Reasoning with Foundation Models: Concepts, Methodologies, and Outlook

Jiankai Sun[1], Chuanyang Zheng[1], Enze Xie[§2], Zhengying Liu[§2], Ruihang Chu[1], Jianing Qiu[1], Jiaqi Xu[1], Mingyu Ding[3], Hongyang Li[4], Mengzhe Geng[1], Yue Wu[2], Wenhai Wang[1], Junsong Chen[2,6], Zhangyue Yin[11], Xiaozhe Ren[2], Jie Fu[5], Junxian He[5], Wu Yuan[1], Qi Liu[3], Xihui Liu[3], Yu Li[1], Hao Dong[7], Yu Cheng[1], Ming Zhang[7], Pheng Ann Heng[1], Jifeng Dai[8,4], Ping Luo[3,4], Jingdong Wang[9], Ji-Rong Wen[10], Xipeng Qiu[11], Yike Guo[5], Hui Xiong[12], Qun Liu[2], Zhenguo Li[2]

[1]The Chinese University of Hong Kong.
[2]Huawei Noah's Ark Lab.
[3]The University of Hong Kong.
[4]Shanghai AI Lab.
[5]Hong Kong University of Science and Technology.
[6]Dalian University of Technology.
[7]Peking University.
[8]Tsinghua University.
[9]Hefei University of Technology.
[10]Renmin University of China.
[11]Fudan University.
[12]Hong Kong University of Science and Technology (Guangzhou).

Abstract

Reasoning, a crucial ability for complex problem-solving, plays a pivotal role in various real-world settings such as negotiation, medical diagnosis, and criminal investigation. It serves as a fundamental methodology in the field of Artificial General Intelligence (AGI). With the ongoing development of foundation models, there is a growing interest in exploring their abilities in reasoning tasks. In this paper, we introduce seminal foundation models proposed or adaptable for

[§]Project Lead. Email: {xie.enze,liuzhengying2}@huawei.com

reasoning, highlighting the latest advancements in various reasoning tasks, methods, and benchmarks. We then delve into the potential future directions behind the emergence of reasoning abilities within foundation models. We also discuss the relevance of multimodal learning, autonomous agents, and super alignment in the context of reasoning. By discussing these future research directions, we hope to inspire researchers in their exploration of this field, stimulate further advancements in reasoning with foundation models, and contribute to the development of AGI. [*†]

Keywords: Reasoning, Foundation Models, Multimodal, AI Agent, Artificial General Intelligence, Formal Methods

[*]We maintain a continuously updated reading list to benefit future research, featuring relevant papers and popular benchmarks on reasoning. GitHub: https://github.com/reasoning-survey/Awesome-Reasoning-Foundation-Models

[†]Preliminary release. We are committed to maintaining the quality and recency of this work.

Contents

1 Introduction

> *"Humans have always done nonmonotonic reasoning, but rigorous monotonic reasoning in reaching given conclusions has been deservedly more respected and admired."*
>
> John McCarthy (2004)

Reasoning is an essential aspect of artificial intelligence, with applications spanning various fields, such as problem-solving, theorem proving, decision-making, and robotics (Manning, 2022). *Thinking, Fast and Slow* (Daniel, 2017) elucidates a dual-system framework for the human mind, consisting of "System 1" and "System 2" modes of thought. "System 1" operates rapidly, relying on instincts, emotions, intuition, and unconscious processes. In contrast, "System 2" operates slower, involving conscious deliberation such as algorithmic reasoning, logical analysis, and mathematical abilities. Reasoning plays a crucial role as one of the key functions of "System 2" (Bengio, 2017; Weston and Sukhbaatar, 2023). Reasoning can be categorized into two broad types: formal language reasoning and natural language reasoning (Reiter, 1975; Berzonsky, 1978; Teig and Scherer, 2016; Yu et al. , 2023b; Zhao et al. , 2023b; Li et al. , 2023r). On one hand, as shown in Figure 1, formal language reasoning is often used in areas like formal verification of software and hardware systems, theorem proving and automated reasoning (Reiter, 1975; Berzonsky, 1978). On the other hand, natural language reasoning enables more intuitive human-computer interactions and supports tasks like question answering (Shao et al. , 2023; Jiang et al. , 2021c), information retrieval (Zhu et al. , 2023d; Ai et al. , 2023), text summarization (Liu et al. , 2023l), and sentiment analysis (Yu et al. , 2023b; Araci, 2019; Barbieri et al. , 2021).

Since their inception, foundation models (Bommasani et al. , 2021) have demonstrated remarkable efficacy across various domains, including natural language processing (Qiao et al. , 2022), computer vision (Wang et al. , 2023d), and multimodal tasks (Li, 2023). However, the burgeoning interest in general-purpose artificial intelligence has sparked a compelling debate regarding whether foundation models can exhibit human-like reasoning abilities. Consequently, there has been a surge of interest in studying the reasoning capabilities of foundation models. While previous surveys have explored the application potential of foundation models from different perspectives (Gu et al. , 2023a; Wang et al. , 2023d; Yin et al. , 2023b; Zong et al. , 2023; Lou et al. , 2023; Charalambous et al. , 2023; Wang et al. , 2023p,s,f), there remains a need for a systematic and comprehensive survey that specifically focuses on recent advancements in multimodal and interactive reasoning, which emulates human reasoning styles more closely. Figure 2 presents an overview of reasoning with regard to tasks and techniques that this article will discuss.

Foundation models typically consist of billions of parameters and undergo (pre-)training using self-supervised learning (Jain et al. , 2023) on a broad dataset (Bommasani et al. , 2021). Once (pre-)trained, foundation models can be adapted to solve numerous downstream tasks through task-specific fine-tuning, linear probing, or prompt engineering, demonstrating remarkable generalizability and impressive accuracy (Bommasani et al. , 2021; Qiu et al. , 2023a). In contrast to the soft attention

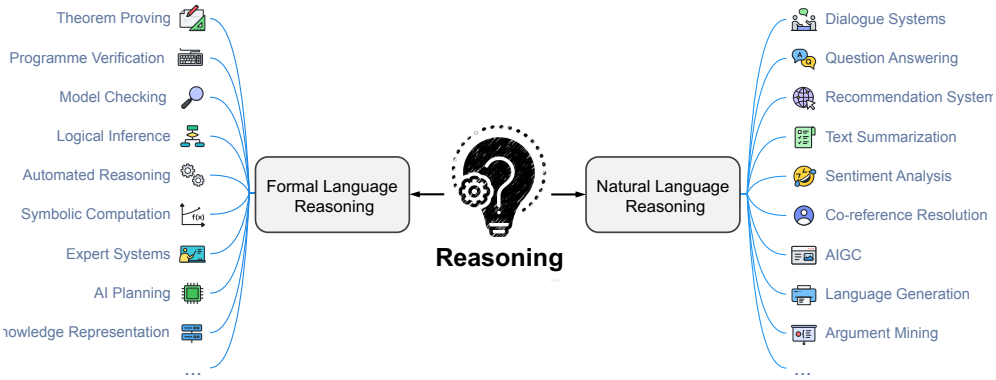


Fig. 1: Two broad types of language reasoning and examples of the supported tasks.

mechanisms utilized in conventional transformers, System 2 Attention (S2A) harnesses the capabilities of Large Language Models (LLMs) to facilitate linguistic reasoning. This method improves the factuality and objectivity of long-form content generation. By integrating logical rules and principles into the learning process (Mao et al. , 2023b), these models can perform complex tasks such as deduction and inference. This allows them to make decisions based on explicit knowledge (Mao et al. , 2023b) and logical reasoning, rather than relying solely on statistical patterns (Yang et al. , 2023e). As a rapidly growing field in artificial intelligence research, reasoning with foundation models aims to develop models capable of understanding and interacting with complex information in a more human-like manner. Built upon a foundation of logical reasoning and knowledge representation, these models make it possible to reason about abstract concepts and make decisions based on logical rules.

First, reasoning with foundation models enables the application of prior knowledge and domain expertise. Logical rules can be derived from expert knowledge or formalized from existing ontologies or knowledge graphs. By leveraging this prior knowledge, models can benefit from a better understanding of the problem domain and make more informed decisions. Second, reasoning with foundation models can enhance the robustness and generalization capabilities. By incorporating the information contained in massive amounts of data, models can better handle situations facing limited data or encountering unseen scenarios during deployment. This enables models to be more reliable and sturdy for robust, real-world usage.

In contrast to current surveys that have primarily focused on specific aspects of foundation models, such as prompts (Qiao et al. , 2022), hallucination (Rawte et al. , 2023), deductive reasoning (Huang and Chang, 2022), logical reasoning (Friedman, 2023a; Yang et al. , 2023e), causal reasoning (Kıcıman et al. , 2023; Stolfo et al. , 2022), health informatics (Qiu et al. , 2023a), or AI agents (Xi et al. , 2023), this paper takes a broader perspective, aiming to connect various research efforts in this area in a cohesive and organized manner. As Figure 2 shows, we provide a concise overview of various reasoning tasks, including **Commonsense Reasoning, Mathematical Reasoning, Logical Reasoning, Causal Reasoning, Visual Reasoning, Audio**

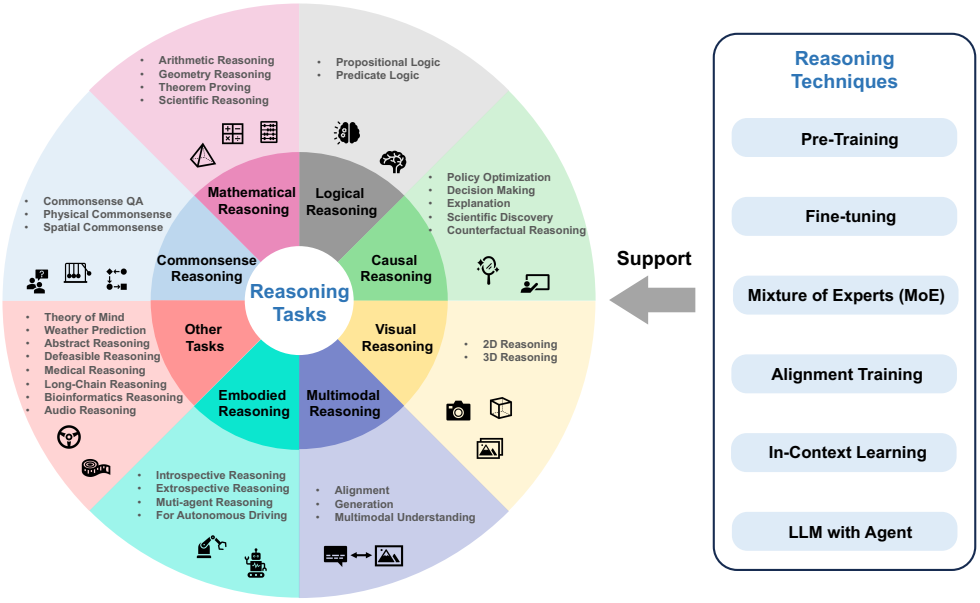


Fig. 2: Left: Overview of the reasoning tasks introduced in this survey, as detailed in Section 3. Right: Overview of the reasoning techniques for foundation models, as detailed in Section 4.

Reasoning, Multimodal Reasoning, Embodied Reasoning, Defeasible Reasoning, and beyond. By doing so, we provide a comprehensive overview highlighting the interconnections and relationships between different aspects of the field to inspire more research efforts to actively engage with and further the advances of reasoning with foundation models.

In summary, we have conducted a survey of over 650 papers on foundation models, primarily focusing on research from the past two years. We discuss different tasks, approaches, techniques, and benchmarks used in these models. We also explore various application domains that can benefit from reasoning with foundation models, such as question-answering, automated reasoning, and knowledge representation. We also discuss the challenges and limitations of current reasoning with foundation models and potential directions for future research. By understanding the advancements and challenges in this field, researchers can explore new avenues for developing intelligent systems that can reason and make decisions in a more human-like and interpretable manner. Overall, this paper aims to provide a comprehensive understanding of reasoning with foundation models, its current state, and future possibilities.

2 Background

This section introduces background knowledge about foundation models for reasoning. We will delve into key aspects such as what reasoning is, recent progress in

Context	Lee found the Northeast to be way too cold. Lee decided to move to Florida.
Question	How would you describe Lee?
Answers	a) happy b) likes cold weather **c) likes the heat**

Table 1: An Example of Commonsense Reasoning Problem from Social IQA (Sap et al. , 2019). The correct answer is in bold.

