Reasoning with Language Model Prompting: A Survey

Shuofei Qiao[1,2]*, Yixin Ou[1,2]*, Ningyu Zhang[1,2]†, Xiang Chen[1,2], Yunzhi Yao[1,2], Shumin Deng[4], Chuanqi Tan[3], Fei Huang[3], Huajun Chen[1,2]†
[1] Zhejiang University & AZFT Joint Lab for Knowledge Engine
[2] Hangzhou Innovation Center, Zhejiang University
[3] Alibaba Group [4] National University of Singapore
{shuofei,ouyixin,zhangningyu,xiang_chen,yyztodd,huajunsir}@zju.edu.cn
shumin@nus.edu.sg {chuanqi.tcq,f.huang}@alibaba-inc.com

Abstract

Reasoning, as an essential ability for complex problem-solving, can provide back-end support for various real-world applications, such as medical diagnosis, negotiation, etc. This paper provides a comprehensive survey of cutting-edge research on reasoning with language model prompting. We introduce research works with comparisons and summaries and provide systematic resources to help beginners. We also discuss the potential reasons for emerging such reasoning abilities and highlight future research directions[1].

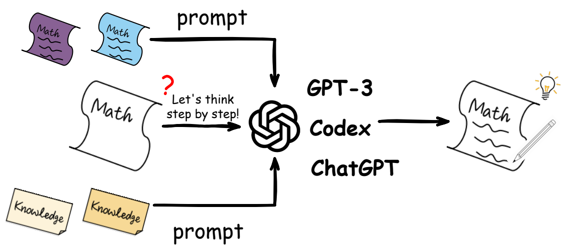

Figure 1: Reasoning with language model prompting. In-context exemplars (colored ●, ●), knowledge (colored ●, ●) or just *Let's think step by step!* are as prompt to enhance language models reasoning.

1 Introduction

Reasoning ability lies at the heart of human intelligence, yet in natural language processing (NLP), modern neural networks can hardly reason from what they are told or have already known (Duan et al., 2020; Wang et al., 2021; Bhargava and Ng, 2022). Fortunately, with the revolutionary development of pre-training (Brown et al., 2020; Chen et al., 2021; Chowdhery et al., 2022), scaling up the size of language models (LMs) has shown to confer a range of reasoning abilities, such as arithmetic (Wang et al., 2022g; Lewkowycz et al., 2022), commonsense (Jung et al., 2022; Liu et al., 2022b) and symbolic (Zhou et al., 2022a; Khot et al., 2022) reasoning. As shown in Figure 1, such abilities may be unlocked by prompting strategies (Liu et al., 2022d) (e.g., *chain-of-thought (CoT) prompting* (Wei et al., 2022b), *generated knowledge prompting* (Liu et al., 2022c), which can dramatically narrow the gap between human and machine intelligence. Likewise, a vast amount of work has been proposed in the NLP community; however, these approaches, scattered among various tasks, have not been systematically reviewed and analyzed.

Organization of This Survey: In this paper, we conduct the first survey of recent progress in reasoning with language model prompting. We first give some preliminaries on this direction (§2) and then propose to organize relevant work by taxonomy (§3). We further provide in-depth comparisons with discussion for insights (§4). To facilitate beginners who are interested in this field, we highlight some open resources (§5) and potential future directions (§6).

2 Preliminaries

In this section, we introduce preliminaries of reasoning with LM prompting. For standard prompting, given the reasoning question Q, prompt T and parameterized probabilistic model p_{LM}, we aim to maximize the likelihood of answer A as:

$$p(A \mid T, Q) = \prod_{i=1}^{|A|} p_{\text{LM}}\left(a_i \mid T, Q, a_{<i}\right) \qquad (1)$$

where a_i is the i-th token of A, and $|A|$ denotes the length of A. For few-shot prompting, T is comprised of K exemplars of (Q, A) pair.

To enhance the reasoning ability with LM prompting, there are two major branches of research. The first one focuses on optimizing the **reasoning strategy** with prompting (§3.1) as shown

* Equal Contribution.
† Corresponding Author.
[1]Work in progress and resources are available at https://github.com/zjunlp/Prompt4ReasoningPapers (updated periodically).

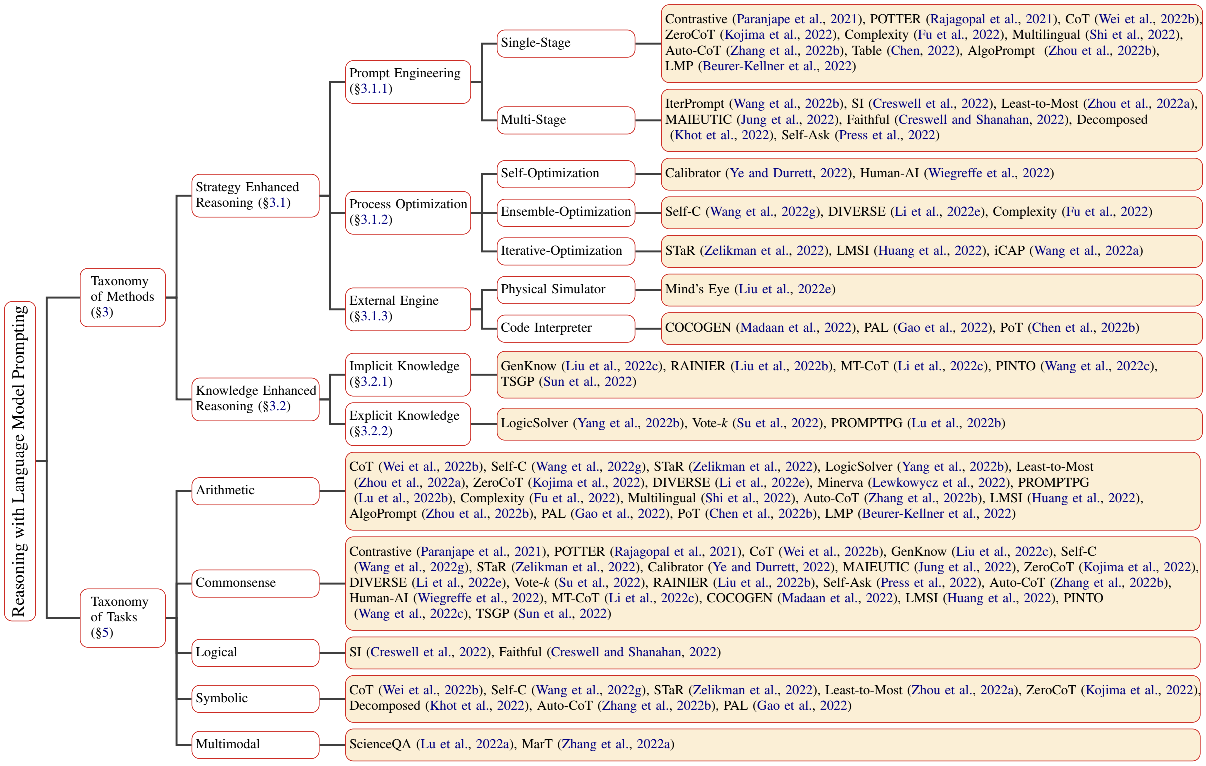

Figure 2: Taxonomy of Reasoning with Language Model Prompting.

in Figure 2, including prompt engineering (§3.1.1) and process optimization (§3.1.2).

