of human intelligence that enables us to navigate our environment, understand others, and make decisions with incomplete information. Benchmarks that can be used for testing commonsense reasoning abilities of LLMs include CSQA (Talmor et al., 2019), StrategyQA (Geva et al., 2021), and ARC (Clark et al., 2018). We refer the reader to Bhargava and Ng (2022)'s survey for more work in this domain.

Symbolic Reasoning. *Symbolic reasoning* is a form of reasoning that involves the manipulation of symbols according to formal rules. In symbolic reasoning, we use abstract symbols to represent concepts and relationships, and then manipulate those symbols according to precise rules in order to draw conclusions or solve problems. Two benchmarks of symbolic reasoning are presented in Wei et al. (2022b), including Last Letter Concatenation and Coin Flip.

Others. In practice, there are many benchmarks that can be used to evaluate reasoning abilities of LLMs (indirectly), as long as the downstream task involves reasoning. BIG-bench (Srivastava et al., 2022), for example, includes over 200 tasks that test a range of reasoning skills, including tasks like Date Understanding, Word Sorting, and Causal Judgement. Other benchmarks, such as SCAN (Lake and Baroni, 2018) and the one proposed by Anil et al. (2022), focus on evaluating generalization ability. LLMs can also be tested on their table reasoning abilities using benchmarks such as WikiTableQA (Pasupat and Liang, 2015), FetaQA (Nan et al., 2022), as suggested by Chen (2022). In addition, there are benchmarks for evaluating LLMs' generative relational reasoning abilities, such as CommonGen (Lin et al., 2020; Liu et al., 2022a) and Open Relation Modeling (Huang et al., 2022b,d).

4.2 Analysis on Reasoning

Although LLMs have demonstrated impressive performance on various reasoning tasks, the extent to which their predictions are based on true reasoning or simple heuristics is not always clear. This is because most existing evaluations focus on their accuracy on end tasks, rather than directly assessing their reasoning steps. While some error analysis has been conducted on the generated rationales of LLMs (Wei et al., 2022b; Kojima et al., 2022, *inter alia*), this analysis has often been limited in depth.

There have been some efforts to develop metrics and benchmarks that enable a more formal/deep analysis of reasoning in LLMs. Golovneva et al. (2022) design ROSCOE, a set of interpretable, detailed step-by-step evaluation metrics covering various perspectives including semantic alignment, logical inference, semantic similarity, and language coherence. Saparov and He (2022) create a synthetic dataset called PrOntoQA that is generated from real or fictional ontologies. Each example in the dataset has a unique proof, which can be converted to simple sentences and back again, allowing for a formal analysis of each reasoning step. Han et al. (2022a) introduce a dataset called FOLIO to test the first-order logic reasoning capabilities of LLMs. FOLIO contains first-order logic reasoning problems that require models to determine the correctness of conclusions given a set of premises. In addition, Wang et al. (2022b) conduct ablation experiments on CoT and find that LLMs may also perform reasoning while prompting with invalid rationals. Their study also suggests that being relevant to the query and correctly ordering the reasoning steps are important for CoT prompting.

In summary, most existing studies primarily report the performance of the models on downstream reasoning tasks, without a detailed examination of the quality of the rationales produced. This leaves open the question of whether the models are actually able to reason in a way that is similar to human reasoning, or whether they are simply able to achieve good performance on the tasks through other means. Further research is needed to more formally analyze the reasoning abilities of LLMs.

5 Findings and Implications

In this section, we summarize the important findings and implications of studies on reasoning in large language models.

Reasoning seems an emergent ability of LLMs. Wei et al. (2022a,b); Suzgun et al. (2022) show that reasoning ability appears to emerge only in large language models like GPT-3 175B, as evidenced by significant improvements in performance on reasoning tasks at a certain scale (e.g., 100 billion

parameters). This suggests that it may be more effective to utilize large models for general reasoning problems rather than training small models for specific tasks. However, the reason for this emergent ability is not yet fully understood. We refer the reader to Wei et al. (2022a); Fu et al. (2022a) for some potential explanations.

