Navigate through Enigmatic Labyrinth
A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future

Zheng Chu[1*], **Jingchang Chen**[1*], **Qianglong Chen**[2*], **Weijiang Yu**[2], **Tao He**[1]
Haotian Wang[1], **Weihua Peng**[2], **Ming Liu**[1,3†], **Bing Qin**[1,3], **Ting Liu**[1]
[1]Harbin Institute of Technology, Harbin, China
[2]Huawei Inc., Shenzhen, China
[3]Peng Cheng Laboratory, Shenzhen, China
`{zchu,jcchen,mliu}@ir.hit.edu.cn, chenqianglong.ai@gmail.com`

Abstract

Reasoning, a fundamental cognitive process integral to human intelligence, has garnered substantial interest within artificial intelligence. Notably, recent studies have revealed that chain-of-thought prompting significantly enhances LLM's reasoning capabilities, which attracts widespread attention from both academics and industry. In this paper, we systematically investigate relevant research, summarizing advanced methods through a meticulous taxonomy that offers novel perspectives. Moreover, we delve into the current frontiers and delineate the challenges and future directions, thereby shedding light on future research. Furthermore, we engage in a discussion about open questions. We hope this paper serves as an introduction for beginners and fosters future research. Resources have been made publicly available at https://github.com/zchuz/CoT-Reasoning-Survey.

1 Introduction

In the realm of human cognition, reasoning stands as the linchpin, essential in the understanding of the world and the formation of our decisions. As the scale of pre-training continues to expand (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a,b), large language models (LLMs) exhibit growing capabilities in numerous downstream tasks (Wei et al., 2022a; Schaeffer et al., 2023; Zhou et al., 2023c). Recently, researchers have discovered that LLMs emerge with the capability for step-by-step reasoning through in-context learning, a phenomenon referred to as chain-of-thought (CoT) reasoning. It is broadly observed that CoT prompting significantly boosts the reasoning abilities of LLMs, especially in complex tasks (Wei et al., 2022b; Cobbe et al., 2021; Geva et al., 2021).

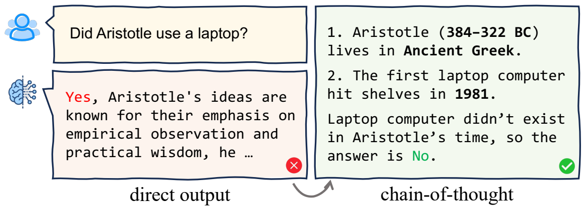

Figure 1: The model tackles complex problems step-by-step under the guidance of chain-of-thought prompting.

Figure 1 illustrates an example of chain-of-thought reasoning. Rather than directly providing the answer, chain-of-thought reasoning offers a step-by-step reasoning trajectory. Specifically, it decomposes intricate problems into manageable steps (*thoughts*), simplifying the overall reasoning process, and creates a linkage (*chain*) among the reasoning steps to ensure no important conditions are overlooked. Additionally, chain-of-thought reasoning offers an observable reasoning process, allowing users to comprehend the model's decision-making trajectory and increase the trustworthiness and interpretability of the final answer.

Benefiting from the remarkable performance of CoT prompting, it has attracted widespread attention across both academia and industry, evolving into a distinct research branch within the field of prompt engineering (Liu et al., 2023d; Qiao et al., 2023). Moreover, it has emerged as a crucial component in the landscape of AI autonomous agents (Wang et al., 2023h; Xi et al., 2023). However, these studies still lack a systematic review and analysis. To fill this gap, we propose this work to conduct a comprehensive and detailed analysis of CoT reasoning. Specifically, this paper delves into the broader scope of chain-of-thought reasoning, which we refer to as generalized chain-of-thought (XoT). The core philosophy of XoT reasoning is the gradual unraveling of complex problems via a step-by-step reasoning approach.

[*] Equal Contribution.
[†] Corresponding Author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1173–1203
August 11-16, 2024 ©2024 Association for Computational Linguistics

Our contributions can be summarized as follows: (1) ***Comprehensive Survey***: This is the first comprehensive survey dedicated for XoT reasoning; (2) ***Meticulous taxonomy***: We introduce a meticulous taxonomy (shown in Figure 2); (3) ***Frontier and Future***: We discuss new frontiers, outline their challenges, and shed light on future research. (4) ***Resources***: We make the resources publicly available to facilitate the research community.

Survey Organization We first give background and preliminary (§2); then present benchmarks (§3) and advanced methods (§4) from different perspectives. Furthermore, we discuss frontier research (§5), and outline challenges as well as future directions (§6). Finally, we give a further discussion about open questions (§A.2).

2 Background and Preliminary

2.1 Background

Over the past few years, as the scale of pre-training continuously increases (Brown et al., 2020; Scao et al., 2022; Touvron et al., 2023b; Zhao et al., 2023b), language models have emerged with numerous new capabilities, such as in-context learning (Wei et al., 2022a; Brown et al., 2020) and chain-of-thought reasoning (Wei et al., 2022b). Accompanying this trend, pre-training then prompting has gradually replaced pre-training then fine-tuning as the new paradigm in natural language processing (Qiu et al., 2020; Zhao et al., 2023b).

2.2 Preliminary

In this section, we provide the preliminary for standard prompting and chain-of-thought reasoning. Referring to Qiao et al. (2023), we define the notations as follows: question $\mathcal{Q}$, prompt $\mathcal{T}$, probabilistic language model p_{LM} and prediction $\mathcal{A}$.

First, we consider the few-shot standard prompting scenario, where prompt $\mathcal{T}_{SP}$ includes instruction I and few-shot demonstrations (several question-answer pairs). The model takes the question and prompt as inputs and produces the answer prediction $\mathcal{A}$ as its output, as shown in Equ. (1,2).

$$\mathcal{T}_{SP} = \{I, (x_1, y_1), \cdots, (x_n, y_n)\} \quad (1)$$

$$p(\mathcal{A} \mid \mathcal{T}, \mathcal{Q}) = \prod_{i=1}^{|\mathcal{A}|} p_{LM}(a_i \mid \mathcal{T}, \mathcal{Q}, a_{<i}) \quad (2)$$

Next, we consider chain-of-thought prompting under few-shot setting, wherein the prompt $\mathcal{T}_{CoT}$ includes instruction, questions, answers, and rationales e_i. In chain-of-thought reasoning, the model no longer directly generates answers. Instead, it generates step-by-step reasoning trajectories $\mathcal{R}$ before giving answers $\mathcal{A}$, as shown in Equ. (3,4,5,6).

$$\mathcal{T}_{\text{CoT}} = \{I, (x_1, e_1, y_1), \cdots, (x_n, e_n, y_n)\} \quad (3)$$

$$p(\mathcal{A}, \mathcal{R}|\mathcal{T}, \mathcal{Q}) = p(\mathcal{A}|\mathcal{T}, \mathcal{Q}, \mathcal{R}) \cdot p(\mathcal{R}|\mathcal{T}, \mathcal{Q}) \quad (4)$$

$$p(\mathcal{R} \mid \mathcal{T}, \mathcal{Q}) = \prod_{i=1}^{|\mathcal{R}|} p_{LM}(r_i \mid \mathcal{T}, \mathcal{Q}, r_{<i}) \quad (5)$$

$$p(\mathcal{A}|\mathcal{T}, \mathcal{Q}, \mathcal{R}) = \prod_{j=1}^{|\mathcal{A}|} p_{LM}(a_i|\mathcal{T}, \mathcal{Q}, \mathcal{R}, a_{<j}) \quad (6)$$

2.3 Advantages of CoT Reasoning

As a novel reasoning paradigm, chain-of-thought gains various advantages. (1) Boosted Reasoning. Chain-of-thought reasoning breaks down complex problems into manageable steps and establishes connections among these steps, thereby facilitating reasoning. (2) Offering Interpretability. Chain-of-thought reasoning provides observable reasoning traces, allowing the user to understand the model's decision, making the reasoning process transparent and trustworthy. (3) Advance Collaboration. Fine-grained reasoning traces facilitate user-system interaction, allowing for altering the model's execution trajectory, thereby fostering the development of autonomous agents powered by LLMs.

3 Benchmarks

In this section, we briefly outline the benchmarks for evaluating reasoning capabilities, including mathematical, commonsense, symbolic, logical, and multi-modal reasoning. The overview of benchmarks is shown in Table 1. For more details about benchmarks, please refer to Appendix B.

Mathematical Reasoning Mathematical reasoning forms the foundation of human intelligence, playing a crucial role in problem-solving, decision-making, and world comprehension. It is commonly used to assess the general reasoning ability of LLMs (Patel et al., 2021; Cobbe et al., 2021; Hendrycks et al., 2021b; Mishra et al., 2022a).

Commonsense Reasoning Commonsense reasoning is essential for the interaction in daily life and the perception of the world, which assesses the world comprehension capacity of language models (Talmor et al., 2019, 2021; Geva et al., 2021).

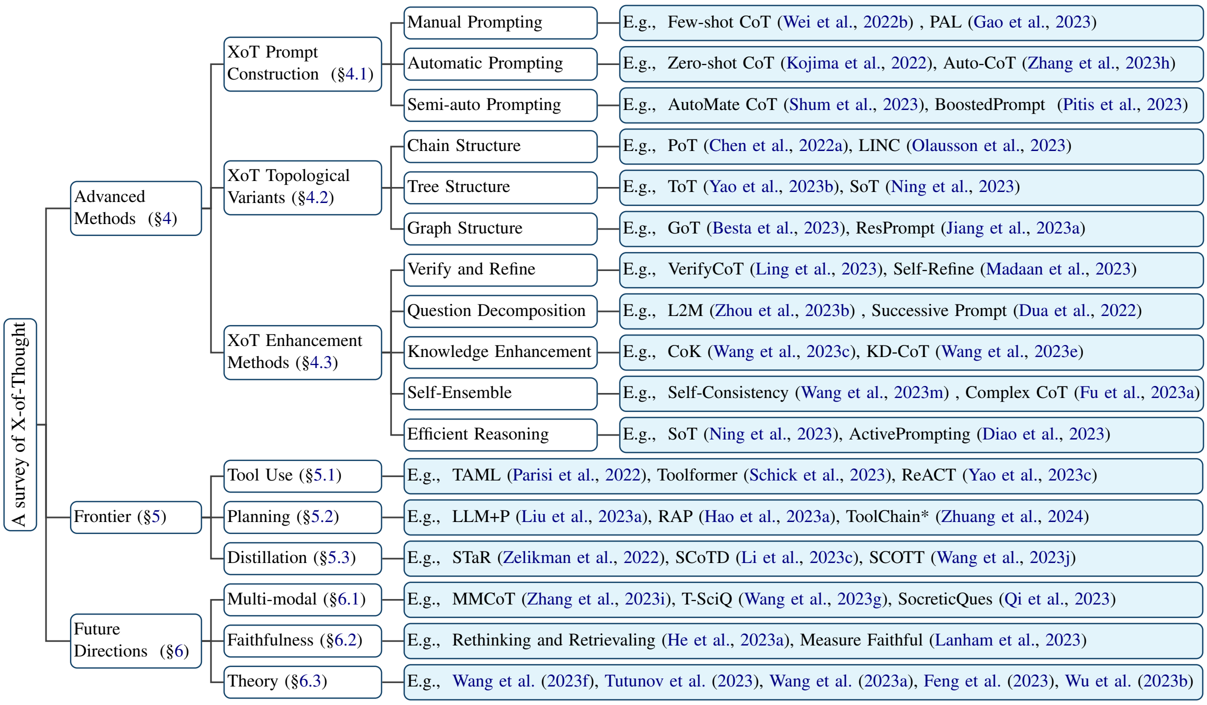

Figure 2: Taxonomy of Advanced Methods, Frontiers and Future Directions (Full version in Figure 8).

Symbolic Reasoning Symbolic reasoning disentangles semantics and serves as a testbed for language models' competence in simulating atomic operations (Wei et al., 2022b; Srivastava et al., 2022; Suzgun et al., 2023).

Logical Reasoning Logical reasoning is of paramount importance as it serves as the bedrock for rational thinking, robust problem-solving and interpretable decision-making (Liu et al., 2020; Yu et al., 2020; Tafjord et al., 2021; Han et al., 2022).

Multi-modal Reasoning Multimodal reasoning seamlessly integrates textual thought with sensory experiences from the natural world, such as visual scenes, and auditory sounds, to create a richer, more comprehensive understanding of information (Zellers et al., 2019; Park et al., 2020; Xiao et al., 2021; Lu et al., 2022; Chen et al., 2023c).

4 Advanced Methods

This section discusses advanced XoT methods from three viewpoints: prompt construction (§4.1), topological variations (§4.2), and enhancement methods (§4.3). The taxonomy is shown in Figure 2.