Problem	A farmer has 3 types of fruits in his garden: apples, oranges, and pears. He has twice as many apples as oranges and three times as many pears as apples. If he has 24 oranges, how many pieces of fruit does he have in total?
Expression	$x = 24 \times 2 + 24 \times 3 \times 2 + 24$
Solution	216

Table 2: A Sample Math Word Problem (MWP).

general foundation models, the architectural design of foundation models, the training methodologies employed, and the transfer learning paradigm that enables their applications for reasoning tasks. By elucidating these fundamental aspects, we hope our readers will understand the underlying principles and techniques driving reasoning with foundation models, setting the stage for the subsequent exploration of recent advancements and methodologies in this field.

2.1 Definition of Reasoning

When the term "reasoning" is brought up, its precise meaning is often unclear to people. To clarify, let us first establish a clear definition of reasoning. "Reasoning" is a broad and multifaceted concept that manifests in various contexts. It encompasses cognitive processes and logical thinking employed to analyze information, make deductions, draw conclusions, and formulate coherent arguments. Reasoning can be observed in diverse domains, such as scientific inquiry, problem-solving, decision-making, and everyday discourse. Its fundamental purpose is to enable individuals to connect pieces of information, evaluate relationships, and arrive at informed judgments or solutions. By exploring the different facets and dimensions of reasoning, we can gain a comprehensive understanding of its significance and explore the mathematical formalisms and techniques employed to elucidate and enhance this fundamental aspect of human cognition.

In addition to its broad conceptual nature, the term "reasoning" carries specific definitions within various fields. Let us briefly touch upon the definitions of reasoning

Example	
Fact1	This animal is a robin.
Rule	All robins are birds.
Fact2	This animal is a bird.
Reasoning Type	**Representation**
Deduction	(Fact1 + Rule → Fact2)
Abduction	(Fact1 + Rule ← Fact2)
Induction	(Fact1 + Fact2 → Rule)

Table 3: Illustration of deductive reasoning, abductive reasoning, and inductive reasoning. In this example, the black text represents the given knowledge, while the red text represents the inferred knowledge. The term "Fact" indicates specific information, while "Rule" denotes a general principle or guideline.

in the domains of philosophy, logic, and Natural Language Processing (NLP) (Clark et al. , 2020; Huang and Chang, 2022; Yang et al. , 2022c; Young et al. , 2022; Yu et al. , 2023a).

Philosophy

Definition 1. *(Cognitive reasoning). Cognitive reasoning refers to modeling the human ability to draw meaningful conclusions despite incomplete and inconsistent knowledge involving among others the representation of knowledge where all processes from the acquisition and update of knowledge to the derivation of conclusions must be implementable and executable on appropriate hardware (Furbach et al. , 2019).*

Logic

Definition 2. *(Logical reasoning). Logical reasoning involves a process of thought where conclusions are methodically drawn based on premises and the relationships between these premises, ensuring that the conclusions are logically implied or necessitated by them (Nunes, 2012).*

NLP

Definition 3. *(Natural language reasoning). Natural language reasoning is a process of integrating multiple knowledge (e.g., encyclopedic knowledge and commonsense knowledge) to derive some new conclusions about the (realistic or hypothetical) world. Knowledge can be derived from sources that are both explicit and implicit. Conclusions are assertions or events assumed to be true in the world, or practical actions (Yu et al. , 2023a).*

We can also get a better understanding of what reasoning is, by categorizing them from different perspectives, as shown in the next sections.

2.1.1 Deductive, Abductive, and Inductive Reasoning

Before delving into recent developments, let us first review the traditional perspectives on reasoning, which categorizes it into three primary types: inductive reasoning, deductive reasoning, and abductive reasoning. This categorization has long been recognized and provides a framework for understanding the different modes of reasoning. By examining each type, we can better understand their distinctive characteristics and applications. So, let us take a closer look at these traditional categories to enhance our comprehension of reasoning processes.

Table 3 provides an example to explain these three reasoning types, respectively. Deductive reasoning is a logical process that derives specific conclusions from general principles or premises. It follows a top-down approach, starting with general principles and applying logical rules to reach specific conclusions. Deductive reasoning aims to provide logically valid and conclusive results.

Inductive reasoning involves drawing general conclusions or patterns based on specific observations or evidence. It moves from specific instances to broader generalizations. Inductive reasoning does not guarantee absolute certainty but provides probable conclusions based on available evidence (Wang et al. , 2023i).

Abductive reasoning is the process of making plausible explanations or hypotheses to account for observed facts or data. It involves inferring the best possible explanation from incomplete or limited information. Abductive reasoning is often used in problem-solving and hypothesis generation.

In commonly used terms of reasoning, for a non-fallacious argument (an argument consisting of a premise and a conclusion) (Flach and Kakas, 2000), a deductive argument is classified as such when the premise can offer conclusive support for the conclusion. In other words, if all the premises of the argument are true, it would be impossible for the conclusion to be false. On the other hand, an inductive argument is characterized by the premise providing only partial support for the conclusion (Salmon et al. , 1989). In the case of inductive arguments, the conclusions extend or surpass the information contained in the premises (Salmon et al. , 1989). Unlike deductive arguments that provide conclusive proof or inductive arguments that offer partial support, abductive arguments aim to provide the most reasonable explanation for a given situation, even if it may not be the only possible explanation.

Typically, in the trio of reasoning types, which includes deduction, abduction, and induction, the most extensively studied and explored is deduction, while research on abduction and induction has remained relatively limited and under-explored (Flach and Kakas, 2000; Yang et al. , 2023e). Encouragingly, progress has been made recently in the field of inductive reasoning. Sinha et al. (2019) propose the CLUTRR dataset for classifying kinship relations in short stories using Natural Language Understanding (NLU). Inductive Relation Induction (Yang et al. , 2022c) investigates the prediction of relation that involves unseen entities. Misra et al. (2022) focus on classifying synthetic language sentences using neural networks, whereas Yang and Deng (2021) have studied rule induction using quasi-natural language (symbolic rather than natural language).

Other taxonomies of reasoning tasks include:

(a) **Formal Reasoning vs. Informal Reasoning** (Evans and Thompson, 2004; Teig and Scherer, 2016): This taxonomy is based on the nature or formality of the

reasoning process. Formal reasoning involves following strict rules, logical frameworks, or formal systems to derive conclusions and often relies on mathematical or deductive reasoning. Informal reasoning, on the other hand, is less structured and more intuitive, relying on personal experiences, common sense, and heuristics.

(b) **Neural Reasoning vs. Symbolic Reasoning vs. Neural-Symbolic Reasoning** (Garcez et al. , 2008, 2015, 2022): This taxonomy is based on the underlying computational framework used for reasoning. Neural reasoning refers to approaches that utilize neural networks or deep learning models for reasoning tasks. Symbolic reasoning involves using symbolic representations, logic-based inference rules, or symbolic manipulation for reasoning. Neural-symbolic reasoning combines elements of both neural networks and symbolic reasoning, aiming to integrate their respective strengths.

(c) **Backward Reasoning vs. Forward Reasoning** (Al-Ajlan, 2015): This taxonomy is based on the direction of the reasoning process. Backward reasoning starts from a goal or desired outcome and works backward by applying rules or evidence to determine the necessary conditions or steps to reach that goal. Forward reasoning starts with initial premises or evidence and progresses step-by-step to derive new conclusions or reach a final outcome.

(d) **Single-step Reasoning vs. Multi-step Reasoning** (Song et al. , 2018; Yu et al. , 2023a): This taxonomy is based on the complexity or number of steps involved in the reasoning process. Multi-step reasoning refers to tasks that require multiple sequential or interconnected steps to arrive at a solution or conclusion. It involves chaining together intermediate steps or inferences to reach the final result.

(e) **Deductive Reasoning vs. Defeasible Reasoning** (Yu et al. , 2023a; Koons, 2005; Pollock, 1987, 1991): The classification criterion for this type of reasoning is based on the nature of the reasoning process and the handling of exceptions or conflicting information. Defeasible reasoning involves reasoning under uncertainty or with incomplete information, where conclusions can be overridden or defeated by new evidence or exceptions. It allows for the revision or re-evaluation of conclusions based on additional information or context.

(f) **Unimodal Reasoning vs. Multimodal Reasoning** (Sowa, 2003; Oberlander et al. , 1996): This taxonomy is based on the input modalities used in the reasoning process. Unimodal reasoning refers to reasoning tasks that involve a single modality of information or input, for example, reasoning tasks that are based solely on language information. Multimodal reasoning, on the other hand, involves integrating and reasoning with multiple modalities of information simultaneously. This could include combining visual, language, textual, auditory, or other types of input for the reasoning process.

In addition to the categorization mentioned above, there are several other ways to classify or categorize information and reasoning, including factual reasoning (Byrne and Tasso, 1999), counterfactual reasoning (Bottou et al. , 2013), plausible (defeasible) reasoning (Collins and Michalski, 1989), default reasoning (Brewka, 2012), and abstract reasoning (Yu et al. , 2021).

2.1.2 Mathematical Representation

By acknowledging the above diverse definitions and perspectives, we gain a richer understanding of reasoning as a multifaceted concept that spans philosophical inquiry, formal logic, and practical applications in fields such as NLP. In this section, we will explore the commonalities and distinct characteristics of reasoning across these domains and investigate the mathematical methodologies employed to advance our understanding and implementation of reasoning processes. Here are examples of illustrating reasoning in different mathematical frameworks:

Propositional Logic

Logical proposition: Let p and q be logical propositions. We can represent their conjunction (AND) as $p \wedge q$. Modus Ponens: If $p \rightarrow q$ and p are true, then we can conclude q. This can be represented as $(p \rightarrow q) \wedge p \rightarrow q$.

Predicate Logic

Quantifier and Predicate: Let $P(x)$ be a predicate representing "x is a prime number." The existential quantifier ($\exists$) can be used to express the existence of a prime number, such as $\exists x P(x)$. Universal Quantifier: Let $Q(x)$ be a predicate representing "x is an even number." The universal quantifier ($\forall$) can be used to express that all numbers are even, such as $\forall x Q(x)$.

Set Theory

Set Intersection: Let A and B be sets. The intersection of A and B is denoted as $A \cap B$. Set Complement: Let A be a set. The complement of A is denoted as A'.

Graph Theory

Graph Representation: Let $G = (V, E)$ be a graph, where V represents the set of nodes and E represents the set of edges. Shortest Path: Let $d(u, v)$ represent the shortest path between nodes u and v in a graph. The shortest path problem can be formulated as finding the minimum value of $d(u, v)$ for all pairs of nodes.

Conditional Probability

Let $P(A)$ represent the probability of event A and $P(B)$ represent the probability of event B. The conditional probability of A given B is denoted as $P(A|B)$ and can be calculated using Bayes' theorem.

Formal Systems

Axiomatic System: Let S be an axiomatic system with a set of axioms and a set of inference rules. A formal proof within the system can be represented as a sequence of statements, where each statement is either an axiom or derived using the inference rules.

These mathematical expressions provide a glimpse into how reasoning can be expressed mathematically in different frameworks. However, it is important to note

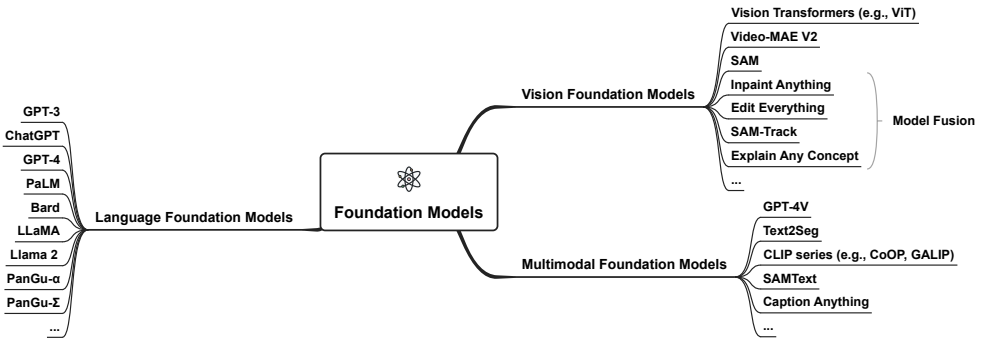


Fig. 3: Foundation models can be mainly categorized into language, vision, and multimodal foundation models, each of which is an actively researched area.

that the complexity of reasoning problems often requires more elaborate mathematical expressions and formalisms.

Despite these traditional categorizations and rigorous mathematical representations, with the advent of foundation models, researchers have increasingly moved away from strict adherence to these restrictions. Instead, they have embraced a more flexible approach to reasoning, considering its various forms and applications in different scenarios.