For prompt engineering (§3.1.1), many approaches try to improve the quality of prompt T, and we call those works **single-stage methods**, while others append c_i into the context of (T, Q) at each reasoning stage or design specific T_{c_i} for each c_i, and we regard those as **multi-stage methods**. For example, Wei et al. (2022b) try to *add reasoning steps* C into prompt where $T = \{(Q_i, C_i, A_i)\}_{i=1}^{K}$, thus Equation 1 can be reformed to:

$$p(A \mid T, Q) = p\,(A \mid T, Q, C)\, p\,(C \mid T, Q) \quad (2)$$

where $p(C \mid T, Q)$ and $p(A \mid T, Q, C)$ are defined as follows:

$$p(C \mid T, Q) = \prod_{i=1}^{|C|} p_{\text{LM}}\,(c_i \mid T, Q, c_{<i})$$

$$p(A \mid T, Q, C) = \prod_{j=1}^{|A|} p_{\text{LM}}\,(a_j \mid T, Q, C, a_{<j})$$

with c_i is one step of total $|C|$ reasoning steps.

For process optimization (§3.1.2), the simplest ways are to bring in an optimizer with parameters $\boldsymbol{\theta}$ to calibrate C when generating A, and we call those works **self-optimization methods**. Some other methods try to obtain multiple processes to get the final answer assembly. We regard those works as **ensemble-optimization methods**. Moreover, the overall optimization process can be iteratively integrated with fine-tuning the p_{LM} on generated triplet (Q, C, A), which are regarded as **iterative-optimization methods**. Besides, some works leverage external reasoning engines (§3.1.3) to produce T or directly execute C for reasoning.

The second one focuses on **knowledge enhancement** with prompting (§3.2). Note that rich **implicit** "modeledge" (Han et al., 2021) in LMs can generate knowledge or rationales as knowledge-informed prompt T (§3.2.1). Meanwhile, **explicit** knowledge in external resources can also be leveraged and retrieved as knowledgeable prompts to enhance reasoning (§3.2.2).

3 Taxonomy of Methods

In this paper, we survey existing reasoning with LM prompting methods, categorizing them into *Strategy Enhanced Reasoning* (§3.1) and *Knowledge Enhanced Reasoning* (§3.2). As shown in Figure 2, we further refine them according to the distinctive features of different methods.

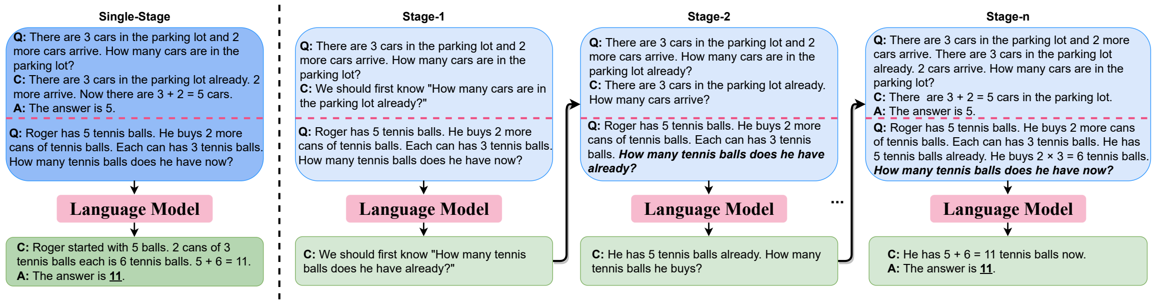

Figure 3: **Single-Stage** (**left**) and **Multi-Stage** (**right**) in Prompt Engineering (§3.1.1) of Strategy Enhanced Reasoning. In each stage, a question (**Q**) prompted with several exemplars containing reasoning steps (**C**) will be fed into the LM. The outputs are reasoning steps and the final answer (**A**) of the given question.

3.1 Strategy Enhanced Reasoning

The primary purpose of this line of work is to design a better reasoning strategy to enhance the performance of LMs reasoning, concretely embodied in *prompt engineering* (§3.1.1), *process optimization* (§3.1.2) and *external engine* (§3.1.3).

3.1.1 Prompt Engineering

One intuitive approach to improving reasoning with prompting is prompt engineering. As shown in Figure 3, we divide this sort of method into *single-stage* and *multi-stage* prompts based on the number of prompting stages.

Single-Stage. Early works leverage template-based prompts (Paranjape et al., 2021; Rajagopal et al., 2021) for reasoning in NLP. Regarding the strong in-context learning ability of large language models (Brown et al., 2020), Wei et al. (2022b) proposes CoT prompting, which adds a series of intermediate reasoning steps, also called CoT, into exemplars of few-shot prompt to induce large language models to generate a reasoning process before answering. Experiments demonstrate that large language models emerge with impressive reasoning abilities with CoT prompting.

In spite of the large improvement brought by CoT, in-context learning is greatly sensitive to the selection of exemplars, and even a tiny change may cause a large drop in model performance (Lu et al., 2022c; Min et al., 2022; Webson and Pavlick, 2022). Hence, the quality of exemplars appears to be particularly important. Fu et al. (2022) indicates that prompts with higher reasoning complexity, e.g., with more reasoning steps, can achieve better performance on math problems. Zhang et al. (2022b) explores the impact of diversity of exemplars in prompt. Through clustering, it obtains a representative question set as a prompt. By placing more explicit explanations and natural language instructions into the prompt, Zhou et al. (2022b) relieves the ambiguity for LMs when facing out-of-distribution (OOD) algorithmic problems. The above work shows that LMs can be outstanding few-shot reasoners. Surprisingly, Kojima et al. (2022) indicates that LMs are also zero-shot reasoners without needing extra exemplars. By only concatenating *"Let's think step by step"*, LMs can consciously generate reasoning steps.

Multi-Stage. When human beings are reasoning, it is usually challenging to come up with the whole reasoning process in one stroke. A more intuitive solution is to decompose a complex problem into simpler sub-problems and reason stage by stage. Similarly, this series of works aims to transform previous one-stage prompting into multi-stage prompting. Press et al. (2022) explicitly defines follow-up questions and intermediate answers in prompts to narrow the compositionality gap in LMs. Jung et al. (2022) regard the output of each stage as a separate new question while Zhou et al. (2022a); Wang et al. (2022b) append it to the whole context to prompt LMs. Khot et al. (2022) first decomposes a task into *split* and *merge* sub-tasks and then constructs specific prompts to tackle each sub-task. Creswell and Shanahan (2022) follows a structure of Selection-Inference (Creswell et al., 2022) which selects specific contexts and inferences based on them at each stage.

3.1.2 Process Optimization

Natural language rationales[2] (Ling et al., 2017a), also called reasoning process in CoT, plays a vital role in CoT prompting (Ye and Durrett, 2022; Lampinen et al., 2022; Min et al., 2022). The

[2] Some references (Ye and Durrett, 2022; Wiegreffe et al., 2022; Zhou et al., 2022b) regard this as explanations.

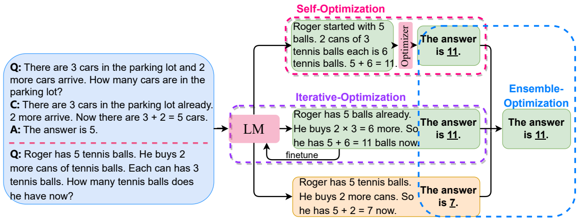

Figure 4: Process Optimization (§3.1.2) of Strategy Enhanced Reasoning. **Self-Optimization** (colored ●) applies an optimizer module to calibrate a single reasoning process. **Ensemble-Optimization** (colored ●) assembles multiple reasoning processes to calibrate the final answer. **Iterative-Optimization** (colored ●) calibrates reasoning processes by iteratively finetuning the language model.

consistency of the reasoning process (Wang et al., 2022g), as well as the continuity between reasoning steps (Li et al., 2022e) both should affect the accuracy of final answers. Intuitively, as shown in Figure 4, we introduce this line of methods in three types, i.e., *self*, *ensemble*, and *iterative* process optimization.