4.1 XoT Prompt Construction

Based on the human effort for constructing chain-of-thought prompting, we divide the construction approaches into three categories: 1) Manual XoT, 2) Automatic XoT, and 3) Semi-automatic XoT.

4.1.1 Manual Prompting

Wei et al. (2022b) first proposes chain-of-thought prompting (Fewshot CoT) by manually annotating natural language form rationales to guide models in stepwise reasoning. Moreover, Fu et al. (2023a) discovers that using complex reasoning chains as demonstrations can further improve reasoning performance. Yet, the NL form reasoning encounters inconsistent reasoning. To mitigate intermediate errors in reasoning, PAL (Gao et al., 2023), PoT (Chen et al., 2022a), MathPrompter (Imani et al., 2023) and NLEP (Zhang et al., 2023d) leverage rationales in programming language form, transforming problem-solving into program generation, and obtaining a deterministic answer through external program executor. Although manual XoT demonstrates better performance, the annotation of rationales incurs a significant increase in cost and introduces dilemmas in demonstration selection.

4.1.2 Automatic Prompting

Some work designs specific instructions to stimulate CoT reasoning under zero-shot, such as appending *Let's think step by step* after questions (Kojima et al., 2022). There are also other types of instructions, including writing programs to solve problems (Chen et al., 2022a), drafting plans before reasoning (Wang et al., 2023i), generating meta instructions based on task information (Crispino et al., 2023) and role playing (Kong et al., 2023a).

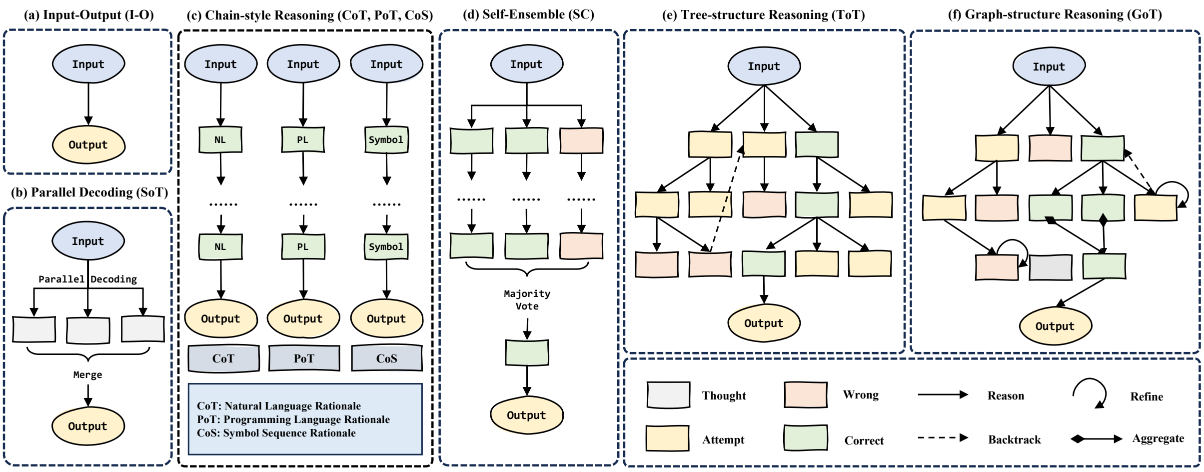

Figure 3: Topological variants emerging in the evolution of XoT. (a) standard I-O prompting, (b) parallel-constrained tree structure variants, (c) chain structure variants with distinct rationale descriptions, (d) chain structure variants with self-ensemble, (e) standard tree structure variants, and (f) standard graph structure variants.

However, due to the lack of guidance from clearly defined demonstrations, instruction-based methods appear extremely unstable. Another route of work conducts few-shot reasoning based on automatically generated rationales (usually by zero-shot CoT), which improves the stability of reasoning. These methods focus on selecting appropriate demonstrations. Zhang et al. (2023h) chooses diverse rationales through clustering, Zou et al. (2023) constructs demonstrations based on the question pattern, improving the generalization, Wan et al. (2023) employs answer entropy as a metric for selection, and Xu et al. (2023) uses Gibbs sampling to iteratively select demonstrations.

4.1.3 Semi-automatic Prompting

Building upon automatic XoT based on few-shot learning, semi-automatic approaches incorporate a small number of human-annotated rationales to obtain supervised signals. They focus on bootstrapping to acquire high-quality rationales and selecting appropriate demonstrations to facilitate reasoning. Shao et al. (2023b) generates high-quality rationales through alternating forward and backward synthetic processes, and Pitis et al. (2023) iteratively expands the examples when encountering challenging questions, which mitigates the issue of limited human supervision. On the other hand, some studies optimize demonstration selection. Shum et al. (2023) and Lu et al. (2023b) utilize policy gradient optimization to learn demonstration selection strategy, while Ye and Durrett (2023) searches the development set and selects proper demonstration using two proxy metrics.

4.1.4 Pros and Cons of Three Approaches

Manual prompting relies on high-quality rationale annotations, which result in better performance. However, it encounters drawbacks such as high labor costs and challenges in domain transfer. In contrast, automatic prompting incurs no labor costs and facilitates free domain transfer. However, it is plagued by errors and instability due to the absence of supervised signals. Semi-automatic prompting strikes a dedicated balance, achieving a trade-off between performance and costs, making it more suitable for downstream applications.

4.2 XoT Topological Variants

The evolution of XoT has led to the development of multiple topological variants[1]. In this section, we will delve into topological variants of XoT: chain structure, tree structure and graph structure.

Chain Structure The description format of rationales significantly influences reasoning execution. PAL (Gao et al., 2023) and PoT (Chen et al., 2022a) use programming languages to depict the reasoning process, transforming problem-solving into code generation. Similarly, formal logic description languages are also used to depict logical reasoning (Olausson et al., 2023; Pan et al., 2023; Ye et al., 2023a). The aforementioned methods decouple the thought generation from execution, thereby eliminating inconsistency reasoning errors. Additionally, algorithmic descriptions (Sel et al., 2023) can offer a high-level reasoning framework

[1]We consider XoT with chain structure and natural language rationales as vanilla CoT (the most primitive one).

instead of details, endowing the model with the ability for global thinking.

Tree Structure Chain structure inherently limits the scope of exploration. Through the incorporation of tree structures and search algorithms, models gain the capability to widely explore and backtrack during reasoning (Long, 2023; Yao et al., 2023b), as shown in Figure 3(e). Chen et al. (2024) iteratively explores and evaluates multiple tree-of-thoughts to further enhance reasoning. Benefiting from the exploration, tree variants have gained preliminary global planning capabilities towards the global optimum. Meanwhile, Mo and Xin (2023); Cao et al. (2023) introduce uncertainty measurement based on Monte Carlo dropout and generation likelihood, respectively, thereby offering a more accurate evaluation of intermediate reasoning processes. Yu et al. (2024) uses a bottom-up approach by building an analogy sub-problems tree. In addition, Ning et al. (2023) initially delivers reasoning drafts, accelerating reasoning by solving tree structure sub-problems in parallel. However, tree-based methods are restricted by demands of explicit question decomposition and state transition, which leads to limitations in task generalization.

Graph Structure Graph structures introduce loops and N-to-1 connections, enabling improved modeling of sub-problem aggregation and self-verification (Besta et al., 2023; Lei et al., 2023a), as illustrated in Figure 3(f). Graph structures outperform tree-based methods in handling complex problems. However, they rely on specially designed state decomposition, leading to poorer generalization. To address this, Jiang et al. (2023a) establishes an implicit graph upon the reasoning process through prompts, avoiding the constraints of explicit topological structures, thereby generalizing to various multi-step reasoning tasks.

The complex topological structure introduces a fine control flow, which facilitates LLMs in tackling harder problems. However, this complexity also limits the application of these methods in general reasoning, posing a significant challenge that needs to be addressed in future research.

4.3 XoT Enhancement Methods

This section introduces five enhanced XoT reasoning approaches, including verify and refine (§4.3.1), question decomposition (§4.3.2), knowledge enhancement (§4.3.3), self-ensemble (§4.3.4) and efficient reasoning (§4.3.5).

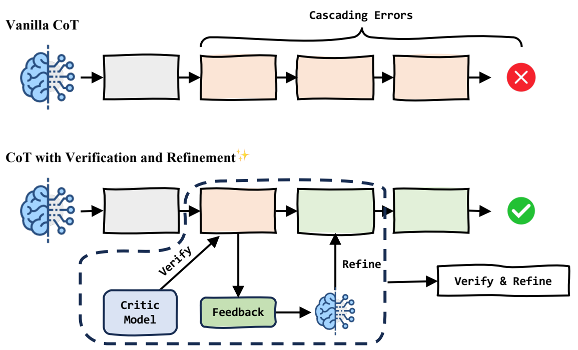

Figure 4: Verification and refinement rectify intermediate errors, which reduce cascading errors in reasoning.

4.3.1 Verify and Refine

LLMs tend to hallucinate, which manifests as factual and faithful errors in reasoning (Huang et al., 2023b). Incorporating verification and refinement can be an effective strategy for mitigating the phenomena. In this section, we primarily focus on mitigating faithful errors, with a separate discussion of factual errors in the following knowledge enhancement section (§4.3.3).

Reasoning can be refined based on critical feedback provided by LLMs. Paul et al. (2024a) trains a small critic model to provide structured feedback, but the quality of the feedback is limited due to the model size. Madaan et al. (2023) employs feedback from itself for iterative self-refinement, Li et al. (2023g) uses finer-grained feedback at the step level, and Shinn et al. (2023) further expands this method by incorporating long and short-term memory to provide more concise feedback. However, recent research suggests that LLMs may not address issues beyond their own capabilities (Kadavath et al., 2022; Yin et al., 2023), which raises doubt on the effectiveness of self-feedback (Huang et al., 2024a). To remedy this, some work incorporates external feedback (Gou et al., 2024a; Nathani et al., 2023) or performs secondary verification on the refined reasoning (Shridhar et al., 2023).

On the other hand, logical reasoning structures are also well-suited for verification. Ling et al. (2023) devises a deductive reasoning form named Natural Program, which guarantees that the conclusion is derived from the designated premises. Wu et al. (2024) applies a deductive filter to verify the entailment relationship between question and reasoning chains. Some studies perform step-wise verification during the beam search decoding stage. Xie et al. (2023) uses the log-probabilities of deduc-

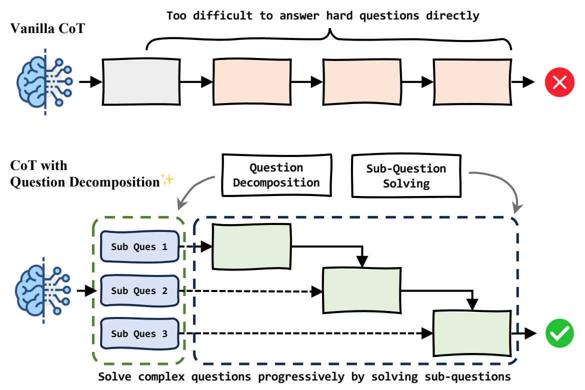

Figure 5: Question decomposition solves complex questions progressively by solving simple sub-questions.

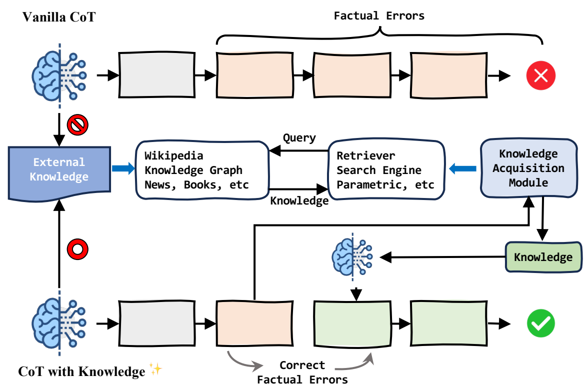

Figure 6: Incorporating knowledge (either internal or external) helps mitigate factual errors in reasoning.

tive reasoning as a search criterion, while Zhu et al. (2024a) trains a deductive discriminator for verification. Besides, backward (abductive) reasoning excels in detecting inconsistencies in reasoning. It reconstructs conditions or variables in the question based on the reasoning chain to discover inconsistencies, thereby refining the reasoning (Xue et al., 2023; Weng et al., 2022; Jiang et al., 2023b).

Reasoning with LLMs is prone to hallucinations, and feedback from intermediate steps plays a crucial role in refining the reasoning. However, the current acquisition of feedback signals still has many shortcomings, which necessitates further research.

4.3.2 Question Decomposition

The philosophy of XoT is to solve questions step-by-step. However, vanilla CoT does not explicitly decompose questions, making it challenging to answer complex questions. To address this, certain approaches address intricate problems by progressively tackling straightforward sub-problems.