In contemporary research, reasoning has evolved to encompass a wide range of tasks and contexts. For instance, Commonsense Reasoning has emerged as a vital area for study, aiming to endow AI systems with the ability to understand and reason about everyday situations, incorporating common knowledge and contextual understanding. An example illustrating Commonsense Reasoning is shown in Table 1. Similarly, Mathematical Reasoning has garnered significant attention, particularly in the context of foundation models. Researchers are exploring ways to enhance models' mathematical reasoning abilities, including solving math word problems. An example showcasing Mathematical Reasoning, specifically a Math Word Problem, is presented in Table 2.

These examples highlight the diverse manifestations of reasoning in different application domains. The focus has shifted from rigid categorizations to addressing specific reasoning challenges and designing models capable of tackling them effectively. By embracing this more flexible and application-driven perspective, researchers aim to broaden the scope of reasoning and advance the development of AI systems capable of exhibiting human-like reasoning capabilities across a wide array of tasks and contexts.

2.2 Foundation Models and Recent Progress

In recent years, the field of artificial intelligence has witnessed rapid development of foundation models. Foundation models have revolutionized various domains, including but not limited to computer vision, natural language processing, and speech recognition. Next, we introduce three main categories of the foundation model and their representative works, as summarized in Figure 3.

2.2.1 Language Foundation Models and Language Prompt

Foundation models, such as GPT-3 (Brown et al. , 2020), herald breakthroughs in natural language understanding and generation tasks first. These models have shown the ability to understand and generate coherent, contextually appropriate responses in natural language and have achieved significant progress in various language-related tasks, including text completion, translation, dialogue, summarization, question answering, and beyond.

Recently, with the advancements in research and refined training methodologies, a variety of advanced large-scale language models (Zhao et al. , 2023b)[§] have emerged. Prominent among them are GPT-4 (OpenAI, 2023a), which powers ChatGPT, and PaLM (Chowdhery et al. , 2022), a crucial component of Bard. Additionally, LLaMA (Touvron et al. , 2023a) and Llama 2 (Touvron et al. , 2023b) have gained popularity as a collection of open-source large language models, varying in parameters from 7B to 65B. The focus on multilingual support has also become a key area of interest in foundation modeling research. For instance, PanGu-α (Zeng et al. , 2021), pre-trained on 1.1 TB of Chinese data and has 200 billion parameters, shows robust language modeling capabilities. Taking the concept further, PanGu-Σ (Ren et al. , 2023) utilizes techniques like Random Routed Experts (RRE) and Expert Computation and Storage Separation (ECSS) to develop a system that trains a trillion-parameter language model, leading to a significant 6.3x increase in training throughput through heterogeneous computing.

2.2.2 Vision Foundation Models and Visual Prompt

Following the remarkable success of foundation models in the language domain, its implications transcend to the realms of the vision field as well.

Vision Transformer (ViT) (Dosovitskiy et al. , 2021) applies the Transformer framework to computer vision tasks, achieving impressive performance in classification and retrieval tasks by leveraging self-attention mechanisms. Swin Transformer (Liu et al. , 2021b) introduces a hierarchical structure with shifted windows, improving the efficiency of processing high-resolution images. It has demonstrated strong performance across various computer vision tasks such as image classification, object detection, and semantic segmentation. Methods like MAE (He et al. , 2022), BEIT (Bao et al. , 2021), and CAE (Chen et al. , 2023g) propose masked modeling as an efficient self-supervised learning strategy to learn general-purpose visual representations. VideoMAE V2 (Wang et al. , 2023e) is an enhanced version of VideoMAE (Tong et al. , 2022), with a billion parameters, designed for video understanding tasks. It utilizes self-supervised learning to learn temporal and spatial dependencies, excelling at tasks like action classification and action detection. As multitask vision foundation models, Florence (Yuan et al. , 2021) and Florence-2 (Ding et al. , 2022; Xiao et al. , 2023a) can be easily adapted for a variety of computer vision tasks, such as classification, retrieval, object detection, visual question answering (VQA), image captioning, video retrieval, and action recognition, etc. Segment Anything Model (SAM) (Kirillov et al. , 2023) excels at producing object masks from input prompts like partial

[§] https://github.com/RUCAIBox/LLMSurvey

masks, points, or boxes. It has the capability to generate masks for all objects in an image. SAM is trained on a vast dataset that includes 11 million images and 1.1 billion masks. Notably, SAM demonstrates zero-shot performance across a wide range of segmentation tasks. As a zero-shot anomaly segmentation, Segment Any Anomaly+ (SAA+) (Cao et al. , 2023) introduces hybrid prompt regularization, leveraging domain-specific expertise and contextual information from the target image to enhance the adaptability of foundational models. By incorporating these elements into the regularization prompt, SAA+ strengthens the prompt's robustness, enabling more precise identification of anomalous regions. Furthermore, Wang et al. (2023a) have also revealed the potential of incorporating domain expert knowledge as prior support in addressing segmentation challenges in complex scenes.

Model Fusion: Enhancing Visual Task through Combination

There is a recent trend in the field of computer vision to combine different pre-trained Vision Foundation Models, each specializing in specific tasks, in order to tackle complex visual tasks more effectively. These approaches take advantage of the increasing power and diversity of these foundation models, leveraging their individual strengths to achieve superior performance in challenging visual tasks.

Inpaint Anything (Yu et al. , 2023d) presents three essential functionalities in image inpainting, namely Remove Anything, Fill Anything, and Replace Anything, which are achieved through the synergistic combination of various foundational models. It leverages click prompts for automatic segmentation, utilizes state-of-the-art inpainting models like LaMa (Suvorov et al. , 2021) and Stable Diffusion (Rombach et al. , 2021) for filling masked regions, and employs AI models with text prompts to generate specific content for filling or replacing voids.

Edit Everything (Xie et al. , 2023a) presents a generative system that combines SAM (Kirillov et al. , 2023), CLIP (Radford et al. , 2021), and Stable Diffusion (Rombach et al. , 2021) to enable image editing guided by both image and text inputs. Initially, Edit Everything (Xie et al. , 2023a) employs SAM to segment the original image into several fragments. Subsequently, the process of image editing is guided by text prompts, leading to a transformation that adjusts the source image to correspond with the target image as described in the given text prompts.

SAM-Track (Cheng et al. , 2023) introduces a video segmentation framework that integrates Grounding-DINO (Liu et al. , 2023h), DeAOT (Yang and Yang, 2022), and SAM (Kirillov et al. , 2023) to facilitate interactive and automated object tracking and segmentation across multiple modalities. The framework allows interactive prompts, including click-prompt, box-prompt, and text-prompt, in the initial frame of the video to guide SAM's segmentation process. Explain Any Concept (EAC) (Sun et al. , 2023a) presents an approach for concept explanation, utilizing SAM for initial segmentation and introducing a surrogate model to enhance the efficiency of the explanation process.

2.2.3 Multimodal Foundation Models

As foundation models continue to exhibit impressive performance on individual modalities, such as language and images, a natural extension arises: Can these models effectively handle multimodal data? This question arises from the recognition that

real-world scenarios often involve multiple modalities, such as text, images, and audio, which collectively provide a more comprehensive and nuanced understanding of the data.

Text2Seg (Zhang et al. , 2023d) introduces a vision-language model that leverages text prompts as input to generate segmentation masks. The model operates by using a text prompt to generate bounding boxes with Grounding DINO (Liu et al. , 2023h), which guides SAM in generating segmentation masks. CLIP (Radford et al. , 2021) learns joint representations of images and text. It achieves this by aligning visual and textual information, enabling cross-modal understanding, and demonstrating impressive capabilities in various vision and language tasks. Similarly, methods (Chen et al. , 2020b; Li et al. , 2020; Zhang et al. , 2021; Zhai et al. , 2022; Yao et al. , 2022b; Jia et al. , 2021; Huo et al. , 2021; Fei et al. , 2022), like ALIGN (Jia et al. , 2021) and WenLan (Huo et al. , 2021), align image and text representations by learning a common feature space. CoOp (Context Optimization) (Zhou et al. , 2022) presents a straightforward technique to customize CLIP-like vision-language models for downstream tasks. CoOp employs learnable vectors to represent the context words in a prompt while maintaining the pre-trained parameters in a fixed state. GALIP (Generative Adversarial CLIPs) (Tao et al. , 2023) is another advancement, specifically developed for the task of text-to-image generation. In CLIP Surgery (Li et al. , 2023q), heatmaps are first generated based on text prompts. Point prompts, which are then sampled from these heatmaps, are then inputted into SAM (Kirillov et al. , 2023) for further processing. Following this, a similarity algorithm utilizing CLIP (Radford et al. , 2021) is employed to produce the final segmentation map. SAMText (He et al. , 2023) presents a flexible approach for creating segmentation masks tailored to scene text. This method initiates by deriving bounding box coordinates from the annotations present in an existing scene text detection model. These coordinates then prompt SAM to generate masks. Caption Anything (Wang et al. , 2023j) presents a foundational model-enhanced framework for image captioning that enables interactive multimodal control from both visual and linguistic aspects. By combining SAM (Kirillov et al. , 2023) with ChatGPT, users gain the flexibility to manipulate images using a variety of prompts, including points prompts or bounding boxes prompts, during interaction. It additionally leverages Large Language Models (LLMs) to refine instructions, ensuring they accurately reflect the user's intended meaning and remain consistent with their intention. GPT-4V(ision) empowers users to interpret and analyze user-provided image inputs (OpenAI, 2023b).

The potential for foundation models to excel in multimodal tasks (text-to-image, text-to-code, and speech-to-text) opens up exciting possibilities in various domains. By seamlessly integrating and processing information from different modalities, these models can enhance tasks such as image captioning, visual question answering, and audio-visual scene understanding. Moreover, multimodal foundation models hold promise in applications that require reasoning and decision-making based on multiple sources of information. By harnessing the power of multimodal data, these models have the potential to unlock new levels of understanding, context awareness, and performance across a wide range of domains, including robotics (Firoozi et al. ,

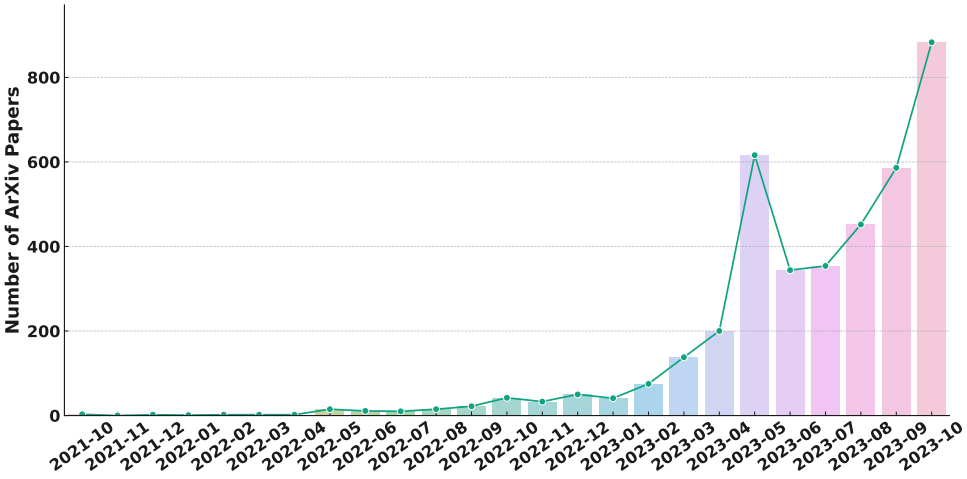


Fig. 4: Number of arXiv Papers on "Reasoning with Large Language Models" over the past two years. It depicts a rising trend in the research interest, with the number of articles surging notably in the months of 2023.

2023), healthcare (Qiu et al. , 2023a), autonomous vehicles (Zhou et al. , 2023d), and multimedia analysis.