Chain of thought elicits "reasoning" of LLMs. The use of chain-of-thought (CoT) prompts (Wei et al., 2022b) has been shown to improve the performance of LLMs on various reasoning tasks, as demonstrated in the experiments of Wei et al. (2022a,b); Suzgun et al. (2022). Additionally, Saparov and He (2022) (§4.2) find that, when using CoT prompts, LLMs are able to produce valid individual proof steps, even when the synthetic ontology is fictional or counterfactual. However, they may sometimes choose the wrong steps when multiple options are available, leading to incomplete or incorrect proofs. Moreover, for many reasoning tasks where the performance of standard prompting grows smoothly with model scale, chain-of-thought prompting can lead to dramatic performance improvement. In addition to these benefits, the use of CoT prompts has been shown to improve the out-of-distribution robustness of LLMs (Wei et al., 2022b; Zhou et al., 2022a; Anil et al., 2022, *inter alia*), an advantage that is not typically observed with standard prompting or fully supervised finetuning paradigms.

LLMs show human-like content effects on reasoning. According to Dasgupta et al. (2022), LLMs exhibit reasoning patterns that are similar to those of humans as described in the cognitive literature. For example, the models' predictions are influenced by both prior knowledge and abstract reasoning, and their judgments of logical validity are impacted by the believability of the conclusions. These findings suggest that, although language models may not always perform well on reasoning tasks, their failures often occur in situations that are challenging for humans as well. This provides some evidence that language models may "reason" in a way that is similar to human reasoning.

LLMs are still unskilled at complex reasoning. Although LLMs seem to possess impressive reasoning capabilities with the techniques described in §3, they still struggle with more complex reasoning tasks or those involving implicature, according to studies such as Valmeekam et al. (2022); Han et al. (2022a); Ruis et al. (2022). For instance, Valmeekam et al. (2022) find that even in relatively simple commonsense planning domains that humans would have no trouble navigating, LLMs such as GPT-3 (Brown et al., 2020) and BLOOM (Scao et al., 2022) struggle to perform effectively. These findings suggest that existing benchmarks may be too simple to accurately gauge the true reasoning abilities of LLMs, and that more challenging tasks may be needed to fully evaluate their abilities in this regard.

6 Reflection, Discussion, and Future Directions

Why reasoning? Reasoning is the process of thinking about something in a logical and systematic way, and it is a key aspect of human intelligence. By incorporating reasoning capabilities into language models, we can enable them to perform tasks that require more complex and nuanced thinking, such as problem solving, decision making, and planning (Huang et al., 2022e,f; Song et al., 2022). This can improve the performance of these models on downstream tasks and increase their out-of-distribution robustness (Wei et al., 2022a,b; Suzgun et al., 2022; Zhou et al., 2022a; Anil et al., 2022). In addition, reasoning can make language models more explainable and interpretable, as it provides explicit rationales for their predictions.

Right task/application? As Valmeekam et al. (2022) point out, current benchmarks may not adequately reflect the reasoning capabilities of LLMs. In addition, tasks such as solving simple math problems and concatenating letters in strings (§4.1) are artificial and do not accurately reflect real-world situations. To truly understand the reasoning ability of LLMs, it is important to consider more realistic and meaningful applications such as decision making (Edwards, 1954), legal reasoning (Levi, 2013), and scientific reasoning (Zimmerman, 2000). Our ultimate goal should not be to enable LLMs to solve simple math problems, which can be simply done with other programs. When conducting relevant research, it is essential to ask *whether the specific task being tackled is meaningful* and *whether the proposed method can be generalized to more realistic tasks and applications*.

Are language models really able to reason? There are several indications that LLMs are able to reason, including 1) high performance on various tasks requiring reasoning (Suzgun et al., 2022);

2) the ability to reason step-by-step with chain-of-thought prompting (Wei et al., 2022b); and 3) the reflection of human-like content effects on reasoning (Dasgupta et al., 2022). However, these findings are not sufficient to conclude that LLMs can truly reason. For 1), it is not clear whether the models are making predictions based on *reasoning* or *heuristics* (Patel et al., 2021). For many existing benchmarks on reasoning, actually, we can design a program with heuristic rules to achieve very high performance. We usually do not think a program relying on heuristic rules is capable of reasoning. For 2), although the models seem to reason step-by-step, the generated rationales may be incorrect and inconsistent. It is possible that the models are "generating reasoning-like response" rather than "reasoning step-by-step". For 3), while LLMs display some human-like reasoning patterns, this does not necessarily mean that they behave like humans.