L2M (Zhou et al., 2023b) initially breaks down the question into sub-questions in a top-down fashion. It then solves one sub-question at a time and leverages its solution to facilitate subsequent sub-questions. Dua et al. (2022) takes a similar approach to L2M, but it uses solutions from previous sub-questions to iteratively decompose questions. Khot et al. (2023) designs a modular task-sharing library that tailors more effective solutions to different classes of sub-questions. Huang et al. (2024b) breaks down the problem into a directed acyclic graph represented by QDMR, and then performs step-wise reasoning based on the graph dependencies. In multi-hop reasoning, iterative decomposition has become a common practice (Wang et al., 2022; Press et al., 2023; Trivedi et al., 2023). Additionally, some methods obtain a dedicated decomposer through supervised training rather than relying on the LLM itself (Li et al., 2023f; Junbing et al., 2023). However, when dealing with tabular reasoning, answering sub-questions may also pose a challenge, particularly when handling large tables. To tackle this issue, certain approaches involve decomposing both the questions and tables simultaneously (Ye et al., 2023b; Cheng et al., 2023; Nahid and Rafiei, 2024).

Bottom-up aggregation is also a viable solution, with a smaller exploration space. Qi et al. (2023) employs Socratic questioning for recursive self-questing to solve complex questions, while Zhang et al. (2024), in a similar fashion, breaks down the conditions of complex problems into small components and resolves them bottom-up.

It should be noted that both decomposition and aggregation are highly dependent on the proper problem division, and reversely, a misaligned division may yield counterproductive results.

4.3.3 Knowledge Enhancement

When dealing with knowledge-sensitive tasks, LLMs often make factual errors. Introducing external knowledge or mining the model's internal knowledge can help alleviate this issue. Some methods explicitly utilize the model's intrinsic knowledge. For example, Dhuliawala et al. (2023); Ji et al. (2023); Zheng et al. (2024) prompt models to output its parametric knowledge, and then reason based on it. Additionally, Zhang et al. (2023f) prompts the model to perform inductive reasoning on its internal knowledge, deriving more general conclusions. Furthermore, Liu et al. (2023c) incorporates reinforcement learning to optimize introspective knowledge-grounded reasoning. Mean-

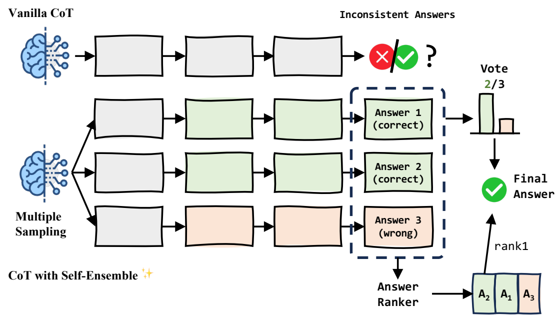

Figure 7: Self-ensemble reduces inconsistency by selecting final answers from multiple samplings.

while, Li and Qiu (2023) leverages model's reasoning traces to construct a memory base, selecting relevant demonstrations whenever needed.

External knowledge is often more reliable than parametric knowledge. Li et al. (2023f); Wang et al. (2023e) generates queries based on the question, utilizing a knowledge base as the external knowledge. Building upon this, Wang et al. (2023c) introduces a verification step for the retrieved knowledge, further ensuring knowledge accuracy. However, when confronted with multi-hop reasoning, direct retrieval using the question can be insufficient. Therefore, Press et al. (2023); Trivedi et al. (2023); Shao et al. (2023a); Yoran et al. (2023) decompose the question and iteratively use sub-question for more precise retrieval.

4.3.4 Self-Ensemble

The sampling during generation introduces uncertainty, which in turn, creates the possibility of improving performance through self-ensemble. Cobbe et al. (2021) trains a verifier to rank answers, and Hu et al. (2024a) utilizes LLMs to self-rank their predictions. SC (Wang et al., 2023m) performs majority voting based on answers across multiple samples, and Fu et al. (2023a) proposes a complexity-based voting strategy on top of SC. Widespread practical evidence indicates that self-ensemble is an effective way to improve performance. However, answer-based ensemble fails to consider intermediate steps. In response, Miao et al. (2024); Yoran et al. (2023); Khalifa et al. (2023) refines the ensemble at the step level, and Yin et al. (2024) introduces hierarchical answer aggregation. Yet another concern is the limited diversity offered by probability sampling. To overcome this limitation, Naik et al. (2023) uses different instructions, Liu et al. (2023e) ensembles various XoT variants, and Qin et al. (2023) ensembles using multi-lingual reasoning chains. Besides, the multi-agent debate (MAD) framework can also be regarded as heterogeneous ensemblings (Liang et al., 2023; Du et al., 2023; Wang et al., 2023b).

Self-ensemble, as a simple yet effective means, has gained widespread favor. Nevertheless, alongside the improvement in performance, there has been a multiplied increase in inference costs, which in turn limits its wide application.

4.3.5 Efficient Reasoning

LLMs are often inefficient in reasoning, such as high latency, substantial annotation costs, and elevated inference costs. To speed up reasoning, Ning et al. (2023) decomposes the questions in parallel and handles them simultaneously, Zhang et al. (2023b) generates a draft to skip intermediate layers during inference, and Leviathan et al. (2023); Chen et al. (2023a) introduce speculative decoding, which employs a smaller model for faster inference. Diao et al. (2023) annotates high-uncertainty samples to reduce human costs, and Aggarwal et al. (2023) dynamically adjusts sampling frequency to reduce inference costs. Further research should focus on efficient reasoning to promote the widespread application of LLMs.

5 Frontiers of Research

5.1 Tool Use

LLMs face difficulties in accessing news, performing calculations, and interacting with the environment. Previous work endows LLMs with the ability to use external tools, enhancing their reasoning capabilities and enabling them to interact with the (multi-modal) external environment (Parisi et al., 2022; Schick et al., 2023; Shen et al., 2023a).

However, these methods have limitations in facilitating multiple tool invocations and rectifying query errors. To tackle this problem, ReAct (Yao et al., 2023c) and Reflexion (Shinn et al., 2023) integrate the strengths of reasoning and action to complement each other. ART (Paranjape et al., 2023) uses a task library to select relevant tools and reasoning demonstrations. MM-REACT (Yang et al., 2023b) further incorporates vision experts to facilitate multi-modal reasoning and action.

Above-mentioned studies focus on leveraging external tools to grant LLMs the capacities they initially lacked, thereby improving their performance across various domains. Tool invocation facilitates interaction with external sources, enabling it to gather additional information, while XoT enables effective elicitation, tracking, and action refining.

5.2 Planning

It is challenging for LLMs to provide accurate responses for complex goals, which requires planning to decompose them into sub-tasks and track the execution process. Plans can be described by code or definition languages. Sun et al. (2023) generates Python code to control the agent, and iteratively refine the plan based on the execution feedback. Liu et al. (2023a); Dagan et al. (2023) leverage the Planning Domain Definition Language (PDDL) (Gerevini, 2020) to describe the planning procedure. PDDL assists in decomposing complex problems and utilizing specialized models for planning before converting the results into natural languages. Zhou et al. (2023d) integrates self-refine (Madaan et al., 2023) with PDDL to achieve a better success rate in long-horizon sequential tasks.

Instead of pre-defined plans, many studies use search algorithms to dynamically plan and explore the action space. Tree-of-Thought explores the problem through DFS or BFS search, and tracks and updates the intermediate states (Yao et al., 2023b). RAP and LATS incorporate Monte Carlo Tree Search based on reasoning trajectories in planning (Hao et al., 2023a; Zhou et al., 2023a), and ToolChain* enables more efficient exploring through heuristic A* search (Zhuang et al., 2024).

LLMs, endowed with robust reasoning capabilities, can devise strategies for achieving complex goals. Furthermore, the integration of planning, reasoning, memory, and tool utilization serves as a cornerstone for LLM-powered autonomous agents.

5.3 Distillation of Reasoning Capabilities

In low-resource scenarios such as edge computing, distillation offers a possibility for deploying LLMs. Some methods employ self-distillation for self-improvement without external supervision. Huang et al. (2023a) employs self-consistency to generate reasoning chains from unlabeled data, followed by fine-tuning, enhancing its generalized reasoning capabilities. Zelikman et al. (2022) improves LM's reasoning capabilities via self-loop bootstrapping.

Despite the powerful reasoning exhibited by CoT, it emerges primarily in large-scale LLMs, with its usage limited in smaller models. Magister et al. (2023) finds that smaller models, after fine-tuning on CoT reasoning data, can also exhibit the capacity for step-by-step reasoning. Following this trend, numerous studies attempt to distill the step-by-step reasoning capabilities of LLMs into smaller models. Hsieh et al. (2023b) employs self-consistency to filter predictions, distilling high-quality reasoning chains from LLMs. Ho et al. (2023); Li et al. (2023c) find that sampling multiple reasoning chains per instance is paramount for improving students' reasoning capability. SCOTT (Wang et al., 2023j) utilizes contrastive decoding (Li et al., 2023e; O'Brien and Lewis, 2023) and counterfactual reasoning objective to tackle the shortcut problem. Li et al. (2024) improves the generalization of reasoning for unseen tasks through LoRA mixture-of-experts distillation.

Recent studies have found that the reasoning capabilities of small models can be further improved by optimizing over preference data. DialCoT (Han et al., 2023) decomposes reasoning steps into a multi-round dialog and optimizes the correct reasoning traces using PPO. Wang et al. (2023k); Feng et al. (2024) train a reward model on automatically generated data, which is designed to rank LLM's reasoning traces, and then optimizes smaller models using PPO. (Xie et al., 2024) utilizes Monte Carlo Tree Search to sample and score reasoning trajectories, generates preference data on the fly, and uses DPO for online preference optimization.

Since code serves as an excellent intermediate representation for reasoning, Zhu et al. (2023) distills program-aided reasoning capability into smaller models. Meanwhile, some studies find that distilling reasoning chains from both natural language and code formats leads to further improvement (Li et al., 2023a; Zhu et al., 2024b). In addition to regular reasoning, Yang et al. (2024a) attempts to distill tabular reasoning capabilities, and Zhao et al. (2024b) seeks to endow smaller models with retrieval-augmented reasoning capabilities.

These studies adopt a shared paradigm that distills smaller models with reasoning chains generated from larger models with superior reasoning capabilities. However, it is worth noting that language models have intricate tradeoffs associated with multi-dimensional capabilities, and distilling task-specific reasoning ability may adversely downgrade the general performance (Fu et al., 2023b).

6 Future Directions

Despite XoT reasoning has showcased remarkable performance on numerous tasks, there are still some challenges that necessitate further research.

6.1 Multi-modal Reasoning

Current XoT research mostly focuses on plain text. However, interacting with the real world necessitates multi-modal capabilities. To facilitate research, SciQA (Lu et al., 2022) and CURE (Chen et al., 2023c) are introduced to emphasize multi-modal CoT reasoning. Through fine-tuning with the combination of vision and language features, Zhang et al. (2023i); Wang et al. (2023g) endow models with multi-modal CoT reasoning capabilities, and Yao et al. (2023d,a) further incorporate graph structures to model multi-hop relationships. Other approaches convert images to captions and use LLM for prompt-based reasoning (Yang et al., 2023b; Zheng et al., 2023b). However, the limited capabilities of vision-language models constrain their performance in multi-step reasoning (Alayrac et al., 2022; Li et al., 2023b; Peng et al., 2023).

Several critical challenges remain to be addressed in future research, which we summarize as follows: (1) Vision-text interaction: How can visual and textual features be effectively integrated, than solely depending on captions? (2) Harnessing VLLMs: How can we better apply LLM-based reasoning techniques to the multi-modal domain? (3) Video Reasoning: How to expand into video reasoning with complex temporal dependencies?

6.2 Faithful Reasoning

Extensive research indicates that LLMs often engage in unfaithful reasoning, such as factual errors and inconsistent reasoning. To address factual errors, one common approach is retrieval augmentation (Trivedi et al., 2023; Zhao et al., 2023a), but it requires appropriate timing and retrieval accuracy. Compared to factual errors, inconsistencies are more difficult to identify (Paul et al., 2024b). Common detection methods include deductive logic (Jiang et al., 2023b; Xue et al., 2023; Ling et al., 2023), post-processing (He et al., 2023a; Lei et al., 2023b), and critic-based approaches (Madaan et al., 2023; Nathani et al., 2023). Among them, Neural-symbolic reasoning (Chen et al., 2022a; Olausson et al., 2023) is a widely used approach for reducing inconsistencies, and question decomposition (Radhakrishnan et al., 2023) has also demonstrated its effectiveness to some degree. Furthermore, Zhang et al. (2023c); Lanham et al. (2023) investigate the factors influencing faithfulness from an empirical perspective.

Faithful reasoning encounters two significant challenges: (1) Detection: How can unfaithful reasoning be accurately identified? (2) Correction: How can one obtain accurate feedback and make correct refinements based on that feedback?