2.2.4 Potential for Applications in Reasoning

Reasoning with foundation models is an emerging field. Recently there has been an influx of research that attempts to apply foundation models to reasoning tasks, and promising results have been achieved. The statistics are presented in Figure 4. Laban et al. (2023) identify challenges in evaluating complex tasks with Large Language Models (LLMs) and highlight the need for improved evaluation benchmarks. Shi et al. (2023) demonstrate that multilingual language models can go beyond language and perform tasks like commonsense reasoning and semantic judgment in a word-in-context setting. Language models serve as multilingual reasoners employing chain-of-thought processes. Self-Taught Reasoner (STaR) (Zelikman et al. , 2022) enhances a model's reasoning abilities by iteratively generating rationales and fine-tuning based on correct answers. MWP-BERT (Liang et al. , 2022b) leverages both BERT (Kenton and Toutanova, 2019) (110M) and RoBERTa (Liu et al. , 2020b) (123M) pre-training to tackle Math Word Problem (MWP) solving. Meanwhile, Minerva (Lewkowycz et al. , 2022), based on the PaLM (Chowdhery et al. , 2022) pre-trained language model, boasts an impressive parameter size of up to 540B. Minerva demonstrates strong performance by accurately answering nearly a third of over two hundred undergraduate-level problems in various disciplines like chemistry, biology, economics, physics, and other sciences that involve quantitative reasoning. Zero-shot-CoT (Kojima et al. , 2022) demonstrates impressive performance across a range of reasoning tasks, including arithmetic challenges such as MultiArith (Patel et al. , 2021a), GSM8K (Cobbe et al. , 2021), AQUA-RAT (Ling et al. , 2017a),

SVAMP (Patel et al. , 2021a), symbolic reasoning, and other logical reasoning tasks like Date Understanding (Srivastava et al. , 2023), Tracking Shuffled Objects (Srivastava et al. , 2023), all without the necessity for handcrafted few-shot examples. Employing just one prompt template, this approach indicates the zero-shot potential and the high-level, multi-task cognitive capacities of LLMs, while also emphasizing the significant prospects for additional research in this field.

However, there is still a need for intelligent systems that can perform more sophisticated forms of reasoning, beyond simple pattern recognition.

3 Reasoning Tasks

In this section, we provide a concise overview of various reasoning tasks, as Figure 2 shows. Here, we present distinct categories of reasoning approaches and tasks:

- Commonsense Reasoning (Section 3.1): Exploring the capacity to infer and apply everyday, intuitive knowledge.
- Mathematical Reasoning (Section 3.2): Focusing on the ability to solve mathematical problems and derive logical conclusions.
- Logical Reasoning (Section 3.3): Examining the process of drawing inferences and making decisions based on formal logic.
- Causal Reasoning (Section 3.4): Investigating the understanding of cause-and-effect relationships and their implications.
- Multimodal Reasoning (Section 3.7): Involving reasoning across multiple data modalities, such as text, images, and sensory information.
- Visual Reasoning (Section 3.5): Focusing on tasks that require the interpretation and manipulation of visual data.
- Embodied Reasoning (Section 3.8): Exploring reasoning in the context of embodied agents interacting with their environment.
- Other Reasoning Tasks (Section 3.9): The discussion of reasoning extends across various contexts, including conceptual frameworks, such as abstract reasoning 3.9.7, defeasible reasoning 3.9.8, as well as applied fields such as medical reasoning 3.9.3, bioinformatic reasoning 3.9.4, among others. We also highlight the immense utility of long-chain reasoning in applications for researchers to explore 3.9.6.

This comprehensive overview provides insights into the diverse landscape of reasoning tasks and approaches within the field. A summary of seminal works in each reasoning sector can be found in Figure 5.

3.1 Commonsense Reasoning

Commonsense reasoning refers to the human-like capacity to make assumptions and inferences about the nature and characteristics of everyday situations that humans encounter on a regular basis[§].

Recent research indicates that language models are capable of acquiring certain aspects of common sense knowledge (Zhao et al. , 2023f; Ye et al. , 2023b). In the

[§] http://www-formal.stanford.edu/leora/commonsense/

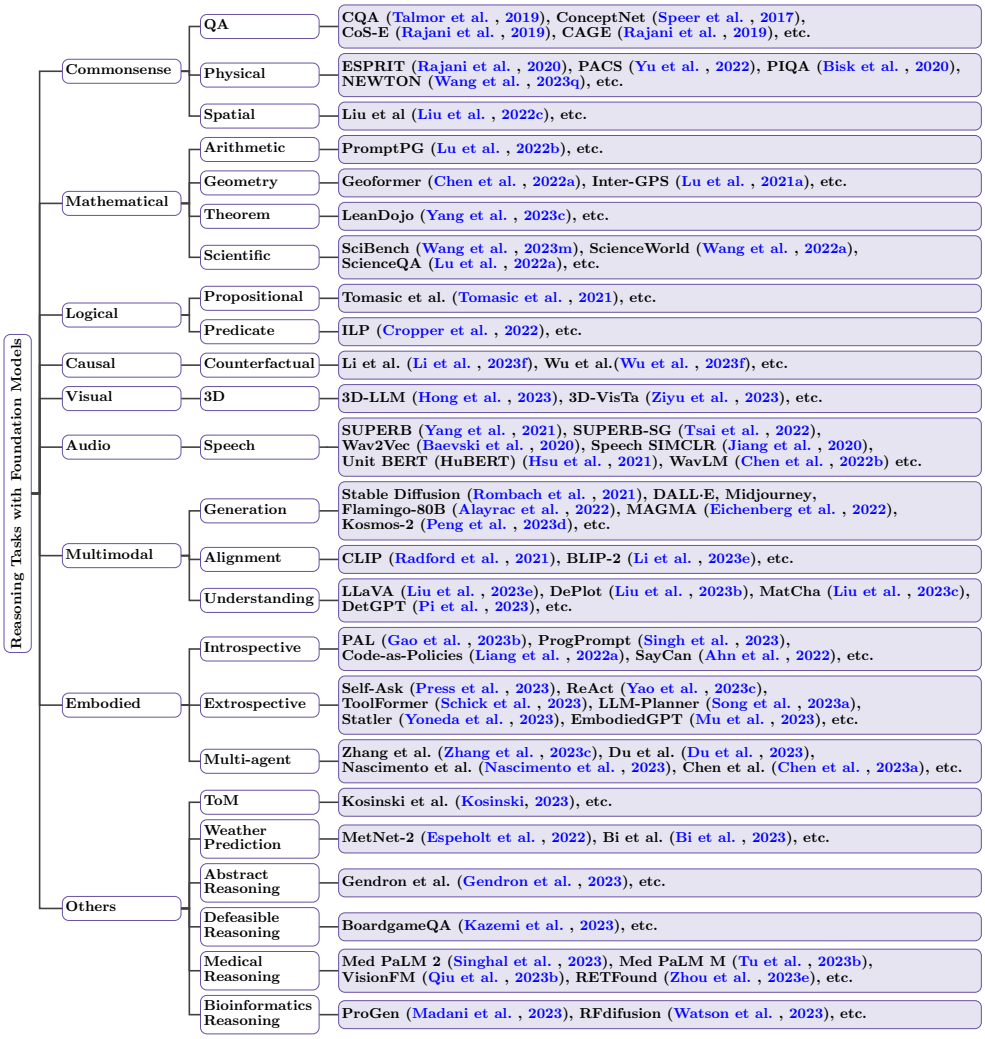


Fig. 5: Taxonomy of Reasoning Tasks with Foundation Models. Only the representative approaches for each type of task are listed.

domain of structured commonsense reasoning, Madaan et al. (2022a) tackle the task by generating a graph based on natural language input. They formalize this problem as a code generation challenge, utilizing large language models that are prompted with code to construct the graph representation. Berglund et al. (2023) also point out that language models often demonstrate a fundamental lapse in logical deduction, failing to generalize a common pattern in their training set, specifically, the likelihood of "B is A" occurring if "A is B" is present. Li et al. (2022e) take a systematic approach to evaluate the performance of large pre-trained language models on various commonsense benchmarks. They conduct zero-shot and few-shot commonsense evaluations

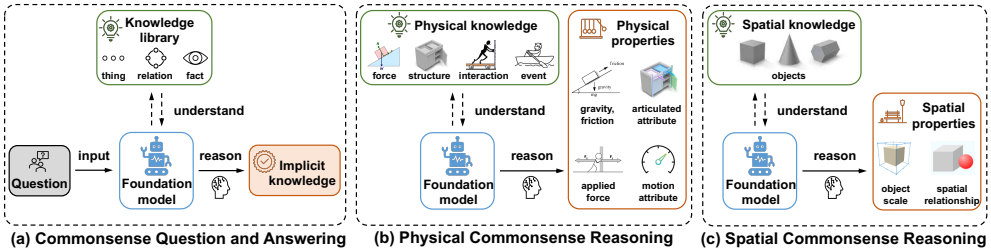


Fig. 6: Three areas of research of foundation models in commonsense reasoning. (a) By understanding everyday knowledge, foundation models can reason about implicit knowledge from questions and deduce answers. (b) Foundation models infer a wide range of physical properties from general physical knowledge. (c) Foundation models reason about spatial properties from a set of objects.

across four different benchmarks, considering six different model sizes. Notably, their evaluation includes a remarkably large language model with 280 billion parameters. Multiple evaluation settings, such as different score functions and prompt formats, are explored to comprehensively assess the models' ability to capture and reason about commonsense knowledge.

Another direction in the field of commonsense reasoning involves combining pre-trained language models with commonsense-specific fine-tuning techniques. Chang et al. (2021) propose several architectural variations, leverage external commonsense corpora, and employ commonsense-specific fine-tuning techniques for the Social IQA task (Sap et al. , 2019). Through their work, they demonstrate that these optimizations can enhance the model's performance in tasks related to social intelligence. Furthermore, Yang et al. (2023a) introduce a two-stage framework designed to connect pre-training and fine-tuning in the task of commonsense generation.

In addition to the above-mentioned works, there are other aspects of commonsense reasoning that have been explored. These include commonsense question answering (QA), physical reasoning, spatial reasoning, and the corresponding benchmarks, as shown in Figure 6. These areas of research contribute to a deeper understanding of how language models can effectively capture and reason about commonsense knowledge in various contexts.

3.1.1 Commonsense Question and Answering (QA)

As a subfield of commonsense reasoning, Commonsense Question Answering (QA) focuses on developing systems capable of answering questions that require a deep understanding of everyday knowledge and human-like reasoning. Unlike traditional fact-based QA, where answers can be derived from explicit information, commonsense QA involves understanding and reasoning about implicit knowledge and everyday human reasoning, as depicted in Figure 6(a).

The Commonsense Question Answering (CQA) dataset (Talmor et al. , 2019) is a challenging multiple-choice dataset specifically designed for commonsense question

answering. It is derived from ConceptNet (Speer et al. , 2017) and consists of approximately 12,000 questions. Every question comes with one correct answer and four additional distractor answers. In addition, the Commonsense Explanations (CoS-E) dataset (Rajani et al. , 2019) contains human commonsense explanations for the CQA dataset. The CoS-E dataset comprises two types of explanations: Selected explanations, which are text spans highlighted in the question that justify the answer choice, and open-ended explanations, which are free-form natural language explanations.

Commonsense Auto-Generated Explanation (CAGE) model (Rajani et al. , 2019) is a framework that involves training a language model to generate useful explanations by fine-tuning it using both the problem input and human-generated explanations.

The development of effective commonsense QA systems is an active area of research, and ongoing advancements in language models, knowledge representation, and reasoning techniques continue to push the boundaries of commonsense understanding in machine intelligence.

3.1.2 Physical Commonsense Reasoning

Commonsense physical reasoning (Ding et al. , 2021b), shown in Figure 6(b), involves utilizing everyday knowledge about the physical world to reason and understand the behavior of objects and their properties. It encompasses reasoning about physical concepts, such as the properties of objects reasoning (gravity, mass, inertia, or friction), their affordances, and how they can be manipulated.

Explaining Solutions to Physical ReasonIng Tasks (ESPRIT) framework (Rajani et al. , 2020) combines commonsense physical reasoning with interpretability via natural language explanations. It operates in two stages: firstly, pinpointing key physical events in tasks, and secondly, crafting natural language descriptions for both the initial scene and these crucial events. The framework aims to provide a unified approach to reasoning about commonsense physical concepts, such as gravity, friction, and collision, while also offering qualitative explanations using natural language. PACS (Physical Audiovisual CommonSense) (Yu et al. , 2022) is a dataset designed for physical audiovisual commonsense reasoning. It comprises 13,400 question-answer pairs, including 1,377 distinct questions and 1,526 videos for physical commonsense. By benchmarking unimodal and multimodal reasoning models, PACS identifies the limitations and areas of improvement in current models, thereby providing valuable opportunities to propel research in physical reasoning by examining multimodal reasoning approaches. PIQA (Physical Interaction: Question Answering) (Bisk et al. , 2020) is a dataset that focuses on multiple-choice question-answering in the domain of physical interactions. The task involves selecting the most appropriate solution from two given options based on a given question. The PIQA dataset consists of over 16,000 training QA pairs, with additional data reserved for development and testing. The questions in PIQA have an average length of 7.8 words, while both correct and incorrect solutions have an average length of 21.3 words. NEWTON (Wang et al. , 2023q) is a comprehensive platform that serves as a repository, pipeline, and benchmark specifically created to assess the physical reasoning capabilities of LLMs.