Additionally, there are several observations that suggest LLMs may not be capable of reasoning: 1) LLMs still struggle with tasks that require complex reasoning (Valmeekam et al., 2022; Han et al., 2022a; Ruis et al., 2022). If LLMs are really decent reasoners, they should handle tasks that can be simply solved by humans through reasoning; 2) LLMs make mistakes in their reasoning, as explained above; 3)[#4] The performance of LLMs on downstream tasks has been found to be sensitive to the frequency of certain terms, such as numbers, in the training data (Razeghi et al., 2022; Jung et al., 2022), which would not be expected if the models were solving mathematical problems through reasoning; 4)[#] Language models have been found to struggle with associating relevant information that they have memorized (Huang et al., 2022c).

Overall, it is still too early to draw a conclusion about the proposed question. In fact, there is also an ongoing debate about whether language models can actually *understand* language or capture *meaning* (Bender and Koller, 2020; Li et al., 2021; Manning, 2022; Piantasodi and Hill, 2022). Further in-depth analysis of factors such as training data, model architecture, and optimization objectives is needed, as well as the development of better benchmarks for measuring the reasoning capabilities of LLMs. However, it is clear that the current models are not yet capable of robust reasoning.

Improving reasoning capabilities of LLMs. While techniques like chain-of-thought prompting (Wei et al., 2022b) may help to elicit reasoning abilities in large language models, they cannot enable the models to solve tasks beyond their current capabilities. To truly enhance reasoning in LLMs, we need to utilize training data, model architecture, and optimization objectives that are designed to encourage reasoning. For example, finetuning a model with a dataset including CoT data has been shown to improve reasoning (Chung et al., 2022), and models can also self-improve through the process of bootstrapping their reasoning (Zelikman et al., 2022; Huang et al., 2022a). There is still much research that needs to be done in this area, and we look forward to future progress in improving reasoning in large language models.

[4] #indicates the finding has not been carefully examined in language models with more than 100 billion parameters.

7 Conclusion

In this paper, we have provided a detailed and up-to-date review of the current state of knowledge on reasoning in large language models. We have discussed techniques for improving and eliciting reasoning in LLMs, methods and benchmarks for evaluating reasoning abilities, and the findings and implications of previous studies in this topic. While LLMs have made significant progress in natural language processing and related fields, it remains unclear to what extent they are capable of true reasoning or whether they are simply using memorized patterns and heuristics to solve problems. Further research is needed to fully understand the reasoning abilities of LLMs, improve LLMs' reasoning capabilities, and determine their potential for use in a variety of applications. We hope that this paper will serve as a useful overview of the current state of the field and stimulate further discussion and research on this interesting and important topic.

Limitations

In this paper, we provide an overview of the current state of knowledge on reasoning in large language models. Reasoning is a broad concept that encompasses various forms, making it impractical to summarize all related work in a single paper. Therefore, we focus on deductive reasoning, as it is the most commonly studied in the literature. Other forms of reasoning such as inductive reasoning (Yang et al., 2022; Misra et al., 2022, *inter alia*) and abductive reasoning (Wiegreffe et al., 2022; Lampinen et al., 2022; Jung et al., 2022, *inter alia*) may not be discussed in depth.

Additionally, given the rapid evolution and significance of reasoning within large language models, it is crucial to note that new contributions may have emerged in the field concurrent with the writing of this paper. An additional resource to consider is a parallel survey by Qiao et al. (2022), which emphasizes reasoning via language model prompting. Our coverage may not extend to papers released during or after 2023 such as evaluation on ChatGPT (Bang et al., 2023; Zheng et al., 2023). As such, we recommend readers to check the papers that cite this survey for a more comprehensive and updated understanding of this field.

Acknowledgements

We would like to thank Jason Wei (OpenAI) and Denny Zhou (Google DeepMind) for their valuable advice and constructive feedback on this work. This material is based upon work supported by the National Science Foundation IIS 16-19302 and IIS 16-33755, Zhejiang University ZJU Research 083650, IBM-Illinois Center for Cognitive Computing Systems Research (C3SR) and IBM-Illinois Discovery Accelerator Institute (IIDAI), gift grants from eBay and Microsoft Azure, UIUC OVCR CCIL Planning Grant 434S34, UIUC CSBS Small Grant 434C8U, and UIUC New Frontiers Initiative. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. Exploring length generalization in large language models. *ArXiv preprint*, abs/2207.04901.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv preprint*, abs/2302.04023.