6.3 Theoretical Perspective

The mechanism behind the CoT and ICL has not been clearly explained so far. Some studies empirically explore the roles of CoT and ICL in reasoning, offering practical insights (Wang et al., 2023a; Madaan and Yazdanbakhsh, 2022; Tang et al., 2023). Another line of work explores from a theoretical perspective. Li et al. (2023h); Feng et al. (2023); Merrill and Sabharwal (2023); Prystawski et al. (2023) investigate why CoT enhances reasoning abilities, while Wu et al. (2023b); Tutunov et al. (2023); Hou et al. (2023); Wang et al. (2023f) examine the mechanisms from a feature-based standpoint (information flow, attention, variables, etc.). Additionally, there have been preliminary explorations of the emergence mechanism (Schaeffer et al., 2023; Zhou et al., 2023c).

At present, the exploration of CoT theories is still limited to the surface level. There are still open questions that require further in-depth investigation. (1) How does the **emergence capability** arise? (2) **In what way** does CoT enhance reasoning compared to standard few-shot prompting?

7 Discussion

We delve into open questions about chain-of-thought reasoning, with the details discussion in Appendix A.2. The discussion encompasses three topics: (a) How does chain-of-thought reasoning ability emerge with large-scale pre-training? (b) How to provide accurate feedback for a model's reasoning and decision-making. (c) The implications of chain-of-thought reasoning for LLM-powered autonomous agents and AGI.

8 Conclusion

In this paper, we conduct a systematic survey of existing research on generalized chain-of-thought reasoning, offering a comprehensive review of the field. Specifically, we meticulously categorize advanced methods, delve into current frontier research, highlight existing challenges, identify potential future research directions, and discuss open questions. This paper is the first systematic survey dedicated to CoT reasoning. We hope that this survey will facilitate further research in this area.

Limitations

This study provides the first comprehensive survey of generalized chain-of-thought (XoT) reasoning. Related work, benchmarks details and further discussion can be found in Appendix A,B.

We have made our best effort, but there may still be some limitations. On one hand, due to page limitations, we can only provide a brief summary of each method without exhaustive technical details. On the other hand, we primarily collect studies from *ACL, NeurIPS, ICLR, ICML, COLING and arXiv, and there is a chance that we may have missed some important work published in other venues. In the benchmarks section, we primarily list widely used datasets, and more complete benchmarks can be found in Guo et al. (2023). As of now, there is no definitive conclusion on open questions. We will stay abreast of discussions within the research community, updating opinions and supplementing overlooked work in the future.

Acknowledgements

The research in this article is supported by the National Key Research and Development Project (2021YFF0901602), the National Science Foundation of China (U22B2059, 62276083), and Shenzhen Foundational Research Funding (JCYJ20200109113441941), Major Key Project of PCL (PCL2021A06). Ming Liu is the corresponding author.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. 2023. Let's sample step by step: Adaptive-consistency for efficient reasoning with llms. *ArXiv preprint*, abs/2305.11860.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

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.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, and Christopher Re. 2023. Ask me anything: A simple strategy for prompting language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of thoughts: Solving elaborate problems with large language models. *ArXiv preprint*, abs/2308.09687.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. Think you have solved direct-answer question answering? try arc-da, the direct-answer AI2 reasoning challenge. *ArXiv preprint*, abs/2102.03315.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

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*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *ArXiv preprint*, abs/2303.12712.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2024. Large language models as tool makers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

Shulin Cao, Jiajie Zhang, Jiaxin Shi, Xin Lv, Zijun Yao, Qi Tian, Lei Hou, and Juanzi Li. 2023. Probabilistic tree-of-thought reasoning for answering knowledge-intensive complex questions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12541–12560, Singapore. Association for Computational Linguistics.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023a. Accelerating large language model decoding with speculative sampling. *CoRR*, abs/2302.01318.

Sijia Chen, Baochun Li, and Di Niu. 2024. Boosting of thoughts: Trial-and-error problem solving with large language models. *CoRR*, abs/2402.11140.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023b. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, Singapore. Association for Computational Linguistics.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023c. Measuring and improving chain-of-thought reasoning in vision-language models. *ArXiv preprint*, abs/2309.04461.

Zhipeng Chen, Kun Zhou, Beichen Zhang, Zheng Gong, Xin Zhao, and Ji-Rong Wen. 2023d. ChatCoT: Tool-augmented chain-of-thought reasoning on chat-based large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14777–14790, Singapore. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022b. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. 2023. Binding language models in symbolic languages. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2023. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *ArXiv preprint*, abs/2311.17667.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv preprint*, abs/2110.14168.

Nicholas Crispino, Kyle Montgomery, Fankun Zeng, Dawn Song, and Chenguang Wang. 2023. Agent instructs large language models to be general zero-shot reasoners. *Preprint*, arXiv:2310.03710.

Gautier Dagan, Frank Keller, and Alex Lascarides. 2023. Dynamic planning with a llm. *ArXiv preprint*, abs/2308.06391.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2023. Implicit chain of thought reasoning via knowledge distillation. *Preprint*, arXiv:2311.01460.

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.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *ArXiv preprint*, abs/2309.11495.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *ArXiv preprint*, abs/2302.12246.

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. *ArXiv preprint*, abs/2207.10342.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *ArXiv preprint*, abs/2301.00234.

Qingxiu Dong, Ziwei Qin, Heming Xia, Tian Feng, Shoujie Tong, Haoran Meng, Lin Xu, Zhongyu Wei, Weidong Zhan, Baobao Chang, Sujian Li, Tianyu Liu, and Zhifang Sui. 2022. Premise-based multimodal reasoning: Conditional inference on joint textual and visual clues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 932–946, Dublin, Ireland. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *ArXiv preprint*, abs/2305.14325.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Benefits of intermediate annotations in reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5627–5634, Online. Association for Computational Linguistics.

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.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Yunlong Feng, Yang Xu, Libo Qin, Yasheng Wang, and Wanxiang Che. 2024. Improving language model reasoning with self-motivated learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 8840–8852. ELRA and ICCL.

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

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023a. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yao Fu, Hao-Chun Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023b. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Alfonso Emilio Gerevini. 2020. An introduction to the planning domain definition language (PDDL): book review. *Artif. Intell.*, 280:103221.

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.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024a. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024b. ToRA: A tool-integrated reasoning agent for mathematical problem solving. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *ArXiv preprint*, abs/2310.19736.

Pranay Gupta and Manish Gupta. 2022. Newskvqa: Knowledge-aware news video question answering. In *Advances in Knowledge Discovery and Data Mining - 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16-19, 2022, Proceedings, Part III*, volume 13282 of *Lecture Notes in Computer Science*, pages 3–15. Springer.

Chengcheng Han, Xiaowei Du, Che Zhang, Yixin Lian, Xiang Li, Ming Gao, and Baoyuan Wang. 2023. DialCoT meets PPO: Decomposing and exploring reasoning paths in smaller language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8055–8068, Singapore. Association for Computational Linguistics.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq R. Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. FOLIO: natural language reasoning with first-order logic. *ArXiv preprint*, abs/2209.00840.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023a. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. 2023b. ToolkenGPT: Augmenting frozen language models with massive tools via tool embeddings. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

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

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023b. Exploring human-like translation strategy with large language models. *ArXiv preprint*, abs/2305.04118.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. 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*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2023. A closer look at the self-verification abilities of large language models in logical reasoning. *CoRR*, abs/2311.07954.

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.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4919, Singapore. Association for Computational Linguistics.

Cheng-Yu Hsieh, Si-An Chen, Chun-Liang Li, Yasuhisa Fujii, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2023a. Tool documentation enables zero-shot tool-usage with large language models. *ArXiv preprint*, abs/2308.00675.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023b. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Chi Hu, Yuan Ge, Xiangnan Ma, Hang Cao, Qiang Li, Yonghua Yang, Tong Xiao, and Jingbo Zhu. 2024a. Rankprompt: Step-by-step comparisons make language models better reasoners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 13524–13536. ELRA and ICCL.

Hanxu Hu, Hongyuan Lu, Huajian Zhang, Wai Lam, and Yue Zhang. 2023a. Chain-of-symbol prompting elicits planning in large langauge models. *ArXiv preprint*, abs/2305.10276.

Mengkang Hu, Yao Mu, Xinmiao Chelsey Yu, Mingyu Ding, Shiguang Wu, Wenqi Shao, Qiguang Chen, Bin Wang, Yu Qiao, and Ping Luo. 2024b. Tree-planner: Efficient close-loop task planning with large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

Pengbo Hu, Ji Qi, Xingyu Li, Hong Li, Xinqi Wang, Bing Quan, Ruiyu Wang, and Yi Zhou. 2023b. Tree-of-mixed-thought: Combining fast and slow thinking for multi-hop visual reasoning. *ArXiv preprint*, abs/2308.09658.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023a. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024a. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

Jinfeng Huang, Qiaoqiao She, Wenbin Jiang, Hua Wu, Yang Hao, Tong Xu, and Feng Wu. 2024b. Qdmr-based planning-and-solving prompting for complex reasoning tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 13395–13406. ELRA and ICCL.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and Lichao Sun. 2023c. Metatool benchmark: Deciding whether to use tools and which to use. *Preprint*, arXiv:2310.03128.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023d. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *ArXiv preprint*, abs/2305.08322.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. Association for Computational Linguistics.

Raer Jack. 2023. Compression for agi. *Stanford MLSys*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection. *ArXiv preprint*, abs/2310.06271.

Dongwei Jiang, Jingyu Zhang, Orion Weller, Nathaniel Weir, Benjamin Van Durme, and Daniel Khashabi. 2024. SELF-[IN]CORRECT: llms struggle with refining self-generated responses. *CoRR*, abs/2404.04298.

Song Jiang, Zahra Shakeri, Aaron Chan, Maziar Sanjabi, Hamed Firooz, Yinglong Xia, Bugra Akyildiz, Yizhou Sun, Jinchao Li, Qifan Wang, et al. 2023a. Resprompt: Residual connection prompting advances multi-step reasoning in large language models. *ArXiv preprint*, abs/2310.04743.

Weisen Jiang, Han Shi, Longhui Yu, Zhengying Liu, Yu Zhang, Zhenguo Li, and James T. Kwok. 2023b. Forward-backward reasoning in large language models for verification. *ArXiv preprint*, abs/2308.07758.

Yan Junbing, Chengyu Wang, Taolin Zhang, Xiaofeng He, Jun Huang, and Wei Zhang. 2023. From complex to simple: Unraveling the cognitive tree for reasoning with small language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12413–12425, Singapore. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared

Kaplan. 2022. Language models (mostly) know what they know. *ArXiv preprint*, abs/2207.05221.

Ehud D. Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tenenholtz. 2022. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *ArXiv preprint*, abs/2205.00445.

Uri Katz, Mor Geva, and Jonathan Berant. 2022. Inferring implicit relations in complex questions with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2548–2566, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Discriminator-guided multi-step reasoning with language models. *ArXiv preprint*, abs/2305.14934.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12685–12708, Singapore. Association for Computational Linguistics.

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

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.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023a. Better zero-shot reasoning with role-play prompting. *CoRR*, abs/2308.07702.

Yilun Kong, Jingqing Ruan, Yihong Chen, Bin Zhang, Tianpeng Bao, Shiwei Shi, Guoqing Du, Xiaoru Hu, Hangyu Mao, Ziyue Li, Xingyu Zeng, and Rui Zhao. 2023b. Tptu-v2: Boosting task planning and tool usage of large language model-based agents in real-world systems. *Preprint*, arXiv:2311.11315.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *ArXiv preprint*, abs/2307.13702.

Soochan Lee and Gunhee Kim. 2023. Recursion of thought: A divide-and-conquer approach to multi-context reasoning with language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 623–658, Toronto, Canada. Association for Computational Linguistics.

Bin Lei, Pei-Hung Lin, Chunhua Liao, and Caiwen Ding. 2023a. Boosting logical reasoning in large language models through a new framework: The graph of thought. *ArXiv preprint*, abs/2308.08614.

Deren Lei, Yaxi Li, Mingyu Wang, Vincent Yun, Emily Ching, Eslam Kamal, et al. 2023b. Chain of natural language inference for reducing large language model ungrounded hallucinations. *ArXiv preprint*, abs/2310.03951.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8769–8784, Online. Association for Computational Linguistics.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR.

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. In *NeurIPS*.

Chenglin Li, Qianglong Chen, Caiyu Wang, and Yin Zhang. 2023a. Mixed distillation helps smaller language model better reasoning. *CoRR*, abs/2312.10730.

Jiangtong Li, Li Niu, and Liqing Zhang. 2022. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 21241–21250. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023c. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Minghao Li, Feifan Song, Bowen Yu, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023d. Api-bank: A benchmark for tool-augmented llms. *ArXiv preprint*, abs/2304.08244.