CATER (Girdhar and Ramanan, 2020) mainly focuses on physics-related visual scenes. CLEVRER (Yi et al. , 2019) is a video question-answering benchmark that

targets the physical and causal relations grounded in dynamic videos of rigid-body collisions. CLEVRER-Humans (Mao et al. , 2022) further extends it to the causal judgment of physical events with human labels. Physion (Bear et al. , 2021), Physion++ (Tung et al. , 2023), and ComPhy (Chen et al. , 2022d) evaluate objects with different latent physical properties (e.g., mass, friction, elasticity, and deformability) from dynamic videos rendered from physics engines.

Based on the above benchmarks, transformer-based foundational models (Ding et al. , 2020; Wu et al. , 2022b) and neuro-symbolic frameworks with differentiable physics (Ding et al. , 2021b) are developed. Aloe (Attention over Learned Object Embeddings) (Ding et al. , 2020) integrates MONet (Burgess et al. , 2019) for unsupervised object segmentation with self-attention mechanisms, facilitating spatio-temporal physical reasoning about objects. SlotFormer (Wu et al. , 2022b), a Transformer-based object-centric dynamics model, is designed to unsupervisedly decipher complex systems and interactions from videos. Utilizing a context encoding provided by Spatial Transformer (Jaderberg et al. , 2016), Generative Structured World Models (G-SWM) (Lin et al. , 2020c) advance object-centric world modeling. They incorporate multimodal uncertainty and situational awareness through a core module known as Versatile Propagation (V-Prop). These frameworks and datasets contribute to the advancement of commonsense physical reasoning by providing resources for model evaluation, interpretability, and understanding physical concepts through explanations and multimodal analysis.

Currently, the physical commonsense reasoning domain based on foundation models is relatively unexplored, offering a ripe avenue for research and development. This presents a unique chance for researchers and practitioners to delve into and expand the boundaries of what's possible with these models, potentially leading to groundbreaking advancements and innovations.

3.1.3 Spatial Commonsense Reasoning

As illustrated in Figure 6(c), spatial commonsense reasoning involves detecting the spatial position of objects and inferring the relationships between visual stimuli to understand the surrounding environment. Within the domain of spatial commonsense reasoning, two significant perspectives are object scales (Aroca-Ouellette et al. , 2021) and spatial relationship (Hudson and Manning, 2019). Liu et al. (2022c) introduce a spatial commonsense benchmark, distinctly highlighting the relative sizes of objects and the spatial interactions between individuals and objects across various actions. They investigate the performance of various models, including pre-trained vision-language models and image synthesis models. Interestingly, they find that the models for synthesizing images demonstrate better capabilities in learning accurate coherent knowledge of spatial relationships compared to other models. Furthermore, the spatial insights obtained through these models for synthesizing images also demonstrate their utility in enhancing natural language understanding tasks that necessitate spatial commonsense reasoning.

3.2 Mathematical Reasoning

Mathematics distinguishes itself as a distinct language that relies on symbolic forms, and precision in meaning and possesses lower dimensionality compared to natural language. This unique characteristic allows us to demonstrate that meaning can be derived from a set of learned rule sets, as exemplified by the symbolic representations of mathematical concepts (Floyd, 2004). Mathematical problems can be effectively programmed when they are represented using symbols and corresponding expressions. By formulating these problems in a computer language that can be translated into machine code, deep learning-based reasoning systems have the ability to train on and acquire the underlying rules (Hinton, 1990; Schmidhuber, 2015; Friedman, 2023b).

Experimental findings suggest that the performance of Large Language Models (LLMs) shows a weak correlation with question difficulty. Ling et al. (2017b) propose an approach to solve algebraic word problems in a way that not only generates the answer but also provides an explanation or rationale for the obtained result. MT2Net (Zhao et al. , 2022b) is a specialized model designed to tackle the MultiHiertt dataset (Zhao et al. , 2022b). It retrieves supporting facts from financial reports and generates executable reasoning programs to answer questions. This approach aims to provide a comprehensive and accurate solution for the given questions.

3.2.1 Arithmetic Reasoning

Math Word Problems (MWPs) are commonly used to evaluate the arithmetic reasoning abilities of language models. While these issues may appear uncomplicated to humans, language models frequently encounter challenges when it comes to tasks involving arithmetic reasoning (Hendrycks et al. , 2021b; Patel et al. , 2021b).

Previous research has explored various approaches to address these challenges. Template-based statistical learning methods like KAZB (Kushman et al. , 2014), ZDC (Zhou et al. , 2015), and similarity-based method SIM (Huang et al. , 2016) have been utilized. Wang et al. (2017) employs a recurrent neural network (RNN) to convert math word problems into equation templates, eliminating the need for complex feature engineering. Additionally, they developed a hybrid model that integrates the RNN with a similarity-based retrieval system, further enhancing its performance. Xie and Sun (2019) introduces an innovative neural approach to construct expression trees in a goal-oriented manner for solving math word problems. Shen et al. (2021a) introduces a novel ranking task for math word problems and presents the Generate & Rank framework, which combines a generative pre-trained language model with multi-task learning. This approach allows the model to learn from its errors and effectively differentiate between correct and incorrect expressions. A notable finding is that employing chain-of-thought prompting, along with a language model containing an impressive 540 billion parameters, yields performance comparable to task-specific fine-tuned models across multiple tasks (Wei et al. , 2022b). Unlike traditional symbolic reasoning tasks such as program synthesis and knowledge graph reasoning, solving MWPs requires additional emphasis on numerical reasoning. PromptPG (Lu et al. , 2022b) takes a different approach by utilizing policy gradient techniques to learn the selection of in-context examples. By dynamically constructing appropriate prompts for each test example, PromptPG facilitates the solving of math word problems. This

adaptive approach enhances the model's ability to handle numerical reasoning tasks effectively.

3.2.2 Geometry Reasoning

GeoS (Seo et al. , 2015) provides a system for mapping geometry word problems into a logical representation, facilitating the process of problem-solving. Chen et al. (2021a) introduce Neural Geometric Solver (NGS) as an approach to addressing challenges posed by geometric problems in the GeoQA benchmark (Chen et al. , 2021a). NGS adopts a holistic approach, adeptly parsing multimodal information and generating interpretable programs. Geoformer (Chen et al. , 2022a) concurrently addresses calculation and proving problems through sequence generation. This approach demonstrates improved reasoning capabilities in both tasks by employing a unified formulation. Additionally, the authors propose the Mathematical Expression Pretraining (MEP) method, predicting mathematical expressions within problem solutions (Chen et al. , 2022a). This technique enhances the model's ability to handle mathematical expressions effectively. Inter-GPS (Lu et al. , 2021a) formulates the geometry-solving task as a problem-goal-searching process. By incorporating theorem knowledge as conditional rules, Inter-GPS enables step-by-step symbolic reasoning, facilitating effective geometry problem-solving.

3.2.3 Automated Theorem Proving

Theorem proving is pivotal in both hardware and software verification (Khan et al. , 2020; Li et al. , 2005). In the context of hardware verification, it has found successful application in the design of integrated circuits (Khan et al. , 2020; Li et al. , 2005). In the realm of software verification, a notable achievement is the development of CertC, a verified C compiler (Berghofer and Strecker, 2004). It is worth mentioning that companies such as Intel have made significant investments in formal methods to ensure the absence of critical floating-point bugs in their processors. A prominent example of the consequences of such bugs is the costly Pentium FDIV bug in 1994, which resulted in a loss of $500 million (Harrison, 2010). Consequently, theorem proving has played a pivotal role in verifying floating-point firmware (Harrison, 2010). Traditionally, theorem proving has relied on highly trained human experts proficient in specific theorem proving tools and their respective application domains. However, the emergence of learnable automated theorem proving holds the potential to revolutionize hardware and software verification in two significant ways. First, it enhances the level of automation in theorem proving, making it less reliant on human expertise and manpower. Second, it increases the adaptability of these methods, broadening their utility and applicability through machine learning.

Researchers create Contemporary mathematical verification systems based on interactive theorem provers (ITPs), including Isabelle (Paulson, 1994), Lean (de Moura et al. , 2015), Coq (Barras et al. , 1997), and Metamath (Megill and Wheeler, 2019). In recent years, various approaches have integrated machine learning with ITPs (Yang and Deng, 2019; Gauthier et al. , 2021). Validated on various datasets (PISA (Jiang et al. , 2021a), miniF2F (Zheng et al. , 2021), LeanDojo (Yang et al. , 2023c), FIMO Liu

et al. (2023a) and TRIGO (Xiong et al. , 2023b)), these approaches leverage advancements in language models (Polu and Sutskever, 2020; Han et al. , 2021; Polu et al. , 2023; jia, 2022; Lample et al. , 2022; Mikuła et al. , 2023) to recommend actions based on the current state of the proof, with a tree search identifying a sequence of correct steps using actions provided by the language model. Methods like Monte Carlo Tree Search (MCTS) (Silver et al. , 2018; Wu et al. , 2021; Laurent and Platzer, 2022) or dynamic-tree MCTS (Wang et al. , 2023c) are employed for this purpose. Previous work has demonstrated the few-shot statement autoformalization capability of large language models (LLMs) (Wu et al. , 2022a). To investigate the applicability of these findings to proof autoformalization, DSP conducted a thorough analysis using Draft, Sketch, and Proof (Jiang et al. , 2023). Subgoal-Learning (Zhao et al. , 2023c) utilizes the subgoal-goal informal proof and demonstration selection. LeanDojo (Yang et al. , 2023c) is an open-source project for Lean (Moura and Ullrich, 2021), which contains toolkits, data, models, and benchmarks. Lyra (Zheng et al. , 2023b) proposes the use of Tool Correction to mitigate LLM hallucinations and Conjecture Correction to improve the quality of generated formal proof conjectures. Following the direction of Lyra, the LEGO-Prover (Xin et al. , 2023) employs a growing skill library containing verified lemmas as skills to enhance the capability of LLMs used in theorem proving.

3.2.4 Scientific Reasoning

Scientific reasoning encompasses the cognitive abilities and problem-solving skills required for formulating, evaluating, and refining hypotheses or theories. In the case of highly developed proficiency, it also involves critical reflection on the process of acquiring and evolving knowledge through these investigative activities (Morris et al. , 2012). As mathematical reasoning forms the foundation of, we mention scientific reasoning here.

Scientific reasoning is closely relevant to AI for Science (AI4Science) (Zhang et al. , 2023i). This relevance extends across a spectrum of fields, including physics, chemistry, quantum mechanics, and more. The integration of foundation models into these domains not only enhances our understanding but also opens up new avenues for exploration and innovation. The potential for foundation models to revolutionize traditional scientific methods, accelerate discoveries, and solve complex problems is immense, making them an indispensable tool in the modern scientific landscape. Subramanian et al. (2023) examine how various factors affect the transfer learning capabilities of foundational models, such as the size of pre-trained models, dataset scale, a blend of models, and parameters outside the training distribution. Their study finds that increasing the number of model parameters can enhance performance. Furthermore, the "pre-train and fine-tune" approach is highly effective for scientific reasoning tasks, particularly in physical systems governed by Partial Differential Equations (PDEs). Horawalavithana et al. (2022) modify OpenAI's GPT-2 transformer decoder architecture to develop a 1.47 billion parameter general-purpose model specifically for chemistry. This large-scale model demonstrates proficiency not only in in-domain tasks but also in out-of-domain challenges. It is trained on a substantial corpus of 670GB of text data, encompassing approximately 53.45 million chemistry-focused scientific articles and abstracts. IBM RXN for Chemistry (Team, 2022; Manica et al. , 2023;

Das et al. , 2021) utilizes foundational models for predicting chemical reactions and procedural methodologies in chemistry. For a more comprehensive exploration of foundational models related to biology, please see Section 3.9.3 and Section 3.9.4. We will not elaborate further on biology foundation models here. Currently, most scientific reasoning research predominantly concentrates on fields like mathematics, physics, biology, and medicine (Qiu et al. , 2023a). In contrast, foundational models in the quantum realm are comparatively scarce. Building scalable foundation models for quantum systems faces several challenges, including the intrinsic complexity of quantum mechanics, limited data availability, the absence of standardized methodologies, and constraints in quantum hardware capabilities. Despite these hurdles, venturing into this promising field presents an intriguing and potentially rewarding area of exploration.