Self-Optimization. Here, self-optimization refers to correcting one process by injecting extra modules. To mitigate the influence of the unreliability of rationales, Ye and Durrett (2022) utilizes a calibrator to tune the probabilities of a prediction based on the score, which reflects the factuality of a rationale. During free-text rationales generation, Wiegreffe et al. (2022) finetunes a sequence-to-sequence model as a filter to predict whether the explanation is acceptable.

Ensemble-Optimization. Due to the limitation of only one reasoning path, the following works rely on ensemble calibration among multiple processes. Wang et al. (2022g) introduce sampling strategies (Ackley et al., 1985; Fan et al., 2018) commonly used in natural language generation to obtain multiple reasoning processes and generate the most consistent answer by majority vote. Based on the motivation of when a reasoning process reaches a wrong answer, not all the steps may undertake the final incorrectness, Li et al. (2022e) proposes a step-aware voting versifier to score each reasoning path. When disorientated majority processes overwhelm reasonable minority processes, the step-aware voting versifier can alleviate the limitation of vanilla majority vote (Wang et al., 2022g). Besides, Wang et al. (2022f) empirically observe that decoder sampling in the output space is the key to robustly improving performance because of the brittleness of manual prompt engineering.

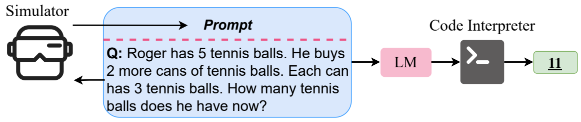

Figure 5: External Engine (§3.1.3) of Strategy Enhanced Reasoning. External engines play the role of prompt producer (**Physical Simulator**) or reasoning executor (**Code Interpreter**) to assist LMs in reasoning.

Iterative-Optimization. Note that LMs can achieve excellent performance in few-shot (Wei et al., 2022b) or zero-shot (Kojima et al., 2022) manners with prompts, another paradigm is to calibrate reasoning processes iteratively with LM finetuning. Specifically, iterative-optimization-based methods try to repeat the process of prompting LMs to generate reasoning processes and use the instances with generated reasoning processes to finetune themselves. Zelikman et al. (2022) initiates with a small set of exemplars to push LMs to produce reasoning steps and answers themselves. Questions and reasoning steps with the correct answers will be directly added to the dataset for finetuning. Incorrect ones will be fed into the model again by being tagged on a hint that labels the correct answer. Compared with Zelikman et al. (2022), Huang et al. (2022) do not need gold labels during self-teaching. Following Wang et al. (2022g), it generates multiple reasoning processes and keeps the ones that lead to the most consistent answer. Then it finetunes the model on these self-generated reasoning-answer data. Wang et al. (2022a) proposes an iterative context-aware prompter which learns to dynamically synthesize prompts conditioned on the contexts of current step.

3.1.3 External Engine

When reasoning with LM prompting, the models should have the ability of semantic understanding (e.g., questions) and complex reasoning (e.g., by generating reasoning processes); however, we cannot have both fish and bear's paw (Hendrycks et al., 2021; Nogueira et al., 2021; Lewkowycz et al., 2022). To tear up the obstacle, external reasoning engines lend a helping hand to LMs (see Figure 5).

Physical Simulator. Given a physical reasoning question, Liu et al. (2022e) utilizes a computational physics engine (Todorov et al., 2012) to simulate the physical process. The simulation results are

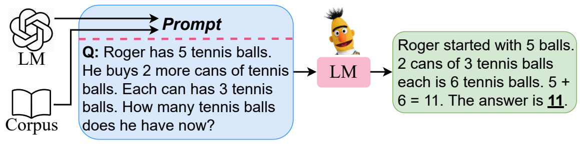

Figure 6: Knowledge Enhanced Reasoning (§3.2). Prompts are generated by LMs (**Implicit Knowledge**) or retrieved from external resources (**Explicit Knowledge**).

treated as prompt to help LMs reason, making up for the lack of physical knowledge in LMs.

Code Interpreter. With the emergence of LMs of code (Chen et al., 2021; Xu et al., 2022), collaborating LMs and codes to tackle specific tasks has recently sprung up (Wang et al., 2022e; Cheng et al., 2022; Wu et al., 2022b). Note that programs yield advantage behaviors in robustness and interpretability and can better illustrate complex structures and deduct complex calculations. Intuitively, Madaan et al. (2022) reframes structured commonsense reasoning tasks as code generation tasks, replacing the natural language with python class code to represent structured graph both in few-shot prompts and LM outputs. Gao et al. (2022) decomposes solving step from LMs to programmatic runtime and mainly remain to learn task for LMs. In few-shot prompts and LM outputs, the reasoning processes are replaced by a mixture of natural and programming language, where natural language is treated as annotations to aid the generation of the program. Similar to Gao et al. (2022), Chen et al. (2022b) proposes *program of thoughts* (PoT) prompting which disentangling computation from reasoning. The main difference is that it also puts forward a zero-shot format of PoT prompting.

3.2 Knowledge Enhanced Reasoning

Knowledge is the cornerstone of reasoning. Knowledge enhanced methods aim to prompt LMs with *implicit* (§3.2.1) or *explicit* (§3.2.2) knowledge to assist in reasoning (see Figure 6).

3.2.1 Implicit Knowledge

Researchers have shown that LMs contain considerable implicit knowledge which can be elicited via conditional generation (Davison et al., 2019; Petroni et al., 2019; Jiang et al., 2020). The following works try to induce such "modeledge" as knowledge-informed prompts for reasoning.

Liu et al. (2022c) applies GPT-3 (Brown et al., 2020) with few-shot prompting to generate knowledge and prompts the downstream LM. ased on this, Liu et al. (2022b) draws support from reinforcement learning (Schulman et al., 2017) to further calibrate the knowledge. Different from the above, which only uses few-shot prompting in the knowledge generation stage, Sun et al. (2022) proposes a two-stage generative prompting which additionally includes answer generation prompts. Li et al. (2022c) and Wang et al. (2022c) both follow the paradigm of generating explanations with prompting a larger LM and then finetuning on a smaller LM. They mainly use the strong generation ability of LMs with few-shot prompting.

3.2.2 Explicit Knowledge

Although large LMs have shown strong generation ability (Wiegreffe et al., 2022; Li et al., 2022c; Wang et al., 2022c), they still have the tendency to hallucinate facts (Rohrbach et al., 2018) and generate inconsistent knowledge (Liu et al., 2022b). Recent works show that retrieving prompts for in-context learning is a nice means to achieve good performance (Liu et al., 2022a; Rubin et al., 2022).

Due to the instability of Liu et al. (2022a) to measure the similarity of structured information, Lu et al. (2022b) propose a dynamic prompt retrieval method based on policy gradient strategy, without brute-force searching. Su et al. (2022) formulates a selective annotation framework to avoid the need for a large labeled retrieval corpus. It develops a graph-based method to construct a diverse and representative small labeled database as much as possible from a large unlabeled corpus. Then the in-context labeled examples can be retrieved from the small database, which largely reduces the cost of annotation and retrieval.