Emily M. Bender and Alexander Koller. 2020. 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*, pages 5185–5198, Online. Association for Computational Linguistics.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pretrained language models: A survey. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *ArXiv preprint*, abs/2108.07258.

Hugo Bronkhorst, Gerrit Roorda, Cor Suhre, and Martin Goedhart. 2020. Logical reasoning in formal and everyday reasoning tasks. *International Journal of Science and Mathematics Education*, 18(8):1673–1694.

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *ArXiv preprint*, abs/2107.03374.

Wenhu Chen. 2022. Large language models are few (1)-shot table reasoners. *ArXiv preprint*, abs/2210.06710.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *ArXiv preprint*, abs/2211.12588.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *ArXiv preprint*, abs/2210.11416.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv preprint*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv preprint*, abs/2110.14168.

Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. *ArXiv preprint*, abs/2208.14271.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *ArXiv preprint*, abs/2205.09712.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *ArXiv preprint*, abs/2207.07051.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A Saurous, Jascha Sohl-Dickstein, et al. 2022. Language model cascades. *ArXiv preprint*, abs/2207.10342.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models. *ArXiv preprint*, abs/2209.15003.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. *ArXiv preprint*, abs/2212.04092.

Ward Edwards. 1954. The theory of decision making. *Psychological bulletin*, 51(4):380.

Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. 2004. *Reasoning about knowledge*. MIT press.

Yao Fu, Hao Peng, and Tushar Khot. 2022a. How does gpt obtain its ability? tracing emergent abilities of language models to their sources.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022b. Complexity-based prompting for multi-step reasoning. *ArXiv preprint*, abs/2210.00720.

Kathleen M Galotti. 1989. Approaches to studying formal and everyday reasoning. *Psychological bulletin*, 105(3):331.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *ArXiv preprint*, abs/2211.10435.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. *ArXiv preprint*, abs/2212.07919.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022a. Folio: Natural language reasoning with first-order logic. *ArXiv preprint*, abs/2209.00840.

Simon Jerome Han, Keith Ransom, Andrew Perfors, and Charles Kemp. 2022b. Human-like property induction is a challenge for large language models.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Rethinking with retrieval: Faithful large language model inference. *ArXiv preprint*, abs/2301.00303.

Chadi Helwe, Chloé Clavel, and Fabian M Suchanek. 2021. Reasoning with transformer-based models: Deep learning, but shallow reasoning. In *3rd Conference on Automated Knowledge Base Construction*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022a. Large language models can self-improve. *ArXiv preprint*, abs/2210.11610.

Jie Huang, Kevin Chang, Jinjun Xiong, and Wen-mei Hwu. 2022b. Open relation modeling: Learning to define relations between entities. In *Findings of the Association for Computational Linguistics: ACL*

2022, pages 297–308, Dublin, Ireland. Association for Computational Linguistics.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022c. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jie Huang, Kerui Zhu, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022d. DEER: Descriptive knowledge graph for explaining entity relationships. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6686–6698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022e. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022f. Inner monologue: Embodied reasoning through planning with language models. In *2022 Conference on Robot Learning*.

Michael Huth and Mark Ryan. 2004. *Logic in Computer Science: Modelling and reasoning about systems*. Cambridge university press.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *The 2022 Conference on Empirical Methods for Natural Language Processing*.

Seyed Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2022. Lambada: Backward chaining for automated reasoning in natural language. *ArXiv preprint*, abs/2212.13894.

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. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *ArXiv preprint*, abs/2210.02406.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2879–2888. PMLR.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Edward H Levi. 2013. *An introduction to legal reasoning*. University of Chicago Press.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *ArXiv preprint*, abs/2206.14858.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022a. Explanations from large language models make small reasoners better. *ArXiv preprint*, abs/2210.06726.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022b. On the advance of making language models better reasoners. *ArXiv preprint*, abs/2206.02336.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Chenzhengyi Liu, Jie Huang, Kerui Zhu, and Kevin Chen-Chuan Chang. 2022a. Dimongen: Diversified generative commonsense reasoning for explaining concept relationships. *ArXiv preprint*, abs/2212.10545.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022b. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *ArXiv preprint*, abs/2212.08410.

Christopher D Manning. 2022. Human language understanding & reasoning. *Daedalus*, 151(2):127–138.

Gary Marcus. 2020. The next decade in ai: four steps towards robust artificial intelligence. *ArXiv preprint*, abs/2002.06177.