Standardization aids in advancing the field of scientific reasoning. Proposing datasets or benchmarks is a process of standardization. Currently, datasets for scientific reasoning are mainly focused on fields such as mathematics, physics, and chemistry, examples of which include SciBench (Wang et al. , 2023m), ScienceWorld (Wang et al. , 2022a), and ScienceQA (Lu et al. , 2022a). SciBench (Wang et al. , 2023m) is a specialized benchmark designed to evaluate the scientific reasoning capabilities, domain knowledge, and advanced calculation skills of LLMs in the context of college-level scientific problems. This comprehensive benchmark encompasses a meticulously curated collection of 695 problems carefully sourced from instructional textbooks. SciBench consists of two datasets. The first dataset constitutes an expansive collection of collegiate-level scientific problems sourced from mathematics, chemistry, and physics textbooks. Its primary objective is to evaluate the LLM's capacity to handle a diverse array of scientific topics and problem categories. The second dataset in SciBench, on the other hand, consists of problems sourced from computer science and mathematics undergraduate exams, forming a closed set. This closed set is intentionally crafted to gauge the LLMs' proficiency in solving precise problem-solving challenges within these particular fields. ScienceWorld (Wang et al. , 2022a) is designed to evaluate agents' scientific reasoning capabilities within an interactive text environment. This environment simulates a standard elementary school science curriculum, featuring 30 high-level task types distributed across 10 different topics. The environment supports multiple states, allowing for diverse interactions and scenarios. By abstracting the world and incorporating a wide range of objects, ScienceWorld provides a complex interactive text environment for agents to navigate and reason therein. It consists of 10 interconnected locations, each containing up to 200 types of objects. These objects span a range of categories, and common environmental items like furniture, books, and paintings. The environment provides a rich and diverse setting for agents to interact with. The action set within ScienceWorld consists of 25 high-level actions, covering actions related to the domain of science and common actions. Each step in ScienceWorld presents approximately 200,000 possible action-object pairs, although only a proportion of these pairs will have actual implications for the task at hand. ScienceQA (Lu et al. , 2022a) is a multimodal dataset comprising 21,208 multiple-choice science questions sourced from elementary and high school

science curricula. The dataset offers a richer domain diversity by covering natural science, language science, and social science topics.

These resources provide valuable platforms for testing the capabilities of foundation models in complex scientific reasoning domains, allowing for a more structured approach to assessing their reasoning abilities. The focus on these traditional sciences highlights the need for expanding the scope of datasets to encompass a wider range of disciplines, potentially leading to more diverse and comprehensive advancements in scientific reasoning.

3.3 Logical Reasoning

Logical reasoning, covering propositional and predicate logic (Table 4), is a rigorous form of thinking that involves using premises and their relations to derive conclusions that are implied by the premises (Nunes, 2012). It can serve as a fundamental basis for various domains in computer science and mathematics.

Previous studies have explored the combination of neural networks and symbolic reasoning in neuro-symbolic methods (Mao et al. , 2019; Pryor et al. , 2023; Tian et al. , 2022; Cai et al. , 2021; Sun et al. , 2021; Manhaeve et al. , 2021; Gupta et al. , 2019). However, these methods often face limitations such as specialized module designs that lack generalizability or brittleness caused by optimization difficulties. In contrast, LLMs exhibit stronger generalization abilities when it comes to logical reasoning. The Logic-LM framework (Pan et al. , 2023a) leverages LLMs and symbolic reasoning to enhance logical problem-solving (Luo et al. , 2023d). It begins by utilizing LLMs to convert natural language problems into symbolic formulations, which are then processed by deterministic symbolic solvers for inference. Additionally, a self-refinement stage is introduced, where error messages from the symbolic solver are utilized to revise the symbolic formalizations. Bubeck et al. (2023) demonstrate that the GPT-4 model can manifest logical reasoning abilities when addressing mathematical and general reasoning problems. These higher-order capabilities, often referred to as emergent properties, result from scaling the model with large datasets (Wei et al. , 2022a). Zhao et al. (2023a) employ language models for multi-step logical reasoning by integrating explicit planning into their inference procedure. This incorporation enables more informed reasoning decisions at each step by considering their future effects. Furthermore, Creswell et al. (2023) propose the Selection-Inference (SI) framework, which employs pre-trained LLMs as general processing modules. The SI framework alternates between selection and inference steps to generate a sequence of interpretable, causal reasoning steps that lead to the final answer.

Recent works leveraging LLMs for logical reasoning tasks can be categorized into two main approaches, as shown in Figure 7. The first approach is in-context learning, where specific prompts are used to elicit step-by-step reasoning from LLMs. Notable methods in this category include chain-of-thought prompting (Wei et al. , 2022b; Wang et al. , 2023n) and the least-to-most prompting approach (Zhou et al. , 2023a). These approaches enable reasoning directly over natural language, providing flexibility. However, the complexity and ambiguity of natural language can result in challenges such as unfaithful reasoning and hallucinations. The second approach is fine-tuning,

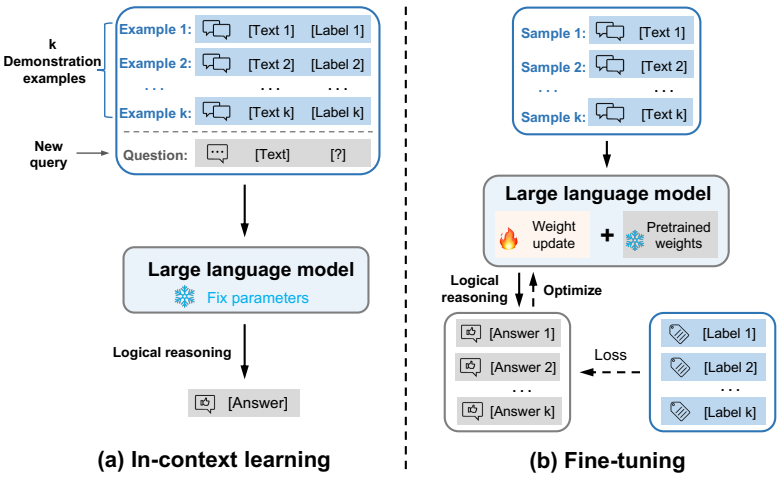


Fig. 7: Two main approaches to enhancing logical reasoning capabilities of large language models. (a) In-context learning leverages specific prompts as a demonstration to elicit logical reasoning. (b) Fine-tuning uses additional training samples to update the specialized model parameters.

	Propositional Logic	Predicate Logic
Basic elements	Atomic propositions, Compound propositions	Atomic propositions, Compound propositions, Variables, Quantifiers, Predicates
Complexity	Lower	Higher
Expressive Power	Limited	More powerful
Applications	Circuit design, Boolean algebra	Natural language processing, Knowledge representation, Database queries
Examples	$p \vee q$; $p \wedge q$; $\neg p$; $p \rightarrow q$	$\forall x, P(x)$; $\exists x, P(x)$

Table 4: Comparison between Propositional Logic and Predicate Logic in terms of basic elements, complexity, expressive power, and applications.

where the reasoning capabilities of LLMs are optimized through fine-tuning or training specialized modules (Clark et al. , 2020; Tafjord et al. , 2022; Yang et al. , 2022b).

3.3.1 Propositional Logic

Propositional logic deals with declarative sentences that can be assigned a truth value, either true or false, without any ambiguity. There are two types of propositional logic: Atomic Propositions and Compound Propositions. Atomic propositions are basic statements that cannot be further broken down, while compound propositions are formed by combining atomic propositions using logical connectives such as conjunction (AND), disjunction (OR), and negation (NOT).

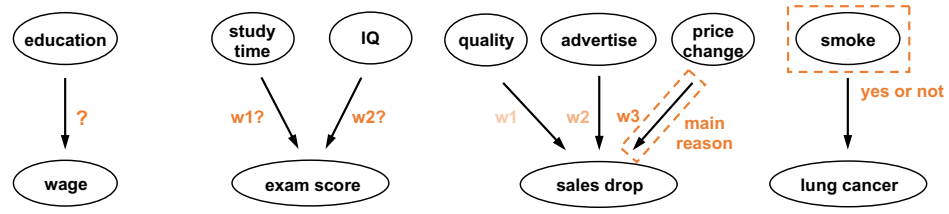


Fig. 8: Examples of causal graphs to reflect different casual reasoning tasks. (a) Causal discovery identifies the underlying causal relationships among variables in a given system. (b) Effect inference estimates the outcome (e.g., weight) of a specific intervention on a system based on known causal relationships. (c) Attribution determines the extent to which a particular cause is responsible for a given effect. (d) Judgment makes decisions based on the perceived consequences and implications of causal relationships.

In the context of propositional logic resolution, Tomasic et al. (2021) performed fine-tuning on the GPT-2 and GPT-3 models, tailoring them for the purpose of simulating propositional logic resolution. This specialized training focuses on non-recursive rules that encompass conjunction, disjunction, and negation connectors. By leveraging these language models, they aimed to enhance the logical reasoning capabilities in propositional logic problems.

The use of language models for propositional logic resolution is intriguing because these models have demonstrated their ability to capture complex patterns and semantic relationships in natural language. By training them to understand and reason with propositional logic, researchers sought to improve their logical reasoning capabilities.

3.3.2 Predicate Logic

Predicate Logic, also known as First-order Logic, can be seen as an extension of propositional logic, allowing for more nuanced expressions. In Predicate Logic, predicates are used to represent properties and provide additional information about the subject of a sentence. It involves variables with a specified domain and encompasses objects, relations, and functions between those objects.

Inductive Logic Programming (ILP) is a specialized domain within the broader field of machine learning (Cropper et al. , 2022). ILP leverages first-order logic to represent hypotheses and data, making logical language a crucial component in knowledge representation and reasoning (De Raedt and Kersting, 2010).

By incorporating predicate logical representations and reasoning, LLMs offer the potential for more interpretable and explainable models (Liu et al. , 2022b). It enables the discovery of logical patterns and rules from data, facilitating the extraction of human-understandable knowledge.

3.4 Causal Reasoning

Causal reasoning refers to the process of understanding and explaining cause-and-effect relationships between events, actions, or variables (Waldmann and Hagmayer, 2013; Liu et al. , 2023j). Causal reasoning tasks can be categorized into causal discovery, effect inference, attribution, judgment, and other tasks (Kıcıman et al. , 2023). Causal discovery is the process of uncovering the directional cause-and-effect relationships between variables. Effect inference involves the characterization of the magnitude and pattern of a known or postulated causal connection (LYU et al. , 2022; Wang et al. , 2021a; Jin et al. , 2023b). Attribution, on the other hand, entails identifying the cause or causes behind a specific change. Judgment tasks expand on attribution tasks by encompassing the assignment of reward or blame for outcomes. Additionally, these tasks encompass various domains such as policy optimization, decision-making, explanation, scientific discovery, and more.

A causal graph, also known as a causal network or causal diagram, is a graphical representation of causal relationships between variables or events (Balashankar and Subramanian, 2021; Schölkopf et al. , 2021). It is a visual tool used to depict cause-and-effect relationships and understand the causal structure of a system or phenomenon. In a causal graph, variables or events are represented by nodes, and causal relationships between them are depicted by directed edges or arrows. In Figure 8, we use causal graphs to illustrate multiple reasoning tasks mentioned above.

Causal Discovery

Causal discovery (Peters et al. , 2017) involves the task of identifying the causal graph (Long et al. , 2022) that represents the underlying process responsible for generating observed data. LLMs have demonstrated competitive performance in discerning pairwise causal connections, although their effectiveness can vary and is influenced by the careful crafting of prompts. Long et al. (2022) investigate the limitations of GPT-3 in understanding causal relationships in the medical context. Within the framework of Neuropathic Pain Diagnosis (Tu et al. , 2019), Tu et al. (2023a) find that ChatGPT tends to make false negative mistakes. The performance of LLMs in causal discovery is not yet stable or consistent, and they may provide different answers to the same question, potentially due to internal model updates. Long et al. (2023) suggest that expert knowledge, including that of LLMs, may be incorrect. They propose leveraging imperfect experts, such as LLMs, to reduce uncertainty in the output of causal discovery algorithms. By incorporating the expertise of LLMs into the statistical analysis of objective data, they aim to improve the accuracy of causal structure learning. Advancing the current research on LLM-driven causal discovery, Ban et al. (2023) integrate knowledge-based LLM causal analysis with data-driven approaches to learning causal structures. They effectively combine the expertise of LLMs regarding existing causal mechanisms with the statistical analysis of objective data. They devise a specialized set of prompts aimed at deriving causal graphs from specific variables. By employing these prompts, they evaluate the impact of LLM-informed causality on deducing causal structures from data. Compared to text-only LLMs, Code-LLMs (Liu et al. , 2023j) with code prompts are better in causal reasoning.