4 Comparison and Discussion

4.1 Comparison of Language Models

Table 1 shows four comparison scopes of different methods. We further illustrate the performance comparison of LMs with different scales on GSM8K (Cobbe et al., 2021) of arithmetic reasoning in Figure 7. Similar results on commonsense reasoning benchmarks are shown in Appendix B.

Wei et al. (2022b) systematically demonstrates that few-shot prompting performs better in almost all tasks as model scale increases, which can be explained by the fact that **LMs with larger model size contain more implicit knowledge for reasoning** (Liang et al., 2022b). Moreover, CoT

Category	Representative Method	Comparison Scope			
		How to Get Prompt	Prompt Type	Language Model	Training Scenario
Prompt Engineering	POTTER (Rajagopal et al., 2021)	Manual	Template	BART/T5	full fine-tune
	CoT (Wei et al., 2022b)	Manual	CoT	UL2/LaMDA/GPT-3/Codex/PaLM	few-shot prompt
	Auto-CoT (Zhang et al., 2022b)	LM Generated	CoT	GPT-3/Codex	few-shot prompt
	Least-to-Most (Zhou et al., 2022a)	Manual	CoT	GPT-3/Codex	few-shot prompt
Process Optimization	Calibrator (Ye and Durrett, 2022)	Manual	Explanations	InstructGPT	few-shot fine-tune
	Self-Consistency (Wang et al., 2022g)	Manual	CoT	UL2/LaMDA/Codex/PaLM	few-shot prompt
	DIVERSE (Li et al., 2022e)	LM Generated	CoT	GPT-3/Codex	few-shot prompt
	LMSI (Huang et al., 2022)	LM Generated	CoT	PaLM	self-train
External Engine	PAL (Gao et al., 2022)	Manual	Code	Codex	few-shot prompt
	PoT (Chen et al., 2022b)	Manual	Code	Codex	few-shot prompt
Implicit Knowledge	RAINIER (Liu et al., 2022b)	LM Generated	Knowledge	UnifiedQA	few-shot prompt
	PINTO (Wang et al., 2022c)	LM Generated	Explanations	ROBERTA/T5	full fine-tune
Explicit Knowledge	PROMPTPG (Lu et al., 2022b)	Retrieval	CoT	GPT-3	few-shot prompt

Table 1: Comparison of reasoning with prompting methods from different scopes.

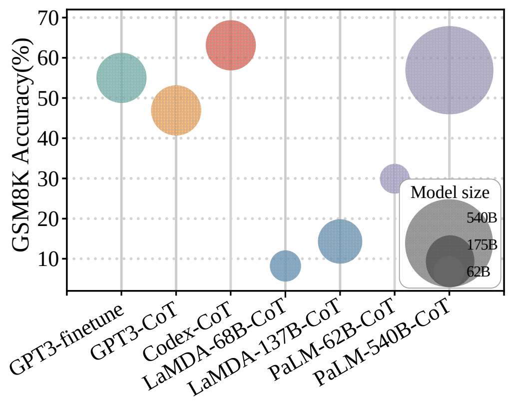

Figure 7: Performance of different language model scales on arithmetic reasoning. Representatively, we show CoT (Wei et al., 2022b) experimental results on GSM8K (Cobbe et al., 2021).

prompting produces much greater increases, with PaLM-540B showing the greatest improvements, as depicted in Figure 7&8. However, when the model scale declines to less than 100B, CoT prompting will yield no performance gain and may even be detrimental. Thus, CoT prompting elicits an emergent ability of model scale, which is defined as abilities of pre-trained LMs which are not present in smaller-scale models but in large-scale models (Wei et al., 2022a). Another intriguing observation is depicted in Figure 7&8 that PaLM-62B (Chowdhery et al., 2022) even performs better than LaMDA-137B (Thoppilan et al., 2022), possibly because it was trained on the higher-quality corpus.

Notably, Figure 7&8 also illustrates that holding the same parameter scale, Codex (Chen et al., 2021) outperforms GPT-3 significantly[3], even though the major difference between them is the training corpus (Codex is a GPT-3 variant training on code). This phenomenon can also be inspected in recent works (Zhou et al., 2022a; Li et al., 2022e; Fu et al., 2022; Zhang et al., 2022b; Madaan et al., 2022; Liang et al., 2022b), indicating that **pre-training on code branch not only enables the ability of code generation/understanding but may also trigger the reasoning ability with CoT**. The exact cause is still elusive, but one theory could be that code is a more reasonable form of text, thinking about procedure-oriented programming is analogous to solving problems step by step, and object-oriented programming is analogous to decomposing complex tasks into simpler ones[4].

4.2 Comparison of Prompts

Table 1 shows the comparison of different methods of reasoning with LM prompting. There are three main sources of prompts for existing methods: 1) **Manual** construction is suitable for template-based prompts and few-shot prompting where the prompt is uncomplicated. 2) **LM Generated** prompt makes up for the shortcomings of manual construction prompt. It can customize specific rationales for each question and provide sufficient knowledge with the prompt for fine-tuning or self-training. 3) **Retrieval**-based prompt often relies on well-annotated external resources (e.g., Wikipedia) and consumes expensive information retrieval, but it can alleviate the unstable issue of the generation.

We observe that no matter how prompt is produced, CoT only works on large LMs under few-shot prompting. Combined with the empirical conclusion in Ye and Durrett (2022), these phenomena reveal that **explicit high-quality reasoning rationales contained in the input context are**

[3]Note that Codex and GPT-3 in our paper refer to code-davinci-002 and text-davinci-002 respectively in OpenAI API.

[4]https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources

the keys for reasoning with LM prompting. Although some works have attempted to explore the in-context learning ability on large LMs (Xie et al., 2022; Min et al., 2022; Akyürek et al., 2022), the reason why CoT prompting can succeed on large LMs is still intriguing to the community and not well-understood. One possible hypnosis is that CoT is a magical side product of training on code and unlocked by prompt. Note that exemplars containing CoT in few-shot prompts can be viewed as a kind of instruction that arouses the reasoning ability hidden in large LMs. Chung et al. (2022) verifies the similar result using CoT in instruction fine-tuning to further advance model performance.

5 Benchmarks and Resources

5.1 Taxonomy of Benchmarks and Tasks

Researchers in the NLP community have released many benchmarks requiring various reasoning skills, including arithmetic reasoning, commonsense reasoning, logical reasoning, symbolic reasoning and multimodal reasoning. In this section, we will give a brief overview of these reasoning benchmarks and tasks. More details of broader benchmarks, as well as reasoning with ChatGPT can be found in Appendix C and D.

Arithmetic Reasoning. Arithmetic reasoning, also referred to as mathematical reasoning, is the ability to perform reasoning on *math word problems* (MWP). Arithmetic reasoning skills are of great importance abilities of human intelligence and are also essential for general-purpose artificial intelligent systems. Early works on this task (Hosseini et al., 2014; Kushman et al., 2014; Roy et al., 2015; Koncel-Kedziorski et al., 2015; Roy and Roth, 2015; Ling et al., 2017b) focus on relatively small datasets consisting of grade school single-step or multi-step MWP, relevant math operations of which cover $+, -, \times, \div$. Later works increase in complexity and scale, and other datasets are proposed to increase the difficulties. Most recently, Mishra et al. (2022a) extends existing datasets to construct a unified benchmark concerning mathematical abilities, language format, language diversity and external knowledge.