Xiang Li, Shizhu He, Jiayu Wu, Zhao Yang, Yao Xu, Yang jun Jun, Haifeng Liu, Kang Liu, and Jun Zhao. 2024. Mode-cotd: Chain-of-thought distillation for complex reasoning tasks with mixture of decoupled lora-experts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11475–11485. ELRA and ICCL.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023e. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Xiaonan Li and Xipeng Qiu. 2023. Mot: Pre-thinking and recalling enable chatgpt to self-improve with memory-of-thoughts. *ArXiv preprint*, abs/2305.05181.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq R. Joty, and Soujanya Poria. 2023f. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *ArXiv preprint*, abs/2305.13269.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023g. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Yingcong Li, Kartik Sreenivasan, Angeliki Giannou, Dimitris S. Papailiopoulos, and Samet Oymak. 2023h. Dissecting chain-of-thought: A study on compositional in-context learning of mlps. *ArXiv preprint*, abs/2305.18869.

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. 2022. Holistic evaluation of language models. *ArXiv preprint*, abs/2211.09110.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *ArXiv preprint*, abs/2305.19118.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

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.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023a. Llm+p: Empowering large language models with optimal planning proficiency. *Preprint*, arXiv:2304.11477.

Hanmeng Liu, Zhiyang Teng, Ruoxi Ning, Jian Liu, Qiji Zhou, and Yue Zhang. 2023b. Glore: Evaluating logical reasoning of large language models. *ArXiv preprint*, abs/2310.09107.

Jiacheng Liu, Ramakanth Pasunuru, Hannaneh Hajishirzi, Yejin Choi, and Asli Celikyilmaz. 2023c. Crystal: Introspective reasoners reinforced with self-feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11557–11572, Singapore. Association for Computational Linguistics.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3622–3628. ijcai.org.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023d. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023e. Plan, verify and switch: Integrated reasoning with diverse X-of-thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2807–2822, Singapore. Association for Computational Linguistics.

Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023f. Plan, verify and switch: Integrated reasoning with diverse X-of-thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2807–2822, Singapore. Association for Computational Linguistics.

Jieyi Long. 2023. Large language model guided tree-of-thought. *ArXiv preprint*, abs/2305.08291.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023a. Chain-of-dictionary prompting elicits translation in large language models. *ArXiv preprint*, abs/2305.06575.

Pan Lu, Swaroop Mishra, Tanglin 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 *NeurIPS*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023c. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631, Toronto, Canada. Association for Computational Linguistics.

Yining Lu, Haoping Yu, and Daniel Khashabi. 2024. GEAR: Augmenting language models with generalizable and efficient tool resolution. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 112–138, St. Julian's, Malta. Association for Computational Linguistics.

Man Luo, Shrinidhi Kumbhar, Ming shen, Mihir Parmar, Neeraj Varshney, Pratyay Banerjee, Somak Aditya, and Chitta Baral. 2023. Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. *Preprint*, arXiv:2310.00836.

Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023. Let's reward step by step: Step-level reward model as the navigators for reasoning. *Preprint*, arXiv:2310.10080.

Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. 2024. At which training stage does code data help LLMs reasoning? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. *ArXiv preprint*, abs/2209.07686.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

William Merrill and Ashish Sabharwal. 2023. The expresssive power of transformers with chain of thought. *Preprint*, arXiv:2310.07923.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024. Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

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.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5807–5832, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Shentong Mo and Miao Xin. 2023. Tree of uncertain thoughts reasoning for large language models. *ArXiv preprint*, abs/2309.07694.

Md Mahadi Hasan Nahid and Davood Rafiei. 2024. Tabsqlify: Enhancing reasoning capabilities of llms through table decomposition. *CoRR*, abs/2404.10150.

Ranjita Naik, Varun Chandrasekaran, Mert Yuksekgonul, Hamid Palangi, and Besmira Nushi. 2023. Diversity of thought improves reasoning abilities of large language models. *ArXiv preprint*, abs/2310.07088.

Deepak Nathani, David Wang, Liangming Pan, and William Wang. 2023. MAF: Multi-aspect feedback for improving reasoning in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6591–6616, Singapore. Association for Computational Linguistics.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Huazhong Yang, and Yu Wang. 2023. Skeleton-of-thought: Large language models can do parallel decoding. *ArXiv preprint*, abs/2307.15337.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *ArXiv preprint*, abs/2309.09117.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *ArXiv preprint*, abs/2303.08774.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. 2023. Art: Automatic multi-step reasoning and tool-use for large language models. *Preprint*, arXiv:2303.09014.

Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm: Tool augmented language models. *ArXiv preprint*, abs/2205.12255.

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.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024a. REFINER: Reasoning feedback on intermediate representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1100–1126, St. Julian's, Malta. Association for Computational Linguistics.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024b. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *CoRR*, abs/2402.13950.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv preprint*, abs/2306.14824.

Silviu Pitis, Michael R. Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *ArXiv preprint*, abs/2304.05970.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Ben Prystawski, Michael Li, and Noah D. Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. 2023. The art of SOCRATIC QUESTIONING: Recursive thinking with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4177–4199, Singapore. Association for Computational Linguistics.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

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

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *ArXiv preprint*, abs/2307.11768.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. 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.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. 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.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. 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.

Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Guoqing Du, Shiwei Shi, Hangyu Mao, Xingyu Zeng, and Rui Zhao. 2023. Tptu: Task planning and tool usage of large language model-based ai agents. *Preprint*, arXiv:2308.03427.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *ArXiv preprint*, abs/2211.05100.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Bilgehan Sel, Ahmad Al-Tawaha, Vanshaj Khattar, Lu Wang, Ruoxi Jia, and Ming Jin. 2023. Algorithm of thoughts: Enhancing exploration of ideas in large language models. *ArXiv preprint*, abs/2308.10379.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023a. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023b. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. *ArXiv preprint*, abs/2302.00618.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023a. HuggingGPT: Solving AI tasks with chatGPT and its friends in hugging face. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yueting Zhuang. 2023b. Taskbench: Benchmarking large language models for task automation. *Preprint*, arXiv:2311.18760.

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. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Kumar Shridhar, Harsh Jhamtani, Hao Fang, Benjamin Van Durme, Jason Eisner, and Patrick Xia. 2023. Screws: A modular framework for reasoning with revisions. *ArXiv preprint*, abs/2309.13075.

Kashun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12113–12139, Singapore. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, abs/2206.04615.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2023. Adaplanner: Adaptive planning from feedback with language models. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of AI through gamification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. Large language models are in-context semantic reasoners rather than symbolic reasoners. *ArXiv preprint*, abs/2305.14825.

Qingyuan Tian, Hanlun Zhu, Lei Wang, Yang Li, and Yunshi Lan. 2023. R^3 prompting: Review, rephrase and resolve for chain-of-thought reasoning in large language models under noisy context. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1670–1685, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Rasul Tutunov, Antoine Grosnit, Juliusz Ziomek, Jun Wang, and Haitham Bou-Ammar. 2023. Why can large language models generate correct chain-of-thoughts? *ArXiv preprint*, abs/2310.13571.

Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Y. Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *ArXiv preprint*, abs/2211.14275.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3493–3514, Toronto, Canada. Association for Computational Linguistics.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2023b. Apollo's oracle: Retrieval-augmented reasoning in multi-agent debates. *Preprint*, arXiv:2312.04854.

Jianing Wang, Qiushi Sun, Nuo Chen, Xiang Li, and Ming Gao. 2023c. Boosting language models reasoning with chain-of-knowledge prompting. *ArXiv preprint*, abs/2306.06427.

Jinyuan Wang, Junlong Li, and Hai Zhao. 2023d. Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2717–2731, Singapore. Association for Computational Linguistics.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023e. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *Preprint*, arXiv:2308.13259.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023f. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023g. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. *ArXiv preprint*, abs/2305.03453.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023h. A survey on large language model based autonomous agents. *ArXiv preprint*, abs/2308.11432.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023i. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023j. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2023k. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *CoRR*, abs/2312.08935.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, and Alessandro Sordoni. 2023l. Guiding language model reasoning with planning tokens. *Preprint*, arXiv:2310.05707.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023m. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023n. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Yuqing Wang and Yun Zhao. 2023. TRAM: benchmarking temporal reasoning for large language models. *ArXiv preprint*, abs/2310.00835.

Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023o. Democratizing reasoning ability: Tailored learning from large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1948–1966, Singapore. 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. *Trans. Mach. Learn. Res.*, 2022.

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 *NeurIPS*.

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.

Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. 2021. STAR: A benchmark for situated reasoning in real-world videos. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Haoyi Wu, Wenyang Hui, Yezeng Chen, Weiqi Wu, Kewei Tu, and Yi Zhou. 2023a. Conic10K: A challenging math problem understanding and reasoning dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6444–6458, Singapore. Association for Computational Linguistics.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. 2023b. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *ArXiv preprint*, abs/2307.13339.

Yexin Wu, Zhuosheng Zhang, and Hai Zhao. 2024. Mitigating misleading chain-of-thought reasoning with selective filtering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11325–11340. ELRA and ICCL.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *ArXiv preprint*, abs/2309.07864.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. 2024. Badchain: Backdoor chain-of-thought prompting for large language models. *CoRR*, abs/2401.12242.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P. Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *CoRR*.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Qizhe Xie. 2023. Self-evaluation guided beam search for reasoning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Weijia Xu, Andrzej Banburski-Fahey, and Nebojsa Jojic. 2023. Reprompting: Automated chain-of-thought prompt inference through gibbs sampling. *Preprint*, arXiv:2305.09993.

Tianci Xue, Ziqi Wang, Zhenhailong Wang, Chi Han, Pengfei Yu, and Heng Ji. 2023. RCOT: detecting and rectifying factual inconsistency in reasoning by reversing chain-of-thought. *ArXiv preprint*, abs/2305.11499.

Bohao Yang, Chen Tang, Kun Zhao, Chenghao Xiao, and Chenghua Lin. 2024a. Effective distillation of table-based reasoning ability from llms. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 5538–5550. ELRA and ICCL.

Hui Yang, Sifu Yue, and Yunzhong He. 2023a. Auto-gpt for online decision making: Benchmarks and additional opinions. *ArXiv preprint*, abs/2306.02224.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. MM-REACT: prompting chatgpt for multimodal reasoning and action. *ArXiv preprint*, abs/2303.11381.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2024b. Language models as inductive reasoners. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 209–225, St. Julian's, Malta. Association for Computational Linguistics.

Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. 2023c. Logical reasoning over natural language as knowledge representation: A survey. *ArXiv preprint*, abs/2303.12023.

Fanglong Yao, Changyuan Tian, Jintao Liu, Zequn Zhang, Qing Liu, Li Jin, Shuchao Li, Xiaoyu Li, and Xian Sun. 2023a. Thinking like an expert:multimodal hypergraph-of-thought (hot) reasoning to boost foundation modals. *Preprint*, arXiv:2308.06207.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023b. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023c. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yao Yao, Zuchao Li, and Hai Zhao. 2023d. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. *ArXiv preprint*, abs/2305.16582.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. 2023a. SatLM: Satisfiability-aided language models using declarative prompting. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

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

Xi Ye and Greg Durrett. 2023. Explanation selection using unlabeled data for in-context learning. *ArXiv preprint*, abs/2302.04813.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023b. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 174–184. ACM.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. CLEVRER: collision events for video representation and reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Tianxiang Sun, Cheng Chang, Qinyuan Cheng, Ding Wang, Xiaofeng Mou, Xipeng Qiu, and Xuanjing Huang. 2024. Aggregation of reasoning: A hierarchical framework for enhancing answer selection in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 609–625. ELRA and ICCL.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5942–5966, Singapore. Association for Computational Linguistics.

Fei Yu, Hongbo Zhang, and Benyou Wang. 2023a. Nature language reasoning, A survey. *ArXiv preprint*, abs/2303.14725.

Junchi Yu, Ran He, and Zhitao Ying. 2024. THOUGHT PROPAGATION: AN ANALOGICAL APPROACH TO COMPLEX REASONING WITH LARGE LANGUAGE MODELS. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiao Yu, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhou Yu. 2023b. Teaching language models to self-improve through interactive demonstrations. *Preprint*, arXiv:2310.13522.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023c. Towards better chain-of-thought prompting strategies: A survey. *Preprint*, arXiv:2310.04959.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*.

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.

Bowen Zhang, Kehua Chang, and Chunping Li. 2023a. Cot-bert: Enhancing unsupervised sentence representation through chain-of-thought. *ArXiv preprint*, abs/2309.11143.

Hugh Zhang and David C. Parkes. 2023. Chain-of-thought reasoning is a policy improvement operator. *Preprint*, arXiv:2309.08589.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. 2023b. Draft & verify: Lossless large language model acceleration via self-speculative decoding. *ArXiv preprint*, abs/2309.08168.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023c. How language model hallucinations can snowball. *ArXiv preprint*, abs/2305.13534.