Type Causality and Actual Causality

Type causality pertains to the inference of causal relationships between variables, which is evident in causal discovery and causal effect estimation. In contrast, actual causality (Halpern, 2016) diverges from causal discovery by shifting the focus from variables and their interrelationships to individual events, with the goal of uncovering their specific causes.

CausaLM (Feder et al. , 2021) has demonstrated that language models like BERT (Kenton and Toutanova, 2019) can obtain a counterfactual representation of a particular concept of interest through the deliberate selection of auxiliary adversarial pre-training tasks. This counterfactual representation enables the prediction of the concept's actual causal effect on model's performance. On the other hand, Zhang et al. (2023a) believe that current LLMs can address causal questions by leveraging existing causal knowledge, akin to combined domain experts. However, these models still struggle to provide satisfactory answers when it comes to discovering new knowledge or performing high-stakes decision-making tasks with a high level of precision. Current LLMs lack the ability to incorporate actual physical data measurements to establish a grounding for their available textual facts (Willig et al. , 2023a,b). As a result, they are unable to engage in actual, inductive inference, similar to classical (causal) structure discovery methods. This limitation raises a crucial societal discussion point regarding the process of learning from facts. It is arguable that the ideal goal should be *understanding* rather than mere *knowing*, as the latter lacks both generalization and justification.

In summary, while LLMs show promise in causal discovery, their performance is still inconsistent and sensitive to prompt engineering. Researchers are exploring ways to address these limitations.

3.4.1 Counterfactual Reasoning

Counterfactuals involve a premise that is false in the real world but assumed to be true in a hypothetical scenario. For example, "If cats were vegetarians," followed by an imaginary consequence like "cats would love cabbages" (Li et al. , 2023f). Counterfactual reasoning involves the consideration of hypothetical scenarios achieved by altering elements or conditions within an actual event or situation (Kahneman and Miller, 1986; Byrne, 2007). It plays a fundamental role in understanding causality, enabling us to explore the potential outcomes that could have arisen under different circumstances. By subjecting language models to counterfactual testing, we can manipulate the factual accuracy and hypothetical nature of statements, thereby evaluating the models' capacity to discern and effectively utilize this information in making predictions. This testing approach enables us to gain insights into the models' aptitude for differentiating between actual and hypothetical scenarios and their ability to leverage this understanding for accurate and contextually appropriate responses.

In the context of language models, each reasoning task can be represented as a mapping function $f_w : X \to Y$, which maps an **input** $x \in X$ using a **world model** $w \in W$ to an **output** $y \in Y$. The world model encapsulates the conditions under which the function evaluation occurs, with the **default world** denoted as w^{default}. Hypothesis h estimates f^w, while counterfactual worlds are represented as w^{cf}. The

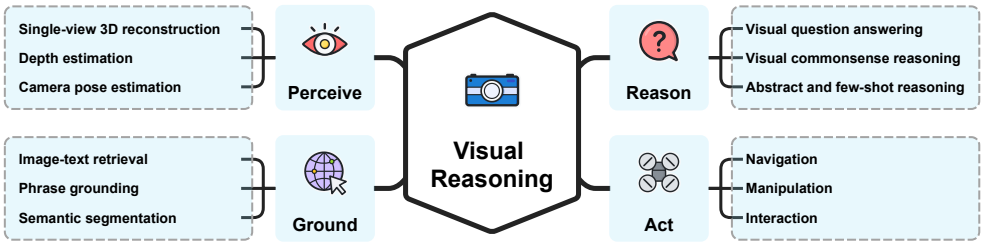


Fig. 9: Four functional domains of a general vision systmem G-VUE (Huang et al. , 2023b) and their corresponding visual tasks.

language model's implementation of f_w for a given instance x can be expressed as:

$$h(f, w, x) = \underset{y'}{\operatorname{argmax}} P_{\mathrm{LM}}(y' | \mathrm{prompt}_f(f, x), \mathrm{prompt}_w(w)), \tag{1}$$

where counterfactual reasoning is denoted as $h(f, w^{\mathrm{cf}}, x)$ and factual reasoning as $h(f, w^{\mathrm{default}}, x)$.

Li et al. (2023f) utilize counterfactual conditionals to examine the ability of pre-trained language models (PLMs) in distinguishing between hypothetical and real-world scenarios. They explore how this capability interfaces with the models' utilization of pre-existing real-world knowledge and associative cues. Their findings reveal that when confronted with counterfactual situations, PLMs tend to generate completions that contradict established world knowledge. As an example, GPT-3 might have developed a nuanced comprehension of how linguistic cues, such as distinguishing between "If/had" and "Because," influence the connection between nearby lexical clues and the following words. This suggests that PLMs might prioritize the influence of immediate contextual cues over broader factual information when responding to counterfactual prompts. Wu et al. (2023f) introduce an evaluation framework that incorporates "counterfactual" task variations. They present a set of 11 counterfactual tasks and assess the capability and performance of GPT-4 (OpenAI, 2023a), Claude (Anthropic, 2023), and PaLM-2 (Anil et al. , 2023) on these tasks, considering both default and counterfactual conditions. The findings indicate that while current language models possess some degree of abstract task-solving abilities, their performance often relies on narrow, context-specific procedures that are not easily transferable across tasks. Notably, the models consistently exhibit a significant decrease in performance when confronted with counterfactual task variants compared to the default settings. These results emphasize the need for a careful interpretation of language model performance, taking into account different aspects of their behaviors and the challenges posed by counterfactual reasoning.

3.5 Visual Reasoning

Visual reasoning refers to the cognitive process of understanding, analyzing, and drawing conclusions from visual information. It involves the ability to perceive, interpret, and reason about visual stimuli such as images, scenes, or other visual representations (Ding et al. , 2023b).

General-purpose Visual Understanding Evaluation (G-VUE) is a comprehensive evaluation framework (Huang et al. , 2023b). It aims to assess the full range of visual cognitive abilities. The framework is divided into four functional domains: Perceive, Ground, Reason, and Act. As shown in Figure 9, the framework encompasses a thoughtfully chosen collection of 11 tasks, including 3D reconstruction, visual reasoning, and manipulation, to represent these domains. G-VUE serves as a standardized and comprehensive platform for assessing the visual understanding capabilities of AI systems. By prioritizing diverse functional domains and carefully selecting tasks, the framework ensures the inclusion of a wide range of visual cognitive abilities. This enables more accurate evaluations of the strengths and weaknesses of AI systems. VLGrammar (Hong et al. , 2021a) is a model that employs compound probabilistic context-free grammars (PCFGs) to simultaneously induce language and image grammar. A novel contrastive learning framework is also proposed, which facilitates the joint learning of these two modules. AeNER (Ding et al. , 2021a) introduces a general neural-network-based approach to dynamic visual spatio-temporal reasoning problems. This approach differs from bespoke methods like modular symbolic components, independent dynamics models, or semantic parsers. AeNER offers a more versatile and adaptable solution for addressing dynamic visual spatio-temporal reasoning challenges.

3.5.1 3D Reasoning

Specifically, 3D reasoning refers to the cognitive process of understanding, analyzing, and reasoning about 3D objects or spatial arrangements.

3D-LLM (Hong et al. , 2023) is designed to process 3D point clouds along with their associated features. It demonstrates remarkable proficiency across a diverse spectrum of 3D-related tasks, encompassing dense captioning, 3D question answering, 3D grounding, 3D-assisted dialogue, navigation, task decomposition, and beyond. PointLLM (Xu et al. , 2023b) is an approach that extends LLMs to understand 3D point clouds, combining geometric, visual, and textual information to interact with and interpret 3D data. It shows superior performance in object classification and captioning tasks compared to 2D baselines. On the other hand, 3D-VisTa (Ziyu et al. , 2023) is a pre-trained Transformer model specifically developed for aligning 3D vision and text. It proves to be highly valuable for 3D vision-language (3D-VL) tasks such as 3D visual grounding, dense captioning, and situated reasoning (Ma et al. , 2023).

Research into 3D reasoning with foundational models is at a fascinating juncture. Models like 3D-LLM, PointLLM, and 3D-VisTa have already showcased their effectiveness in diverse 3D tasks, blending geometric, visual, and textual data. Despite these advancements, the field is still burgeoning, with much room for exploration and enhancement. Future directions could include refining model capabilities for more intricate 3D scene interpretations, expanding applications in real-world scenarios such as navigation agents for visually impaired or blind people (Qiu et al. , 2022), and bridging gaps in current methodologies.

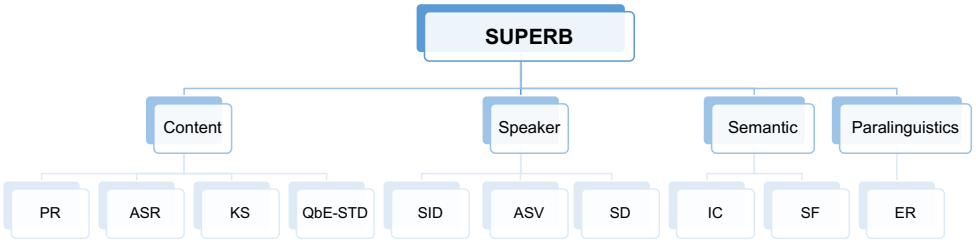


Fig. 10: Four evaluation areas of the SUPERB (Yang et al. , 2021) focus on the discriminative abilities of foundation models and the corresponding tasks. PR: phone recognition, ASR: automatic speech recognition, KS: keyword spotting, QbE-STD: query by example spoken term detection, SID: speaker identification, ASV: automatic speaker verification, SD: speaker diarization, IC: intent classification, SF: slot filling, ER: emotion recognition.

3.6 Audio Reasoning

Audio reasoning pertains to the cognitive mechanism of comprehending, examining, and deriving conclusions from auditory data, of which speech is the major source. Speech representations learned in a self-supervised fashion provide a promising solution in this direction, where a single foundation model is trained and can be applied to a wide spectrum of downstream tasks (Mohamed et al. , 2022).

3.6.1 Speech

The field of speech processing can be broadly classified into two distinct categories: discriminative tasks and generative tasks. Discriminative tasks entail the process of making discrete decisions based on continuous speech, while generative tasks involve the generation of continuous speech from diverse input sources. The Speech processing Universal PERformance Benchmark (SUPERB) (Yang et al. , 2021) is a widely adopted framework for evaluating the discriminative abilities of the foundation model. As demonstrated in Figure 10, it encompasses ten tasks covering four elements of speech: Content, Speaker, Semantics, and Paralinguistics.

The enhanced Speech processing Universal PERformance Benchmark (SUPERB-SG) (Tsai et al. , 2022) further introduces a framework to evaluate the generative abilities of the foundation model with five tasks: speech translation (ST), out-of-domain automatic speech recognition (OOD-ASR), voice conversion (VC), speech separation (SS), and speech enhancement (SE).

The foundation models for learning self-supervised speech representation can be categorized into three major types: 1) **generative models** that reconstruct the input speech sequence leveraging on restricted or corrupted views, for example, vector-quantized variational autoencoder (VQ-VAE) (Van Den Oord et al. , 2017), autoregressive predictive coding (APC) (Chung et al. , 2019), and masked acoustic model (MAM) (Liu et al. , 2020a); 2) **contrastive models** that differentiate a target positive sample from distracting negative samples, for example, contrastive predictive

coding (CPC) (Oord et al. , 2018), Wav2Vec 2.0 (Baevski et al. , 2020), and Speech SIMCLR (Jiang et al. , 2020); and 3) **predictive models** that follow the settings similar to teacher-student learning (Li et al. , 2017), for example, Hidden Unit BERT (HuBERT) (Hsu et al. , 2021), WavLM (Chen et al. , 2022b) and Data2Vec (Baevski et al. , 2022). The Transformer-Encoder (Dong et al. , 2018) architecture and the Conformer-Encoder (Gulati et al. , 2020) architecture are widely adopted in speech foundation models.