Commonsense Reasoning. Commonsense knowledge and commonsense reasoning are some of the major issues in machine intelligence (Storks et al., 2019; Bhargava and Ng, 2022). When answering a question, people often draw upon their rich world knowledge. For LMs, the major challenge of performing commonsense reasoning lies in how to involve physical and human interactions under the presumption of general background knowledge (Bhargava and Ng, 2022). Many benchmark datasets and tasks (Clark et al., 2018; Mihaylov et al., 2018; Talmor et al., 2019; Bisk et al., 2020; Geva et al., 2021) are designed to evaluate the ability of machines to learn commonsense knowledge in order to reason natural language text. The most widely used benchmark today is CommonsenseQA (Talmor et al., 2019), which focuses on commonsense question answering, based on knowledge encoded in ConceptNet (Speer et al., 2017).

Logical Reasoning. Common forms of logical reasoning include deductive reasoning and inductive reasoning. Deductive reasoning is performed by going from general information to specific conclusions; typical datasets in this field consist of synthetic rule bases plus derived conclusions (Clark et al., 2020; Tafjord et al., 2021). Recently, Dalvi et al. (2021) creatively proposes a dataset to contain multi-step entailment trees, aiming to fulfill models with the ability to generate explanations showing the line of reasoning from what is known to the answer. As opposed to deductive reasoning, inductive reasoning aims to draw conclusions by going from the specific to the general. Sinha et al. (2019) constructs a diagnostic benchmark requiring LM's abilities of both extracting relations between entities as well as generating the logical rules.

Symbolic Reasoning. Symbolic reasoning here only refers to a narrow collection of simple tasks that test a diverse set of symbolic manipulation functions, rather than symbolic AI, which is a more general concept implemented by rules engines or expert systems, or knowledge graphs. The construction of these tasks are usually well-defined by human; thus, it's easy to split the test set into in-domain test sets as well as out-of-domain test sets. Typical symbolic reasoning tasks include last letter concatenation, reverse list and coin flip (Wei et al., 2022b).

Multimodal Reasoning. Most existing benchmarks for reasoning are restricted to the textual-only modality and limited domain diversity. However, humans utilize the information available across different modalities when performing reasoning. To this end, multimodal reasoning benchmarks are presented to narrow this gap. Zellers et al. (2019) seeks to answer cognition-level questions from images, and Park et al. (2020) checks how well PLMs reason about the dynamic context from a static image and an event. Recently, Lu et al. (2022a) present ScienceQA, a large-scale multimodal multiple choice dataset that consists of diverse questions of science topics with corresponding answers and explanations. Zhang et al. (2022a) proposes the new task of multimodal analogical reasoning over knowledge graphs, which requires multimodal reasoning ability with the help of background knowledge.

Apart from the above-mentioned specific reasoning tasks, there are some benchmarks (Lake and Baroni, 2017; Srivastava et al., 2022) that can evaluate the model's more diverse and generalized reasoning capabilities, which can also be included in the category of reasoning tasks. Most recently, Yu et al. (2022) introduces ALERT, a benchmark that spans over 20 datasets and covers 10 different reasoning skills, to assess different LMs on fine-grained reasoning skills.

5.2 Resources

Thanks to the open-source spirit of the NLP community, numerous resources are publicly available alongside papers for researchers to experiment with. ThoughtSource is a central, open resource and community around data and tools related to CoT reasoning in large language models[5]. The LangChain library is designed to help developers build applications using LLMs combined with other sources of computation or knowledge[6]. Helwe et al. (2022) develops a PyTorch-based library called LogiTorch for logical reasoning on natural language.

6 Future Directions

Even though numerous works have been proposed for reasoning with language model prompting, there remain some potential directions:

Theoretical Principle of Reasoning. LMs have been demonstrated to have emergent zero-shot learning and reasoning abilities (Wei et al., 2022b; Wang et al., 2022g; Wei et al., 2022a). To uncover the mystery of such a success, many researchers have empirically explored the role of in-context learning (Ye and Durrett, 2022; Liu et al., 2022a) and rationales (Min et al., 2022; Lampinen et al., 2022). Another line of works tries to investigate the architecture of Transformers via knowledge neurons (Dai et al., 2022) or skill neurons (Wang et al., 2022d). However, the potential theoretical principle of reasoning for LM prompting still needs to be better understood. As discussed in §4, there is a hypothesis that there may have correlations between code and reasoning ability/CoT. More recent works, Wang et al. (2022e) and Madaan et al. (2022) demonstrate that pre-trained LMs of code are better handling structured commonsense reasoning and structure prediction than LMs of natural language, even when the downstream task does not involve source code at all. Thus, the truth may be close, and we argue that it is beneficial to study the theoretical principle to advocate for a transparent view of reasoning with LM prompting and further decipher the dark matter of intelligence by highlighting the counterintuitive continuum across language, knowledge, and reasoning[7].

However, the code-based pre-training (or restructured pre-training (Yuan and Liu, 2022)) still has limitations since it has to utilize off-the-shelf structure (e.g., existing aligned corpus or build from scratch via syntax tree or AMR (Banarescu et al., 2013)) to reformulate plain texts. Thus, we envision developing unified foundation architectures instead of Transformers (Vaswani et al., 2017) to encode multi-grained/multimodal data, possibly inspired by physics, cognitive, or brain science, and some cutting-edge technologies such as spiking neural network (DeWolf, 2021), diffusion models (Dhariwal and Nichol, 2021), ordinary differential equations (Li et al., 2022a) can also provide insightful solutions.

Efficient Reasoning. To be noted, existing methods mainly depend on large LMs, which may consume high computing resources. Regarding practicality, it is necessary to study reasoning with small LMs or develop efficient reasoning methodologies which pay attention to carbon emission and energy usage during model training and inference (Xu et al., 2021). One feasible way may be developing models that can enable generalization across a range of evaluation scenarios such as Flan-T5 (Chung et al., 2022), which finetune both with and without exemplars (i.e., zero-shot and few-shot) and with and without CoT. Another way is con-

[5] https://github.com/OpenBioLink/ThoughtSource

[6] https://github.com/hwchase17/langchain

[7] Quoted from the keynote talk on ACL 2022 entitled "2082: An ACL Odyssey: The Dark Matter of Intelligence and Language.

sulting *language model cascades* (Dohan et al., 2022) where we can cascade several smaller LMs to compete with a larger one. Recently, an intuitive approach has been proposed to transfer the reasoning capabilities of large LMs to smaller LMs via knowledge distillation (Magister et al., 2022). Other promising directions include retrieval augmentation (Li et al., 2022b), model editing (Cao et al., 2021; Mitchell et al., 2022a,b), delta-tuning (He et al., 2022; Mao et al., 2022; Pal et al., 2022; Ding et al., 2022), etc.

Robust and Interpretable Reasoning. Robustness and interpretability have long been pursued by the field of deep learning, especially in tasks that require strong logic, like reasoning. Creswell and Shanahan (2022) leverages a selection-inference (Creswell et al., 2022) multi-stage architecture for faithful reasoning, but there is still a lack of interpretability within each stage. Code-based works (Madaan et al., 2022; Gao et al., 2022; Chen et al., 2022b) reach robustness and interpretability to some extent, but they have the aid of an external engine. There is still a long way to achieve true robustness and interpretability with LMs. Fortunately, Dohan et al. (2022) provides a new idea for utilizing a probabilistic program with LMs to tackle various language-based reasoning problems. Other possible solutions may be utilizing neural-symbolic methodologies (Du et al., 2021; Li et al., 2022d; Ouyang et al., 2021; Feng et al., 2022) or reinforcement learning from human feedback like ChatGPT (OpenAI, 2022).