Sarah J. Zhang, Reece Shuttleworth, Derek Austin, Yann Hicke, Leonard Tang, Sathwik Karnik, Darnell Granberry, and Iddo Drori. 2022a. A dataset and benchmark for automatically answering and generating machine learning final exams. *ArXiv preprint*, abs/2206.05442.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. OPT: open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068.

Tianhua Zhang, Jiaxin Ge, Hongyin Luo, Yung-Sung Chuang, Mingye Gao, Yuan Gong, Xixin Wu, Yoon Kim, Helen Meng, and James Glass. 2023d. Natural language embedded programs for hybrid language symbolic reasoning. *ArXiv preprint*, abs/2309.10814.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2024. Cumulative reasoning with large language models. In *ICLR 2024 Workshop on*

Bridging the Gap Between Practice and Theory in Deep Learning.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023e. Siren's song in the AI ocean: A survey on hallucination in large language models. *ArXiv preprint*, abs/2309.01219.

Zhebin Zhang, Xinyu Zhang, Yuanhang Ren, Saijiang Shi, Meng Han, Yongkang Wu, Ruofei Lai, and Zhao Cao. 2023f. IAG: Induction-augmented generation framework for answering reasoning questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1–14, Singapore. Association for Computational Linguistics.

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, and Hai Zhao. 2023g. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *CoRR*, abs/2311.11797.

Zhuosheng Zhang and Aston Zhang. 2023. You only look at screens: Multimodal chain-of-action agents. *Preprint*, arXiv:2309.11436.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023h. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023i. Multimodal chain-of-thought reasoning in language models. *ArXiv preprint*, abs/2302.00923.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023a. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Zheng Gong, Beichen Zhang, Yuanhang Zhou, Jing Sha, Zhigang Chen, Shijin Wang, Cong Liu, and Ji-Rong Wen. 2022. Jiuzhang: A chinese pre-trained language model for mathematical problem understanding. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 4571–4581. ACM.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023b. A survey of large language models. *ArXiv preprint*, abs/2303.18223.

Xufeng Zhao, Mengdi Li, Wenhao Lu, Cornelius Weber, Jae Hee Lee, Kun Chu, and Stefan Wermter. 2024a. Enhancing zero-shot chain-of-thought reasoning in large language models through logic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6144–6166, Torino, Italia. ELRA and ICCL.

Yichun Zhao, Shuheng Zhou, and Huijia Zhu. 2024b. Probe then retrieve and reason: Distilling probing and reasoning capabilities into smaller language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 13026–13032. ELRA and ICCL.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023a. Progressive-hint prompting improves reasoning in large language models. *ArXiv preprint*, abs/2304.09797.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023b. DDCot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. 2024. Take a step back: Evoking reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2023a. Language agent tree search unifies reasoning acting and planning in language models. *Preprint*, arXiv:2310.04406.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023b. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2023c. The mystery and fascination of llms: A comprehensive survey on the interpretation and analysis of emergent abilities. *ArXiv preprint*, abs/2311.00237.

Zhehua Zhou, Jiayang Song, Kunpeng Yao, Zhan Shu, and Lei Ma. 2023d. Isr-llm: Iterative self-refined large language model for long-horizon sequential task planning. *Preprint*, arXiv:2308.13724.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. 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 3277–3287, Online. Association for Computational Linguistics.

Tinghui Zhu, Kai Zhang, Jian Xie, and Yu Su. 2024a. Deductive beam search: Decoding deducible rationale for chain-of-thought reasoning. *CoRR*, abs/2401.17686.

Xuekai Zhu, Biqing Qi, Kaiyan Zhang, Xingwei Long, and Bowen Zhou. 2023. Pad: Program-aided distillation specializes large models in reasoning. *CoRR*, abs/2305.13888.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024b. Improving small language models' mathematical reasoning via equation-of-thought distillation. *CoRR*, abs/2401.11864.

Yuchen Zhuang, Xiang Chen, Tong Yu, Saayan Mitra, Victor Bursztyn, Ryan A. Rossi, Somdeb Sarkhel, and Chao Zhang. 2024. Toolchain*: Efficient action space navigation in large language models with a* search. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net.

Jin Ziqi and Wei Lu. 2023. Tab-CoT: Zero-shot tabular chain of thought. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10259–10277, Toronto, Canada. Association for Computational Linguistics.

Anni Zou, Zhuosheng Zhang, and Hai Zhao. 2024. Aurora: A one-for-all platform for augmented reasoning and refining with task-adaptive chain-of-thought prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 1801–1807. ELRA and ICCL.

Anni Zou, Zhuosheng Zhang, Hai Zhao, and Xiangru Tang. 2023. Meta-cot: Generalizable chain-of-thought prompting in mixed-task scenarios with large language models. *ArXiv preprint*, abs/2310.06692.

A Appendix

A.1 Related Survey

Zhao et al. (2023b) primarily focuses on the development of contemporary LLMs, while Qiu et al. (2020) surveys about early PLMs. Some works discuss reasoning in specific domains, such as mathematical reasoning (Lu et al., 2023c), commonsense reasoning (Talmor et al., 2019), and logical reasoning (Yang et al., 2023c). Huang et al. (2023b); Zhang et al. (2023e) conducts an investigation into potential hallucination phenomena in LLM's reasoning. Dong et al. (2023) discusses in-context learning techniques in the era of LLMs, and Yu et al. (2023a) conducts a macroscopic investigation into natural language reasoning. Liu et al. (2023d) discusses prompt tuning, Qiao et al. (2023); Yu et al. (2023c); Huang and Chang (2023) focus on prompt engineering and reasoning strategies, and Zhang et al. (2023g) highlights the development from chain-of-thought reasoning to autonomous agents. This repository[2] also collects chain-of-thought reasoning papers.

Distinct from the above-mentioned surveys, this paper focuses on generalized chain-of-thought (XoT) reasoning in the era of LLMs. This is the first systematic investigation into XoT reasoning, and we hope our work can serve as an overview to facilitate future research.

A.2 Further Discussion

Open Question: Does CoT ability originate from code data pre-training? This is a pending question, initially summarized by Fu and Khot (2022) and widely circulated in the research community. In the early stages, LLMs like GPT3 (Brown et al., 2020) (davinci) and OPT (Zhang et al., 2022b) usually do not possess CoT capabilities, and they do not use or only incorporate a small amount of code data (not specialized) during pre-training. Recent models often incorporate specialized code data during pre-training, such as GPT-3.5, LLaMA2 (Touvron et al., 2023b) (with approximately 8% of code data during pre-training) and they all possess strong CoT capabilities. Additionally, Gao et al. (2023); Chen et al. (2022a) have found that the use of programming language form rationales can significantly enhance the model's performance on complex reasoning tasks. Various indications point towards the source of CoT abilities lying in code data during pre-training.

[2] Timothyxxx/Chain-of-ThoughtsPapers

Recently, Ma et al. (2024) investigates the impact of code data on LLMs at different training stages, reaching the first qualitative conclusion supported by quantitative experimental results. They find that mixing code data during the pre-training stage enhances general reasoning abilities, while doing that in the instruction fine-tuning stage endows task-specific reasoning abilities.

Open Question: How to provide precise feedback on model's reasoning or decisions? When dealing with multi-step reasoning or decision-making tasks, errors often occur in intermediate steps, and if these errors are not corrected promptly, they may lead to cascading errors. Currently, the primary methods for obtaining feedback include feedback from model itself (Madaan et al., 2023; Shinn et al., 2023), feedback from other models (Paul et al., 2024a), feedback from the external environment (Nathani et al., 2023; Gou et al., 2024a), and feedback based on reinforcement learning (Uesato et al., 2022; Lightman et al., 2024; Ma et al., 2023). However, some studies have raised doubts about the ability of LLMs to provide self-feedback (Huang et al., 2024a; Jiang et al., 2024). Generally speaking, certain issues exist in current methods. (1) How dependable is the feedback generated by the model itself? (2) Is there a fundamental distinction between feedback from other language models and self-feedback? (3) Does the feedback quality still remain constrained by the model's capability boundaries? (4) How is external feedback for various scenarios pre-defined, and how can this be expanded to different scenarios?

In summary, there is currently no fully satisfying feedback approach and more research attention is needed on how to accurately obtain feedback signals from the model's intermediate reasoning.

Discussion: Towards (early) AGI AGI has been the long-standing ultimate aspiration in the realm of artificial intelligence. Recent research on LLM-powered autonomous agents has successfully demonstrated a preliminary implementation of nascent artificial general intelligence (AGI).

Synergy between reasoning and interaction. Equipped with robust language comprehension capabilities, LLMs can interact with the external world through text-based interactions using plugins (tools, KB query, search engine, etc.) (Schick et al., 2023; Shen et al., 2023a; Qin et al., 2024). Combining powerful reasoning capabilities, LLMs have made significant strides in various planning and decision-making tasks (Shinn et al., 2023; Yao et al., 2023b; Zhuang et al., 2024), catalyzing research on LLM-based autonomous agents (Wang et al., 2023h; Xi et al., 2023; Zhang et al., 2023g).

LLM acts as the Brain (Controller). In contrast to traditional AI, which concentrates on specific tasks, AGI seeks the ability to understand general tasks (Devlin et al., 2019; Dosovitskiy et al., 2021), covering a widespread spectrum. Within LLM-powered AI, the LLM typically serves as the brain (or central controller), handling reasoning, planning and decision-making, while delegating specific execution to dedicated modules (tools, weak AI, etc.) (Shen et al., 2023a; Yang et al., 2023a). LLM-powered AI has already diverged significantly from weak AI and is progressing toward human cognition and thinking.

While some studies suggest that LLMs represent an early manifestation of AGI (Bubeck et al., 2023; Jack, 2023), there are also scholars who contend that LLMs may not progress into AGI due to factors such as auto-regressive modeling and limited memory. As of now, there is still intense debate on whether LLMs can evolve into AGI. But regardless, LLM-powered AI has embarked on a distinctly different path from traditional AI, evolving towards a more generalized direction.

A.3 Early Attempts and Efforts in Specific Domains

In this section, we list the early attempts of XoT reasoning and efforts focused on specific domains.

Before the concept of CoT was introduced (Wei et al., 2022b), some efforts were made to enhance reasoning performance through the use of rationales (Marasovic et al., 2022; Rajani et al., 2019a,b; Dua et al., 2020). After that, certain work has empirically demonstrated the effectiveness of chain-of-thought prompting (Lampinen et al., 2022; Ye and Durrett, 2022; Arora et al., 2023) and Shi et al. (2023) explores multi-lingual CoT reasoning. Other work focuses on specific domains, such as machine translation (He et al., 2023b), sentiment analysis (Fei et al., 2023), sentence embeddings (Zhang et al., 2023a), summarization (Wang et al., 2023n), arithmetic (Lee and Kim, 2023), tabular reasoning (Chen, 2023; Ziqi and Lu, 2023), and backdoor attack (Xiang et al., 2024), etc. Katz et al. (2022); Zhang et al. (2022a) provide benchmarks and resources. Besides, some research utilizes specific pre-training to enhance reasoning (Lewkowycz et al., 2022; Zhao et al., 2022).

A.4 Empirical Results

We statistic the performance of various XoT methods in mathematics, commonsense, and symbolic reasoning, as shown in Table 2. We primarily collect the performance of GPT series models and the results are mainly from corresponding papers (some results are used as baselines in other papers). It is worth noting that due to variations in model checkpoints and experimental setups, even the methods with the same backbone LLM may not be fairly comparable. Therefore, this table only provides a rough trend of performance.

B Details of Benchmarks

B.1 Mathematical Reasoning

Mathematical reasoning is often used to measure the reasoning power of a model. Early benchmarks contain simple arithmetic operations (Hosseini et al., 2014; Koncel-Kedziorski et al., 2015; Roy and Roth, 2015; Koncel-Kedziorski et al., 2016). Ling et al. (2017) labels the reasoning process in natural language form, and Amini et al. (2019) builds on AQUA by labeling the reasoning process in program form. Later benchmarks (Miao et al., 2020; Patel et al., 2021; Cobbe et al., 2021; Gao et al., 2023) contain more complex and diverse questions. (Zhu et al., 2021; Chen et al., 2021, 2022b) require reasoning based on the table content. There are also competition-level benchmarks (Hendrycks et al., 2021b; Mishra et al., 2022a,b) and reading comprehension form benchmarks (Dua et al., 2019; Chen et al., 2023b).