Conor McHugh and Jonathan Way. 2018. What is reasoning? *Mind*, 127(505):167–196.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online. Association for Computational Linguistics.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, Florence, Italy. Association for Computational Linguistics.

Kanishka Misra, Julia Taylor Rayz, and Allyson Ettinger. 2022. A property induction framework for neural language models. *ArXiv preprint*, abs/2205.06910.

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. 2022. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2022. Show your work: Scratchpads for intermediate computation with language models. In *Deep Learning for Code Workshop*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

John Arthur Passmore. 1961. Philosophical reasoning.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8864–8880, Online. Association for Computational Linguistics.

Xinyu Pi, Qian Liu, Bei Chen, Morteza Ziyadi, Zeqi Lin, Yan Gao, Qiang Fu, Jian-Guang Lou, and Weizhu Chen. 2022. Reasoning like program executors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Steven T Piantasodi and Felix Hill. 2022. Meaning without reference in large language models. *ArXiv preprint*, abs/2208.02957.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *ArXiv preprint*, abs/2210.03350.

Ben Prystawski, Paul Thibodeau, and Noah Goodman. 2022. Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. *ArXiv preprint*, abs/2209.08141.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *ArXiv preprint*, abs/2212.09597.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv preprint*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning. *ArXiv preprint*, abs/2202.07206.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. *ArXiv preprint*, abs/2210.14986.

Jacob Russin, Randall C O'Reilly, and Yoshua Bengio. 2020. Deep learning needs a prefrontal cortex. *Work Bridging AI Cogn Sci*, 107:603–616.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *ArXiv preprint*, abs/2210.01240.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv preprint*, abs/2211.05100.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *ArXiv preprint*, abs/2210.03057.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2022. Distilling multi-step reasoning capabilities of large language models into smaller models via semantic decompositions. *ArXiv preprint*, abs/2212.00193.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2022. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *ArXiv preprint*, abs/2212.04088.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, abs/2206.04615.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *ArXiv preprint*, abs/2210.09261.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. In *Advances in Neural*

Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *ArXiv preprint*, abs/2211.09085.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Boshi Wang, Xiang Deng, and Huan Sun. 2022a. Iteratively prompt pre-trained language models for chain of thought. In *The 2022 Conference on Empirical Methods for Natural Language Processing*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2022b. Towards understanding chain-of-thought prompting: An empirical study of what matters. *ArXiv preprint*, abs/2212.10001.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022c. Self-consistency improves chain of thought reasoning in language models. *ArXiv preprint*, abs/2203.11171.

Peter C Wason. 1968. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281.

Peter Cathcart Wason and Philip Nicholas Johnson-Laird. 1972. *Psychology of reasoning: Structure and content*, volume 86. Harvard University Press.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Yixuan Weng, Minjun Zhu, Shizhu He, Kang Liu, and Jun Zhao. 2022. Large language models are reasoners with self-verification. *ArXiv preprint*, abs/2212.09561.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2022. Language models as inductive reasoners. *ArXiv preprint*, abs/2212.10923.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*.

Ping Yu, Tianlu Wang, Olga Golovneva, Badr Alkhamissy, Gargi Ghosh, Mona Diab, and Asli Celikyilmaz. 2022. Alert: Adapting language models to reasoning tasks. *ArXiv preprint*, abs/2212.08286.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STar: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022a. Opt: Open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. *ArXiv preprint*, abs/2210.03493.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers? *ArXiv preprint*, abs/2304.10513.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022a. Least-to-most prompting enables complex reasoning in large language models. *ArXiv preprint*, abs/2205.10625.

Fan Zhou, Haoyu Dong, Qian Liu, Zhoujun Cheng, Shi Han, and Dongmei Zhang. 2022b. Reflection of thought: Inversely eliciting numerical reasoning in language models via solving linear systems. *ArXiv preprint*, abs/2210.05075.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022c. Teaching algorithmic reasoning via in-context learning. *ArXiv preprint*, abs/2211.09066.

Corinne Zimmerman. 2000. The development of scientific reasoning skills. *Developmental review*, 20(1):99–149.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
See the limitation section

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
See the abstract and introduction section

☑ A4. Have you used AI writing assistants when working on this paper?
We wrote part of the appendix with ChatGPT assistance (e.g., to generate an initial description for commonsense reasoning). The generated text is carefully revised and examined by the authors.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

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

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

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

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.