Multimodal (Interactive) Reasoning. Textual reasoning is restricted to what can be expressed through natural language. A more promising direction is multimodal reasoning regarding the information diversity of the real world of human reasoning. A recent work Lu et al. (2022a) generates CoT when dealing with a multimodal dataset; however, it simply extracts textual descriptions from images, and it is still a textual reasoning task indeed. Intuitively, it is beneficial to integrate multimodal information into reasoning processes such as images, audio, videos, etc., and design a unified multimodal CoT. Apart from unified multimodal models, it is also promising to model chains (Wu et al., 2022a) to conduct interactive reasoning among models of different modalities. Besides, (Sap et al., 2022) show that one of today's largest language models (GPT-3 (Brown et al., 2020)) lacks the skill to reason about the mental states, and reactions of all people involved. Thus, interactive reasoning methodologies should be noted by inspiring from other domains (e.g., Cognitive Science (Hollenstein et al., 2019), Social Intelligence (Krishna et al., 2022)), which may have potential guidance for reasoning in NLP since only increasing the scale of LMs is likely not the most effective way to create AI systems.

Generalizable (True) Reasoning. Generalization is one of the most significant symbols of models to attain true reasoning abilities. Given a reasoning task, we hope LMs can handle not only the problem itself but solve a group of similar reasoning tasks (not seen during training). Zhou et al. (2022b); Anil et al. (2022) explore the OOD problem on the length of reasoning questions, but the true generalization is still far from satisfactory. Meanwhile, Kejriwal et al. (2022) highlights that more comprehensive evaluation methods grounded in theory (e.g., naive physics (Gardin and Meltzer, 1989) and commonsense psychology (Gordon and Hobbs, 2004)) should be proposed. We argue that the generalizable reasoning may be closely related to analogy reasoning (Chen et al., 2022a), causal reasoning (Feder et al., 2022), compositional reasoning (Yang et al., 2022a), etc. Besides, generalizable reasoning abilities may be separated from the theoretical principle, robustness, and interpretability mentioned above, which should be considered from multiple perspectives.

7 Conclusion and Vision

In this paper, we provide a review of reasoning with language model prompting, including comprehensive comparisons, summaries, and several research directions. Note that reasoning in NLP has the potential advantages of complex problem-solving and should better utilize dark matters in cross-disciplines (e.g., Theory of Mind (Sap et al., 2022)). In the future, we envision a more potent synergy between the methodologies from the NLP and other domains (e.g., Cognitive Science (Hollenstein et al., 2019), Social Intelligence (Krishna et al., 2022)) for reasoning. We hope sophisticated and efficient LM prompting models will increasingly contribute to improving reasoning performance. On the converse, since just increasing the scale of LMs is probably not the most efficient strategy for building AI systems, we anticipate that theories in other domains can provide valuable guidance for reasoning in NLP.

Limitations

In this study, we provide a survey of reasoning with language model prompting. Due to the page limit, we may miss some important references (**we will maintain the Github resources**[8]) and cannot afford all the technical details. Moreover, we mainly review the cutting-edge works within two years (mostly in 2022), mainly from the ACL, EMNLP, NAACL, NeurIPS, ICLR, etc. Besides, our work (under this version) may miss some kind of reasoning tasks such as reasoning with generics (Allaway et al., 2022), default inheritance reasoning (Brewka, 1987), non-monotonic reasoning (Ginsberg, 1987) in NLP, and will try our best to fulfill this gap. We discuss the related surveys in Appendix A and will continue adding more related approaches with more detailed analysis.

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cogn. Sci.*, 9(1):147–169.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What learning algorithm is in-context learning? investigations with linear models. *CoRR*, abs/2211.15661.

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen R. McKeown, Doug Downey, and Yejin Choi. 2022. Penguins don't fly: Reasoning about generics through instantiations and exceptions. *CoRR*, abs/2205.11658.

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 V. Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. Exploring length generalization in large language models. *CoRR*, abs/2207.04901.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, LAW-ID@ACL 2013, August 8-9, 2013, Sofia, Bulgaria*, pages 178–186. The Association for Computer Linguistics.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2022. Prompting is programming: A query language for large language models. *arXiv preprint arXiv:2212.06094*.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 12317–12325. AAAI Press.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Gerhard Brewka. 1987. The logic of inheritance in frame systems. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence. Milan, Italy, August 23-28, 1987*, pages 483–488. Morgan Kaufmann.

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*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022a. E-KAR: A benchmark for rationalizing natural language analogical

[8]https://github.com/zjunlp/Prompt4ReasoningPapers

reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3941–3955. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

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

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

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2022. Binding language models in symbolic languages. *CoRR*, abs/2210.02875.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, 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. *CoRR*, abs/1803.05457.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

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

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

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

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 1173–1178. Association for Computational Linguistics.

Travis DeWolf. 2021. Spiking neural networks take control. *Sci. Robotics*, 6(58).

Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2022. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *CoRR*, abs/2203.06904.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-Dickstein, Kevin Murphy, and Charles Sutton. 2022. Language model cascades. *CoRR*, abs/2207.10342.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2021. Excar: Event graph knowledge enhanced explainable causal reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2354–2363. Association for Computational Linguistics.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5436–5443. ijcai.org.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Duan, Duyu Tang, and Ming Zhou. 2020. Machine reasoning: Technology, dilemma and future. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, EMNLP 2020, Online, November 19-20, 2020*, pages 1–6. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Trans. Assoc. Comput. Linguistics*, 10:1138–1158.

Yufei Feng, Xiaoyu Yang, Xiaodan Zhu, and Michael A. Greenspan. 2022. Neuro-symbolic natural logic with introspective revision for natural language inference. *Trans. Assoc. Comput. Linguistics*, 10:240–256.

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

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

Francesco Gardin and Bernard Meltzer. 1989. Analogical representations of naive physics. *Artif. Intell.*, 38(2):139–159.

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.

Matthew L Ginsberg. 1987. Readings in nonmonotonic reasoning.

Andrew S. Gordon and Jerry R. Hobbs. 2004. Formalizations of commonsense psychology. *AI Mag.*, 25(4):49–62.

Zhen Guo, Zelin Wan, Qisheng Zhang, Xujiang Zhao, Feng Chen, Jin-Hee Cho, Qi Zhang, Lance M. Kaplan, Dong H. Jeong, and Audun Jøsang. 2022. A survey on uncertainty reasoning and quantification for decision making: Belief theory meets deep learning. *CoRR*, abs/2206.05675.

Kyle Hamilton, Aparna Nayak, Bojan Bozic, and Luca Longo. 2022. Is neuro-symbolic AI meeting its promise in natural language processing? A structured review. *CoRR*, abs/2202.12205.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chadi Helwe, Chloé Clavel, and Fabian Suchanek. 2022. Logitorch: A pytorch-based library for logical reasoning on natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

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 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing NLP with cognitive language processing signals. *CoRR*, abs/1904.02682.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533, Doha, Qatar. Association for Computational Linguistics.