B.2 Commonsense Reasoning

Commonsense reasoning entails the process of drawing inferences, forming judgments, and gaining insights based on widely known and commonly accepted world knowledge. Acquiring and understanding commonsense knowledge presents a significant challenge for models engaged in commonsense reasoning. Various benchmarks have been put forward to address these challenges, including commonsense understanding (Talmor et al., 2019, 2021; Bhakthavatsalam et al., 2021; Mihaylov et al., 2018; Geva et al., 2021; Huang et al., 2019; Bisk et al., 2020), event temporal commonsense reasoning (Rashkin et al., 2018; Zhou et al., 2019) , and commonsense verification (Wang et al., 2019).

B.3 Symbolic Reasoning

Symbolic reasoning here refers specifically to the simulation of some simple operations, which are simple for humans yet challenging for LLMs. Last letter concatenation, coin flip, and reverse list (Wei et al., 2022b) are the most commonly used symbolic reasoning tasks. In addition, the collaborative benchmark BigBench (Srivastava et al., 2022) and BigBench-Hard (Suzgun et al., 2023) also contain several symbolic reasoning datasets, such as state tracking and object counting.

B.4 Logical Reasoning

Logical reasoning encompasses deductive reasoning, inductive reasoning, and abductive reasoning. Deductive reasoning derives conclusions from general premises (Liu et al., 2020; Yu et al., 2020; Tafjord et al., 2021; Han et al., 2022; Hong et al., 2023). Inductive reasoning derives general conclusions from special cases (Yang et al., 2024b). Abductive reasoning gives rational explanations for observed phenomena (Saparov and He, 2023).

B.5 Multi-modal Reasoning

In the real world, reasoning also involves information in modalities other than text, with visual modalities being the most prevalent. To this end, many benchmarks for visual multi-modal reasoning are proposed (Zellers et al., 2019; Park et al., 2020; Dong et al., 2022; Lu et al., 2022), and among them, ScienceQA (Lu et al., 2022) annotates reasoning process and is the most commonly used visual multi-modal reasoning benchmark. Video multi-modal reasoning (Lei et al., 2020; Yi et al., 2020; Wu et al., 2021; Xiao et al., 2021; Li et al., 2022; Gupta and Gupta, 2022) is more challenging as it introduces additional temporal information compared to visual multi-modal reasoning.

B.6 Comprehensive Benchmarks

Apart from the aforementioned individual datasets, there are also some comprehensive evaluation benchmarks. Some works aim to provide a holistic evaluation of the general reasoning capabilities (Srivastava et al., 2022; Suzgun et al., 2023; Hendrycks et al., 2021a; Huang et al., 2023d; Liang et al., 2022). In addition, there are also some multi-task benchmarks that focus on specific reasoning abilities, such as logical reasoning (Luo et al., 2023; Liu et al., 2023b) and temporal reasoning (Chu et al., 2023; Wang and Zhao, 2023).

Task	Dataset	Size	Input	Output	Rationale	Description
Mathematical Reasoning	AddSub (Hosseini et al., 2014)	395	Question	Number	Equation	Simple arithmetic
	SingleEq (Koncel-Kedziorski et al., 2015)	508	Question	Number	Equation	Simple arithmetic
	MultiArith (Roy and Roth, 2015)	600	Question	Number	Equation	Simple arithmetic
	MAWPS (Koncel-Kedziorski et al., 2016)	3,320	Question	Number	Equation	Simple arithmetic
	AQUA-RAT (Ling et al., 2017)	100,000	Question	Option	Natural Language	Math reasoning with NL rationale
	ASDiv (Miao et al., 2020)	2,305	Question	Number	Equation	Multi-step math reasoning
	SVAMP (Patel et al., 2021)	1,000	Question	Number	Equation	Multi-step math reasoning
	GSM8K (Cobbe et al., 2021)	8,792	Question	Number	Natural Language	Multi-step math reasoning
	GSM-Hard (Gao et al., 2023)	936	Question	Number	Natural Language	GSM8K with larger number
	MathQA (Amini et al., 2019)	37,297	Question	Number	Operation	Annotated based on AQUA
	DROP (Dua et al., 2019)	96,567	Question+Passage	Number+Span	Equation	Reading comprehension form
	TheoremQA (Chen et al., 2023b)	800	Question+Theorem	Number	✗	Answer based on theorems
	TAT-QA (Zhu et al., 2021)	16,552	Question+Table+Text	Number+Span	Operation	Answer based on tables
	FinQA (Chen et al., 2021)	8,281	Question+Table+Text	Number	Operation	Answer based on tables
	ConvFinQA (Chen et al., 2022b)	3,892	Question+Table+Dialog	Number	Operation	Multi-turn dialogs
	MATH (Hendrycks et al., 2021b)	12,500	Question	Number	Natural Language	Challenging competition math problems
	NumGLUE (Mishra et al., 2022b)	101,835	Question+Text	Number+Span	✗	Multi-task benchmark
	LILA (Mishra et al., 2022a)	133,815	Question+Text	Free-form	Program	Multi-task benchmark
Commonsense Reasoning	ARC (Bhakthavatsalam et al., 2021)	7,787	Question	Option	✗	From science exam
	OpenBookQA (Mihaylov et al., 2018)	5,957	Question+Context	Option	✗	Open-book knowledges
	PIQA (Bisk et al., 2020)	21,000	Goal+Solution	Option	✗	Physical commonsense knowledge
	CommonsenseQA (Talmor et al., 2019)	12,247	Question	Option	✗	Derived from ConceptNet
	CommonsenseQA 2.0 (Talmor et al., 2021)	14,343	Question	Yes/No	✗	Gaming annotation with high quality
	Event2Mind (Rashkin et al., 2018)	25,000	Event	Intent+Reaction	✗	Intension commonsense reasoning
	McTaco (Zhou et al., 2019)	13,225	Question	Option	✗	Event temporal commonsense reasoning
	CosmosQA (Huang et al., 2019)	35,588	Question+Paragraph	Option	✗	Narrative commonsense reasoning
	ComValidation (Wang et al., 2019)	11,997	Statement	Option	✗	Commonsense verification
	ComExplanation (Wang et al., 2019)	11,997	Statement	Option/Free-form	✗	Commonsense explanation
	StrategyQA (Geva et al., 2021)	2,780	Question	Yes/No	✗	Multi-hop commonsense reasoning
Symbolic Reasoning	Last Letter Concat. (Wei et al., 2022b)	-	Words	Letters	✗	Rule-based
	Coin Flip (Wei et al., 2022b)	-	Statement	Yes/No	✗	Rule-based
	Reverse List (Wei et al., 2022b)	-	List	Reversed List	✗	Rule-based
	BigBench (Srivastava et al., 2022)	-	-	-	✗	Contains multiple symbolic reasoning datasets
	BigBench-Hard (Suzgun et al., 2023)	-	-	-	✗	Contains multiple symbolic reasoning datasets
Logical Reasoning	ReClor (Yu et al., 2020)	6,138	Question+Context	Option	✗	Questions from GMAT and LSAT
	LogiQA (Liu et al., 2020)	8,678	Question+Paragraph	Option	✗	Questions from China Civil Service Exam
	ProofWriter (Tafjord et al., 2021)	20,192	Question+Rule	Answer+Proof	Entailment Tree	Reasoning process generation
	FOLIO (Han et al., 2022)	1,435	Conclusion+Premise	Yes/No	✗	First-order logic
	DEER (Yang et al., 2024b)	1,200	Fact	Rule	✗	Inductive reasoning
	PrOntoQA (Saparov and He, 2023)	-	Question+Context	Yes/No+Proccess	First-Order Logic	Deductive reasoning
Multimodal Reasoning	VCR (Zellers et al., 2019)	264,720	Question+Image	Option	Natural Language	Visual commonsense reasoning
	VisualCOMET (Park et al., 2020)	1,465,704	Image+Event	Action+Intent	✗	Visual commonsense reasoning
	PMR (Dong et al., 2022)	15,360	Image+Background	Option	✗	Premise-based multi-modal reasoning
	ScienceQA (Lu et al., 2022)	21,208	Q+Image+Context	Option	Natural Language	Multi-modal reasoning with NL rationales
	VLEP (Lei et al., 2020)	28,726	Premise+Video	Option	✗	Video event prediction
	CLEVRER (Yi et al., 2020)	305,280	Question+Video	Option/Free-form	Program	Video temporal and causal reasoning
	STAR (Wu et al., 2021)	600,000	Question+Video	Option	✗	Video situated reasoning
	NEXT-QA (Xiao et al., 2021)	47,692	Question+Video	Option	✗	Video temporal,causal,commonsense reasoning
	Causal-VidQA (Li et al., 2022)	107,600	Question+Video	Free-form	Natural Language	Video causal and commonsense reasoning
	News-KVQA (Gupta and Gupta, 2022)	1,041,352	Q+V+KG	Option	✗	Video reasoning with external knowledge

Table 1: An overview of benchmarks and tasks on reasoning.

Method	Setting	Backbone	Mathematical				Commonsense		Symbolic	
			GSM8K	SVAMP	Asdiv	AQuA	CSQA	StrategyQA	LastLetterConcat	CoinFlip
I-O Prompting (Brown et al., 2020)	fewshot	text-davinci-002	19.7	69.9	74	29.5	79.5	65.9	5.8	49.0
Fewshot CoT (Wei et al., 2022b)	fewshot	text-davinci-002	63.1	76.4	80.4	45.3	73.5	65.4	77.5	99.6
PoT (Chen et al., 2022a)	fewshot	text-davinci-002	80	89.1	-	58.6	-	-	-	-
Complex CoT (Fu et al., 2023a)	fewshot	text-davinci-002	72.6	-	-	-	-	77	-	-
Automate CoT (Shum et al., 2023)	fewshot	text-davinci-002	49.7	73.3	74.2	37.9	76.1	67.9	58.9	-
Fewshot CoT (Wei et al., 2022b)	fewshot	text-davinci-003	66.83	69.06	-	29.13	-	-	-	-
PHP (Zheng et al., 2023a)	fewshot	text-davinci-003	79	84.7	-	58.6	-	-	-	-
Self-consistency (Wang et al., 2023m)	fewshot	text-davinci-003	67.93	83.11	-	55.12	-	-	-	-
Active Prompt (Diao et al., 2023)	fewshot	text-davinci-003	65.6	80.5	79.8	48	78.9	74.2	71.2	-
Synthetic Prompt (Shao et al., 2023b)	fewshot	text-davinci-003	73.9	81.8	80.7	-	-	-	-	-
FOBAR (Jiang et al., 2023b)	fewshot	text-davinci-003	79.5	86	-	58.66	-	-	-	-
Boosted Prompting (Pitis et al., 2023)	fewshot	text-davinci-003	71.6	-	-	55.1	-	-	-	-
Fewshot CoT (Wei et al., 2022b)	fewshot	code-davinci-002	60.1	75.8	80.1	39.8	79	73.4	70.4	99
Self-Consistency (Wang et al., 2023m)	fewshot	code-davinci-002	78	86.8	87.8	52	81.5	79.8	73.4	99.5
PAL (Gao et al., 2023)	fewshot	code-davinci-002	72	79.4	79.6	-	-	-	-	-
Resprompt (Jiang et al., 2023a)	fewshot	code-davinci-002	66.6	-	-	45.3	-	-	-	-
DIVERSE (Li et al., 2023g)	fewshot	code-davinci-002	82.3	87	88.7	-	79.9	78.6	-	-
Least-to-Most (Zhou et al., 2023b)	fewshot	code-davinci-002	68.01	-	-	-	-	-	94	-
Boosted Prompting (Pitis et al., 2023)	fewshot	code-davinci-002	83.3	88.6	-	61.7	-	-	-	-
Fewshot CoT (Wei et al., 2022b)	fewshot	gpt-3.5-turbo	76.5	81.9	-	54.3	78	63.7	73.2	99
Self-consistency (Wang et al., 2023m)	fewshot	gpt-3.5-turbo	81.9	86.4	-	62.6	-	-	-	-
MetaCoT (Zou et al., 2023)	fewshot	gpt-3.5-turbo	75.1	88.6	-	54.7	72.4	64.5	77.2	100
Verify CoT (Ling et al., 2023)	fewshot	gpt-3.5-turbo	86	-	-	69.5	-	-	92.6	-
Active Prompting (Diao et al., 2023)	fewshot	gpt-3.5-turbo	81.8	82.5	87.9	55.3	-	-	-	-
RCoT (Xue et al., 2023)	fewshot	gpt-3.5-turbo	84.6	84.9	89.3	57.1	-	-	-	-
FOBAR (Jiang et al., 2023b)	fewshot	gpt-3.5-turbo	87.4	87.4	-	57.5	-	-	-	-
Memory-of-Thought (Li and Qiu, 2023)	fewshot	gpt-3.5-turbo	-	-	-	54.1	-	-	-	-
Adaptive-consistency (Aggarwal et al., 2023)	fewshot	gpt-3.5-turbo	82.7	85	83	-	-	67.9	-	-
Boosted Prompting (Pitis et al., 2023)	fewshot	gpt-3.5-turbo	87.1	-	-	72.8	-	-	-	-
Zeroshot CoT (Kojima et al., 2022)	zeroshot	text-davinci-002	40.5	63.7	-	31.9	64	52.3	57.6	87.8
PoT (Chen et al., 2022a)	zeroshot	text-davinci-002	57	70.8	-	43.9	-	-	-	-
AutoCoT (Zhang et al., 2023h)	zeroshot	text-davinci-002	47.9	69.5	-	36.5	74.4	65.4	59.7	99.9
COSP (Aggarwal et al., 2023)	zeroshot	code-davinci-001	8.7	-	-		55.4	52.8	-	-
Plan-and-Solve (Wang et al., 2023i)	zeroshot	text-davinci-003	58.2	72	-	42.5	65.2	63.8	64.8	96.8
Agent-Instruct (Crispino et al., 2023)	zeroshot	gpt-3.5-turbo	73.4	80.8	-	57.9	74.1	69	99.8	95.2
Self-Refine (Madaan et al., 2023)	zeroshot	gpt-3.5-turbo	64.1	-	-	-	-	-	-	-
RCoT (Xue et al., 2023)	zeroshot	gpt-3.5-turbo	82	79.6	86	55.5	-	-	-	-

Table 2: The performance of various XoT methods in commonly used mathematical, commonsense and symbolic reasoning benchmarks. It is worth noting that, due to variations in the experimental setups of different methods, their performances are not directly comparable. The table is used to provide an overall empirical insight.

- A survey of X-of-Thought
  - Advanced Methods (§4)
    - XoT Prompt Construction (§4.1)
      - Manual Prompting
        - Few-shot CoT (Wei et al., 2022b), PAL (Gao et al., 2023), PoT (Chen et al., 2022a), MathPrompter (Imani et al., 2023), Complex CoT (Fu et al., 2023a)
      - Automatic Prompting
        - Zero-shot CoT (Kojima et al., 2022), PoT (Chen et al., 2022a), Plan-and-Solve (Wang et al., 2023i), Auto-CoT (Zhang et al., 2023h), RePrompting (Xu et al., 2023), Agent-Instruct (Crispino et al., 2023), MetaCoT (Zou et al., 2023), COSP (Wan et al., 2023), LogiCoT (Zhao et al., 2024a), Role-Play Prompting (Kong et al., 2023a)
      - Semi-automatic Prompting
        - Synthetic Prompting (Shao et al., 2023b), AutoMate CoT (Shum et al., 2023), Explanation-Selection (Ye and Durrett, 2023), BoostedPrompt (Pitis et al., 2023), DynamicPrompt (Lu et al., 2023b), SPCoT (Wang et al., 2023d)
    - XoT Topological Variants (§4.2)
      - Chain Structure
        - PoT (Chen et al., 2022a), PAL (Gao et al., 2023), LINC (Olausson et al., 2023), LogicLM (Pan et al., 2023), SatisfiedLM (Ye et al., 2023a), AoT (Sel et al., 2023), CoS (Hu et al., 2023a),
      - Tree Structure
        - ToT^1 (Yao et al., 2023b), ToT^2 (Long, 2023), ToUT (Mo and Xin, 2023), Skeleton-of-Thought (Ning et al., 2023), ProbTree (Cao et al., 2023), Thought-Propagation (Yu et al., 2024)
      - Graph Structure
        - GoT^1 (Besta et al., 2023), GoT^2 (Lei et al., 2023a), ResPrompt (Jiang et al., 2023a), Cascades Mixture-of-Thought (Dohan et al., 2022)
    - XoT Enhancement Methods (§4.3)
      - Verify and Refine
        - Self-Refine (Madaan et al., 2023), DIVERSE (Li et al., 2023g), Reflexion (Shinn et al., 2023), R^{3}Prompt (Tian et al., 2023), REFINER (Paul et al., 2024a), SCREWS (Shridhar et al., 2023), CRITIC (Gou et al., 2024a), MAF (Nathani et al., 2023), CannotSelfCorrect (Huang et al., 2024a), Verify-and-Edit (Zhao et al., 2023a), VerifyCoT (Ling et al., 2023), RCoT (Xue et al., 2023), Self-Verification (Weng et al., 2022), FOBAR (Jiang et al., 2023b), AuRoRA (Zou et al., 2024), SelF-Reasoner (Wu et al., 2024), Guided Decoding (Xie et al., 2023), DBS (Zhu et al., 2024a)
      - Question Decompose
        - Least-to-Most (Zhou et al., 2023b), Decomposed Prompting (Khot et al., 2023), Successive Prompting (Dua et al., 2022), PHP (Zheng et al., 2023a), CogTree (Junbing et al., 2023), iCAP (Wang et al., 2022), Self-Ask (Press et al., 2023), IRCot (Trivedi et al., 2023), SocraticQuestion (Qi et al., 2023), CumulativeReasoning (Zhang et al., 2024), Binder (Cheng et al., 2023), VersatileDecomposer (Ye et al., 2023b)
      - Knowledge Enhancement
        - CoD (Lu et al., 2023a), CoK^1 (Li et al., 2023f), CoK^2 (Wang et al., 2023c), Memory-of-Thought (Li and Qiu, 2023), KD-CoT (Wang et al., 2023e), IAG (Zhang et al., 2023f), Self-Ask (Press et al., 2023), Iter-RetGen (Shao et al., 2023a), MCR (Yoran et al., 2023), Crystal (Liu et al., 2023c), Chain-of-Verification (Dhuliawala et al., 2023), StepbackPrompt (Zheng et al., 2024)
      - Self-Ensemble
        - Verifiers (Cobbe et al., 2021), Self-Consistency (Wang et al., 2023m), ComplexCoT (Fu et al., 2023a), GRACE (Khalifa et al., 2023), Self-Check (Miao et al., 2024), MCR (Yoran et al., 2023), Diversity-of-Thought (Naik et al., 2023), DiverseXoT (Liu et al., 2023e), CLP (Shi et al., 2023), MAD^1 (Liang et al., 2023), MAD^2 (Du et al., 2023), MAD^3 (Wang et al., 2023b), Aggregation-of-Reasoning (Yin et al., 2024), Rank Prompt (Hu et al., 2024a),
      - Efficient Reasoning
        - Adaptive Consistency (Aggarwal et al., 2023), Skeleton-of-Thought (Ning et al., 2023), ActivePrompting (Diao et al., 2023), DraphVerify (Zhang et al., 2023b), SpeculativeDecoding (Leviathan et al., 2023)
  - Frontier (§5)
    - Tool Use (§5.1)
      - MRKL (Karpas et al., 2022), TAML (Parisi et al., 2022), HuggingGPT (Shen et al., 2023a), Toolformer (Schick et al., 2023), ToolkenGPT (Hao et al., 2023b), ChatCoT (Chen et al., 2023d), LATM (Cai et al., 2024), GEAR (Lu et al., 2024), ToolLLM (Qin et al., 2024), ToolDoc (Hsieh et al., 2023a), MINT (Wang et al., 2024), ReACT (Yao et al., 2023c), ART (Paranjape et al., 2023), MMREACT (Yang et al., 2023b), API-Bank (Li et al., 2023d), MetaTool (Huang et al., 2023c), TaskBench (Shen et al., 2023b)
    - Planning (§5.2)
      - AdaPlanner (Sun et al., 2023), LLM+P (Liu et al., 2023a), LLM+DP (Dagan et al., 2023), ISRLLM (Zhou et al., 2023d), ReAct (Yao et al., 2023c), Self-Refine (Madaan et al., 2023), Reflexion (Shinn et al., 2023), Plan, Verify and Switch (Liu et al., 2023f), ToT^1 (Yao et al., 2023b), ToT^2 (Long, 2023), Tree-Planner (Hu et al., 2024b), RAP (Hao et al., 2023a), LATS (Zhou et al., 2023a), ToolChain* (Zhuang et al., 2024), TPTU (Ruan et al., 2023), TPTUv2 (Kong et al., 2023b), AgentInstruct (Crispino et al., 2023), ToRA (Gou et al., 2024b), AutoUI (Zhang and Zhang, 2023)
    - Distillation (§5.3)
      - LMSI (Huang et al., 2023a), STaR (Zelikman et al., 2022), Magister et al. (2023), SCoTD (Li et al., 2023c), SECToR (Zhang and Parkes, 2023), Distilling Step-by-Step (Hsieh et al., 2023b), SCOTT (Wang et al., 2023j), DialCoT (Han et al., 2023), Planning-Token (Wang et al., 2023l), TailoredLearning (Wang et al., 2023o), Yu et al. (2023b), ImplicitCoT (Deng et al., 2023), Fu et al. (2023b), CoT Collection (Kim et al., 2023), MoDE-CoTD (Li et al., 2024), PRR (Zhao et al., 2024b), PaD (Zhu et al., 2023), EoTD (Li et al., 2023a), Mixed Distillation (Zhu et al., 2024b), Math-Shepherd (Wang et al., 2023k), MCTS-IPL (Xie et al., 2024)
  - Future Directions (§6)
    - Multi-modal (§6.1)
      - Multi-modalCoT (Zhang et al., 2023i), GoT^3 (Yao et al., 2023d), ToMT (Hu et al., 2023b), Hypergraph-of-Thought (Yao et al., 2023a), T-SciQ (Wang et al., 2023g), SocraticQuestion (Qi et al., 2023), MMReact (Yang et al., 2023b)
    - Faithfulness (§6.2)
      - Rethinking and Retrievaling (He et al., 2023a), Verify-and-Edit (Zhao et al., 2023a), CoK (Li et al., 2023f), Verify-Edit (Zhao et al., 2023a), Chain-of-NLI (Lei et al., 2023b), Radhakrishnan et al. (2023), Lanham et al. (2023), Zhang et al. (2023c)
    - CoT Theory (§6.3)
      - Wang et al. (2023a), Madaan and Yazdanbakhsh (2022), Tang et al. (2023), Merrill and Sabharwal (2023), Wu et al. (2023b), Li et al. (2023h), Feng et al. (2023), Tutunov et al. (2023), Hou et al. (2023), Wang et al. (2023f), Schaeffer et al. (2023), Zhou et al. (2023c)
  - Benchmarks (§3)
    - Mathematical (§B.1)
      - AddSub (Hosseini et al., 2014), SingleEq (Koncel-Kedziorski et al., 2015), MultiArith (Roy and Roth, 2015), MAWPS (Koncel-Kedziorski et al., 2016), AQUA-RAT (Ling et al., 2017), ASDiv (Miao et al., 2020), SVAMP (Patel et al., 2021), GSM8K (Cobbe et al., 2021), GSM-Hard (Gao et al., 2023), MathQA (Amini et al., 2019), DROP (Dua et al., 2019), TheoremQA (Chen et al., 2023b), TAT-QA (Zhu et al., 2021), FinQA (Chen et al., 2021), ConvFinQA (Chen et al., 2022b), MATH (Hendrycks et al., 2021b), NumGLUE (Mishra et al., 2022b), LILA (Mishra et al., 2022a), Conic10K (Wu et al., 2023a)
    - Commonsense (§B.2)
      - CSQA (Talmor et al., 2019), CSQA 2.0 (Talmor et al., 2021), ARC (Bhakthavatsalam et al., 2021), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), Event2Mind (Rashkin et al., 2018), McTaco (Zhou et al., 2019), CosmosQA (Huang et al., 2019), ComValidation (Wang et al., 2019), ComExplanation (Wang et al., 2019), StrategyQA (Geva et al., 2021)
    - Symbolic (§B.3)
      - Last Letter Concat. (Wei et al., 2022b), Coin Flip (Wei et al., 2022b), Reverse List (Wei et al., 2022b), BigBench (Srivastava et al., 2022), BigBench-Hard (Suzgun et al., 2023)
    - Logical (§B.4)
      - ReClor (Yu et al., 2020), LogiQA (Liu et al., 2020), ProofWriter (Tafjord et al., 2021), FOLIO (Han et al., 2022), PrOntoQA (Saparov and He, 2023), LogiGLUE (Luo et al., 2023), GLORE (Liu et al., 2023b), FALLACIES (Hong et al., 2023)
    - Multi-modal (§B.5)
      - Image-Text
        - ScienceQA (Lu et al., 2022), VCR (Zellers et al., 2019), VisualCOMET (Park et al., 2020), PMR (Dong et al., 2022), CURE (Chen et al., 2023c)
      - Video-Text
        - VLEP (Lei et al., 2020), CLEVRER (Yi et al., 2020), STAR (Wu et al., 2021), NExT-QA (Xiao et al., 2021), Causal-VidQA (Li et al., 2022), News-KVQA (Gupta and Gupta, 2022)

Figure 8: Taxonomy of Advanced Methods, Frontiers, Future Directions, and Benchmarks (Full Edition).