Danqing Huang, Shuming Shi, Chin-Yew Lin, Jian Yin, and Wei-Ying Ma. 2016. How well do computers solve math word problems? large-scale dataset construction and evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 887–896, Berlin, Germany. Association for Computational Linguistics.

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

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

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. *CoRR*, abs/2205.11822.

Mayank Kejriwal, Henrique Santos, Alice M. Mulvehill, and Deborah L. McGuinness. 2022. Designing a strong test for measuring true common-sense reasoning. *Nat. Mach. Intell.*, 4(4):318–322.

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. *CoRR*, abs/2210.02406.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *CoRR*, abs/2205.11916.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing Algebraic Word Problems into Equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.

Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S. Bernstein. 2022. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119(39):e2115730119.

Nate Kushman, Luke Zettlemoyer, Regina Barzilay, and Yoav Artzi. 2014. Learning to automatically solve algebra word problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 271–281. The Association for Computer Linguistics.

Brenden M. Lake and Marco Baroni. 2017. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *CoRR*, abs/1711.00350.

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. 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? *CoRR*, abs/2204.02329.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V. Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam

Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *CoRR*, abs/2206.14858.

Bei Li, Quan Du, Tao Zhou, Yi Jing, Shuhan Zhou, Xin Zeng, Tong Xiao, JingBo Zhu, Xuebo Liu, and Min Zhang. 2022a. ODE transformer: An ordinary differential equation-inspired model for sequence generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8335–8351. Association for Computational Linguistics.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022b. A survey on retrieval-augmented text generation. *CoRR*, abs/2202.01110.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022c. Explanations from large language models make small reasoners better. *CoRR*, abs/2210.06726.

Xiao Li, Gong Cheng, Ziheng Chen, Yawei Sun, and Yuzhong Qu. 2022d. Adalogn: Adaptive logic graph network for reasoning-based machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7147–7161. Association for Computational Linguistics.

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

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. 2022a. Reasoning over different types of knowledge graphs: Static, temporal and multi-modal. *arXiv preprint arXiv:2212.05767*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022b. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017a. 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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 158–167. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017b. 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.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022a. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. 2022b. Rainier: Reinforced knowledge introspector for commonsense question answering. *CoRR*, abs/2210.03078.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022c. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3154–3169. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022d. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, abs/2107.13586.

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M. Dai. 2022e. Mind's eye: Grounded language model reasoning through simulation. *CoRR*, abs/2210.05359.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *CoRR*, abs/2209.09513.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *CoRR*, abs/2209.14610.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022c. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *CoRR*, abs/2210.07128.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabsa. 2022. Unipelt: A unified framework for parameter-efficient language model tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6253–6264. Association for Computational Linguistics.

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.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *CoRR*, abs/2202.12837.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022a. Lila: A unified benchmark for mathematical reasoning. *CoRR*, abs/2210.17517.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022b. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland. Association for Computational Linguistics.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. Investigating the limitations of the transformers with simple arithmetic tasks. *CoRR*, abs/2102.13019.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt/`.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2021. Fact-driven logical reasoning. *CoRR*, abs/2105.10334.

Vaishali Pal, Evangelos Kanoulas, and Maarten de Rijke. 2022. Parameter-efficient abstractive question answering over tables or text. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, DialDoc@ACL 2022, Dublin, Ireland, May 26, 2022*, pages 41–53. Association for Computational Linguistics.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4179–4192. Association for Computational Linguistics.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pages 508–524. Springer.

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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

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

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *CoRR*, abs/2003.08271.

Dheeraj Rajagopal, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, Andrew E. Fano, and Eduard H. Hovy. 2021. Cross-domain reasoning via template filling. *CoRR*, abs/2111.00539.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4035–4045. Association for Computational Linguistics.

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.

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about Quantities in Natural Language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2655–2671. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *EMNLP*, abs/2210.13312.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *CoRR*, abs/2210.03057.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *CoRR*, abs/1904.01172.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2022. Selective annotation makes language models better few-shot learners. *CoRR*, abs/2209.01975.

Yueqing Sun, Yu Zhang, Le Qi, and Qi Shi. 2022. TSGP: two-stage generative prompting for unsupervised commonsense question answering. *CoRR*, abs/2211.13515.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*, pages 5026–5033. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Boshi Wang, Xiang Deng, and Huan Sun. 2022a. Iteratively prompt pre-trained language models for chain of thought.

Boshi Wang, Xiang Deng, and Huan Sun. 2022b. Shepherd pre-trained language models to develop a train of thought: An iterative prompting approach. *CoRR*, abs/2203.08383.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022c. PINTO: faithful language reasoning using prompt-generated rationales. *CoRR*, abs/2211.01562.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. 2021. From LSAT: the progress and challenges of complex reasoning. *CoRR*, abs/2108.00648.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022d. Finding skill neurons in pre-trained transformer-based language models. *CoRR*, abs/2211.07349.

Xingyao Wang, Sha Li, and Heng Ji. 2022e. Code4struct: Code generation for few-shot structured prediction from natural language. *CoRR*, abs/2210.12810.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2022f. Rationale-augmented ensembles in language models. *CoRR*, abs/2207.00747.

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

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854, Copenhagen, Denmark. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2300–2344. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *CoRR*, abs/2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, December 6-14, 2022, virtual*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark O. 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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 632–658. Association for Computational Linguistics.

Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J. Cai. 2022a. Promptchainer: Chaining large language model prompts through visual programming.

In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022, Extended Abstracts*, pages 359:1–359:10. ACM.

Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022b. Autoformalization with large language models. *CoRR*, abs/2205.12615.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Frank F. Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *MAPS@PLDI 2022: 6th ACM SIGPLAN International Symposium on Machine Programming, San Diego, CA, USA, 13 June 2022*, pages 1–10. ACM.

Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. 2021. A survey on green deep learning. *CoRR*, abs/2111.05193.

Jingfeng Yang, Haoming Jiang, Qingyu Yin, Danqing Zhang, Bing Yin, and Diyi Yang. 2022a. SEQZERO: few-shot compositional semantic parsing with sequential prompts and zero-shot models. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 49–60. Association for Computational Linguistics.

Zhicheng Yang, Jinghui Qin, Jiaqi Chen, Liang Lin, and Xiaodan Liang. 2022b. Logicsolver: Towards interpretable math word problem solving with logical prompt-enhanced learning. *CoRR*, abs/2205.08232.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot in-context learning. *CoRR*, abs/2205.03401.

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

Weizhe Yuan and Pengfei Liu. 2022. restructured pretraining. *CoRR*, abs/2206.11147.

Eric Zelikman, Yuhuai Wu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *CoRR*, abs/2203.14465.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. 2022a. Multimodal analogical reasoning over knowledge graphs.

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

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

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

Appendix

A Related Survey

As this area is relatively nascent, only a few surveys exist. Closest to our work, Bhargava and Ng (2022) covers methods for commonsense knowledge reasoning and generation with pre-trained LMs. Liang et al. (2022a) surveys knowledge graph reasoning tracing from static to temporal and then to multi-modal knowledge graphs. Hamilton et al. (2022) conducts a survey of studies implementing neural-symbolic (NeSy) NLP approaches for reasoning and so on. Guo et al. (2022) provides a survey of several popular works dealing with uncertainty reasoning. Other surveys focusing on prompt learning (Liu et al., 2022d) or pre-trained models (Qiu et al., 2020; Du et al., 2022) are also related to our work. Unlike those surveys, in this paper, we conduct a review of reasoning with LM prompting, hoping to systematically understand the methodologies, compare different methods and inspire new ideas.

B Performance Comparison of LMs with Different Scales

To show the generalization of discussions in §4.1 on different reasoning tasks, we additionally display the performance comparison of LMs with different scales on CommonsenseQA (Talmor et al., 2019) of commonsense reasoning in Figure 8.

C Detailed Information of Reasoning Benchmarks

In § 5.1, we give a brief overview on benchmarks and tasks requiring various reasoning skills. We list more benchmarks and show their key statistics in Table 2.

Task	Dataset	Size			
		Train	Valid	Test	All
Arithmetic Reasoning	AddSub (Hosseini et al., 2014)	395	-	-	395
	SingleOp (Roy et al., 2015)	562	-	-	562
	SingleEq (Koncel-Kedziorski et al., 2015)	508	-	-	508
	MultiArith (Roy and Roth, 2015)	600	-	-	600
	Dophin18k (Huang et al., 2016)	18,460	-	-	18,460
	MAWPS (Koncel-Kedziorski et al., 2016)	1,921	-	-	1,921
	Math23k (Wang et al., 2017)	23,161	-	-	23,161
	AQUA-RAT (Ling et al., 2017b)	97,467	-	254	97,721
	MathQA (Amini et al., 2019)	29,807	4,471	2,981	37,259
	DROP (Dua et al., 2019)	5,850	-	-	5,850
	ASDiv (Miao et al., 2020)	1,217	-	-	1,217
	GSM8K (Cobbe et al., 2021)	7,473	-	1,319	8,792
	SVAMP (Patel et al., 2021)	1,000	-	-	1,000
	MATH (Hendrycks et al., 2021)	7,500	-	5,000	12,500
	NumGLUE (Mishra et al., 2022b)	101,835	-	-	101,835
	Lila (Mishra et al., 2022a)	133,815	-	-	133,815
Symbolic Reasoning	Last Letter Concatenation (Wei et al., 2022b)	-	-	-	-
	Coin Flip (Wei et al., 2022b)	-	-	-	-
	Reverse List (Wei et al., 2022b)	-	-	-	-
Commonsense Reasoning	ARC (Clark et al., 2018)	3,370	869	3,548	7,787
	OpenBookQA (Mihaylov et al., 2018)	4,957	500	500	5,957
	CommonsenseQA (Talmor et al., 2019)	9,741	1,221	1,140	12,102
	PIQA (Bisk et al., 2020)	16,000	2,000	3,000	21,000
	StrategyQA (Geva et al., 2021)	2,290	-	490	2,780
Logical Reasoning	RuleTaker (Clark et al., 2020)	14,135	2,019	3,038	20,192
	ProofWriter (Tafjord et al., 2021)	-	-	-	-
	EntailmentBank (Dalvi et al., 2021)	1,313	187	340	1,840
	CLUTRR (Sinha et al., 2019)	6,016	-	-	6,016
Multimodal Reasoning	VCR(Zellers et al., 2019)	212,923	26,534	25,263	264,720
	Visual Commonsense Graphs (Park et al., 2020)	1,174,063	146,332	145,309	1,465,704
	ScienseQA (Lu et al., 2022a)	12,726	4,241	4,241	21,208

Table 2: An overview of benchmarks and tasks on reasoning.

D Reasoning with ChatGPT

Recently, ChatGPT (OpenAI, 2022) is an AI chatbot system released in November, which has attracted tremendous users. It is trained based on Reinforcement Learning from Human Feedback (Ouyang et al., 2022). The backbone of ChatGPT is from a model in the GPT-3.5 large LM series[9]. In order to savor the reasoning ability of large LMs more realistically, we conduct some case tests on ChatGPT. Concretely, we pick out a piece of data from GSM8K (Cobbe et al., 2021), CommonsenseQA (Talmor et al., 2019) and Last Letter Concatenation (Wei et al., 2022b) which respectively represent arithmetic reasoning, commonsense reasoning, and symbolic reasoning. Then we test each of the selected data on ChatGPT directly. Results can be seen in Figure 9-11.

Figure 9 shows that given a math problem in GSM8K (Cobbe et al., 2021), ChatGPT outputs a reasoning process and a correct answer without in-context exemplars. This blazes its powerful arithmetic reasoning ability. The reasoning process has the same format as the gold label in GSM8K, indicating that GSM8K may be contained in the training corpus of ChatGPT.

In Figure 10, we test ChatGPT on a piece of data in CommsonsenseQA (Talmor et al., 2019). It not only gives the correct answer but additionally details why each option is right or wrong, which does not appear in the gold label of the dataset. This demonstrates the strong commonsense reasoning ability of ChatGPT.

Figure 11 is a case in Last Letter Concatenation (Wei et al., 2022b). We observe that although ChatGPT gives a detailed and accurate description of last letter concatenation, it fails to answer the given question, showing that its symbolic reasoning capability is not as excellent as the above two.

[9] https://beta.openai.com/docs/model-index-for-researchers

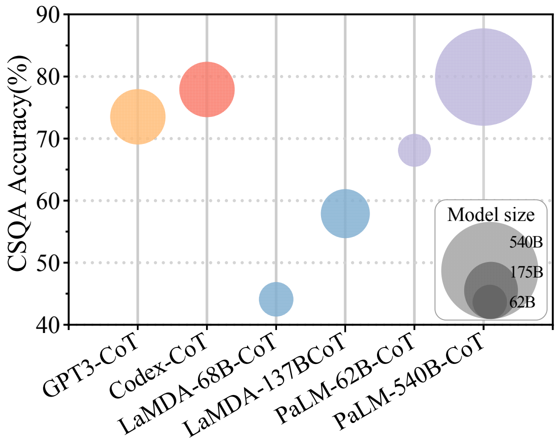

Figure 8: Performance of different language model scales on commonsense reasoning. Representatively, We show CoT (Wei et al., 2022b) experimental results on CommonsenseQA (Talmor et al., 2019).

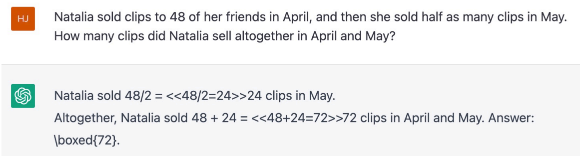

Figure 9: A test case from GSM8K (Cobbe et al., 2021) on ChatGPT (OpenAI, 2022).

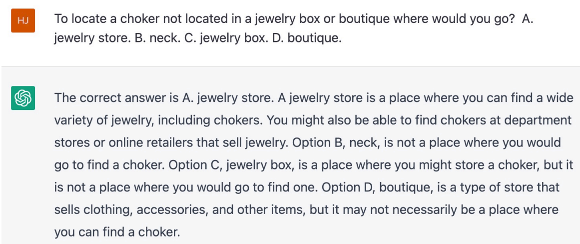

Figure 10: A test case from CommonsenseQA (Talmor et al., 2019) on ChatGPT (OpenAI, 2022).

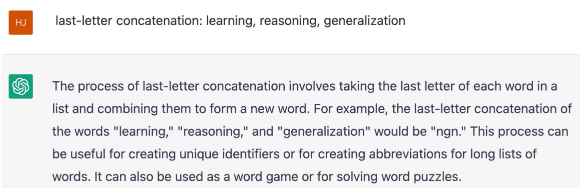

Figure 11: A test case from Last Letter Concatenation (Wei et al., 2022b) on ChatGPT (OpenAI, 2022).