3.7 Multimodal Reasoning

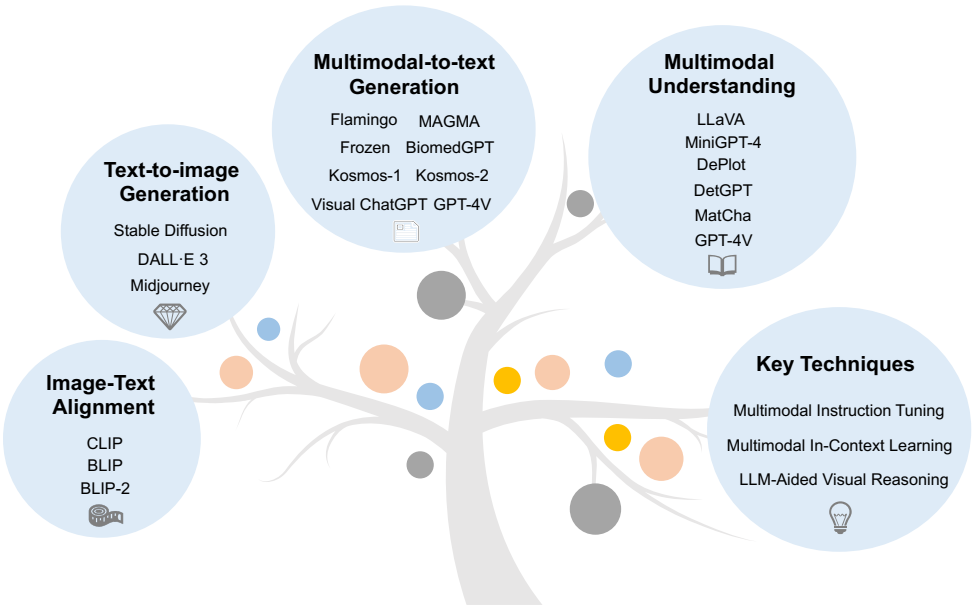


Fig. 11: Multimodal reasoning tasks can be broadly categorized into image-text alignment, text-to-image generation, multimodal-to-text generation, and multimodal understanding. Current multimodal foundation models mainly involves three key techniques to approach reasoning tasks, including multimodal instruction tuning, multimodal in-context learning, and LLM-aided visual reasoning. The figure style credits from tutorial (Li, 2023).

Multimodal reasoning refers to the cognitive process of integrating and reasoning across multiple modalities of information, such as text, images, videos, and other sensory inputs, to enhance understanding and perform complex reasoning tasks (Yin et al. , 2023b; Zong et al. , 2023)[§].

[§] https://github.com/atfortes/Awesome-Multimodal-Reasoning

In the pursuit of developing Artificial General Intelligence (AGI), multimodal reasoning represents a promising advancement over unimodal approaches for several reasons. Firstly, multimodal reasoning aligns more closely with the way humans perceive the world. Humans naturally receive inputs from multiple senses, which often complement and cooperate with each other. As a result, leveraging multimodal information is anticipated to enhance the intelligence of Multimodal Foundation Models. Secondly, multimodal reasoning provides a more user-friendly interface. By incorporating support for multimodal input, users can interact and communicate with intelligent assistants in a more flexible, diverse and potentially more intuitive manner, improving the overall user experience. Thirdly, multimodal reasoning facilitates a more comprehensive problem-solving capability. While unimodal language models typically excel in natural language processing (NLP) tasks, Multimodal Foundation Models have the potential to support a broader spectrum of tasks, making them more versatile and effective as task-solvers. Key techniques and applications of Multimodal Foundation Models encompass various areas, including Multimodal Instruction Tuning (M-IT), which focuses on fine-tuning models based on multimodal instructions; Multimodal In-Context Learning (M-ICL), which leverages contextual information to enhance multimodal reasoning; and LLM-Aided Visual Reasoning (LAVR), which utilizes LLMs to enhance visual reasoning capabilities. Figure 11 shows multiple multimodal reasoning tasks and the key techniques behind, which are introduced as follows.

3.7.1 Alignment

Image-Text Alignment

CLIP (Radford et al. , 2021) utilizes a learning method that enables the creation of cohesive representations for both images and text. By aligning visual and textual information, CLIP fosters cross-modal comprehension and demonstrates exceptional proficiency across a wide range of vision and language tasks. In a similar vein, BLIP-2 (Li et al. , 2023e) adopts a strategy to facilitate efficient cross-modal alignment without fine-tuning the vision encoder. Instead, it introduces a Querying Transformer (Q-Former) that extracts visual features from a fixed image encoder. These extracted query embeddings serve as soft visual prompts for the alignment process. Flamingo (Alayrac et al. , 2022) bridges pretrained vision and language backbones by token fusion with cross-attentions.

3.7.2 Generation

Text-to-image Generation

Stable Diffusion (Rombach et al. , 2021) integrates cross-attention layers to the model architecture, transforming diffusion models into robust and adaptable generative models for diverse conditional inputs like text and bounding boxes. The application of latent diffusion models (LDMs) represents a significant breakthrough in image inpainting, while also delivering impressive results in unconditional content generation, super-resolution image generation, and other tasks. Notably, LDMs offer substantial reductions in computational demands compared to pixel-based diffusion models, while

maintaining highly competitive performance. DALL·E[§] is an advanced AI system that has the capability to generate realistic images and artwork based on natural language descriptions. Likewise, Midjourney is another AI system that specializes in generating images based on natural language descriptions, which are referred to as "prompts". By leveraging the power of AI, Midjourney[§] can translate textual prompts into visual compositions, providing a visual representation of the given description.

Multimodal-to-text Generation

Flamingo-80B (Alayrac et al. , 2022) comprises a family of Visual Language Models (VLMs) equipped with in-context few-shot learning capabilities. These models undergo thorough evaluation across a wide array of tasks, including open-ended ones like visual question-answering and captioning, as well as closed-ended tasks such as multiple-choice visual question-answering. Frozen (Tsimpoukelli et al. , 2021) accomplishes few-shot learning ability within a multimodal context by preserving the language capabilities of a Language Model (LM) while incorporating visual information as a prefix. Frozen achieves this by freezing the LM and training a separate vision encoder to represent images. In the Frozen approach, visual information is represented as a sequence of embeddings, serving as a visual prefix. MAGMA (Eichenberg et al. , 2022) follows a similar approach to Frozen by incorporating a new image prefix encoder while keeping the language model frozen. It trains a series of Visual Language models capable of generating text autoregressively from combined visual and textual inputs. Visual ChatGPT (Wu et al. , 2023a) and GPT-4 (OpenAI, 2023a) represent advancements in extending chatbot capabilities to encompass multimodal applications that support both image and text prompts. Visual ChatGPT builds upon the foundation of ChatGPT and incorporates visual models. It incorporates a Prompt Manager that manages the histories of various visual foundation models, enabling a comprehensive multimodal conversation experience. On the other hand, GPT-4 takes a different approach by accepting prompts that consist of both images and texts. This flexibility empowers users to specify vision and language tasks by generating text outputs in response to arbitrarily interlaced text and image prompts. Microsoft has also proposed a series of Multimodal Foundation Models, including Kosmos-1 (Huang et al. , 2023d) and Kosmos-2 (Peng et al. , 2023d). These models further contribute to the development of multimodal capabilities and facilitate rich interactions involving both images and text. Furthermore, there are ongoing efforts to adapt GPT to specific domains, such as BiomedGPT (Zhang et al. , 2023f), which focuses specifically on biomedical research. These domain-specific adaptations aim to enhance the language model's performance and applicability within specialized fields.

3.7.3 Multimodal Understanding

Visual Instruction Tuning (Liu et al. , 2023e) presents a groundbreaking approach that utilizes GPT-4 to generate multimodal language-image instruction-following data. This approach has the potential to reduce the reliance on manual annotation of large multimodal datasets. Expanding on this foundation, LLaVA (Large Language and

[§] https://openai.com/dall-e-3
[§] https://www.midjourney.com

Vision Assistant) (Liu et al. , 2023e) represents an extensively trained, large-scale multimodal model. It seamlessly integrates a vision encoder with Vicuna (Chiang et al. , 2023), facilitating versatile visual and language comprehension for general-purpose applications. LLaVA excels across a diverse spectrum of tasks necessitating multimodal understanding, encompassing visual question-answering, image captioning, and instruction-following. Notably, it achieves impressive performance on Science QA (Lu et al. , 2022a), a multimodal reasoning dataset in the science domain.

In the domain of reasoning on charts, DePlot (Liu et al. , 2023b) presents a few-shot solution for visual language reasoning. It tackles the challenge through a two-step process: first, translating the plot into text, and then performing reasoning over the translated text. The authors also investigate the combination of DePlot with LLMs to further enhance performance. MatCha (Math reasoning and Chart derendering pretraining) (Liu et al. , 2023c) introduces a comprehensive framework for visual language understanding in the chart domain. It highlights the importance of two critical components: understanding layout, including number extraction and organization, and mathematical reasoning. To enhance visual language understanding, the authors propose two complementary pretraining tasks: chart derendering, which involves generating the underlying data table or code used to create a given plot or chart, and math reasoning.

DetGPT (Pi et al. , 2023) revolutionizes object detection through its reasoning-based approach. It enables the automatic localization of objects of interest based on user-expressed desires, even in cases where the object is not explicitly mentioned. This innovative method incorporates reasoning capabilities to enhance the object detection process. Furthermore, Q-Bench (Wu et al. , 2023c) demonstrates that the multimodal foundation models can perceive low-level visual attributes and provide image quality understanding.

Integrating diverse data types such as text, images, tables, and audio presents distinct challenges for multimodal foundation models compared to their unimodal counterparts. A primary obstacle lies in effectively merging these varied data formats, a task complicated by issues like inconsistency and incompleteness in datasets, where mismatches between image content and corresponding descriptions, or missing data, can adversely affect model performance. Additionally, multimodal foundation models typically demand substantial computational resources for training. Exploring efficient training methods for these models thus emerges as a valuable area of research, crucial for advancing the capabilities of multimodal AI systems. These multimodal foundation models are also instrumental in learning universal representations applicable to fields like materials science, chemistry, and biology (Team, 2022; Manica et al. , 2023).

3.8 Agent Reasoning

Agent reasoning, is an important capability for the Autonomous Language Agents, which refers to a cognitive process that integrates perception, action, and interaction with the physical environment or simulated environment to support reasoning and problem-solving. Autonomous Agents in the context of Large Language Models have

the ability to perform a wide range of tasks, such as task decomposition, generating code, answering questions, engaging in dialogue, providing recommendations, and more. Autonomous Agents, often known as AI Agents, harness the power of Large Language Models to autonomously perform tasks, utilizing their extensive knowledge, reasoning skills, and vast informational resources (Alibali et al. , 2014).

Several works have investigated the use of language for planning purposes (Jansen, 2020; Li et al. , 2022d; Sharma et al. , 2021; Zeng et al. , 2023; Huang et al. , 2022b; Ahn et al. , 2022; Mu et al. , 2023; Hu et al. , 2023a; Zhou et al. , 2023b). Recent methods in task planning utilize pre-trained autoregressive foundation models to break down abstract, high-level instructions into executable, low-level step sequences for an agent, applying a zero-shot approach (Huang et al. , 2022b; Ahn et al. , 2022). Specifically, Huang et al. (2022b) prompt GPT-3 (Brown et al. , 2020) and Codex (Chen et al. , 2021c) to create actions for agents, where each action step is semantically converted into a permissible action through Sentence-RoBERTa (Liu et al. , 2019; Reimers and Gurevych, 2019). In contrast, SayCan (Ahn et al. , 2022) grounds the actions and language by combining the probability of each candidate action, as determined by FLAN (Wei et al. , 2021), with the action's value function. The latter acts as a surrogate for measuring affordance (Shah et al. , 2021). However, both approaches assume the successful execution of each proposed step by the agent, without considering potential intermediate failures in dynamic environments or accounting for the performance of lower-level policies. SwiftSage (Lin et al. , 2023) is a framework influenced by the dual-process theory of human cognition, tailored for superior performance in action planning within intricate interactive reasoning tasks. This framework is structured around two main components: the SWIFT module and the SAGE module. The SWIFT module represents fast and intuitive thinking and is responsible for action planning based on the oracle agent's action trajectories. It is implemented as a small encoder-decoder language model that has been fine-tuned specifically for this purpose. On the other hand, the SAGE module emulates deliberate thought processes and utilizes LLMs such as GPT-4 for subgoal planning and grounding. This module leverages the power of language models to perform more sophisticated reasoning tasks within the framework. Another noteworthy approach in this regard is Reasoning via Planning (RAP) (Hao et al. , 2023a), which capitalizes on the language model's dual role as both a world model and a reasoning agent. RAP incorporates a well-founded planning algorithm, specifically based on Monte Carlo Tree Search, to facilitate strategic exploration within the expansive realm of reasoning. The effectiveness of RAP is evaluated across various tasks, including plan generation, mathematical reasoning (e.g., GSM8K (Cobbe et al. , 2021)), and logical reasoning (e.g., PrOntoQA (Saparov and He, 2023)). The evaluations demonstrate RAP's proficiency in addressing diverse reasoning challenges, effectively showcasing its versatility as a capable reasoning agent.

Introspective Reasoning, Extrospective Reasoning, Embodied Reasoning, and Multiagent Reasoning, along with their interconnected aspects, play pivotal roles in the advancement of agent reasoning systems (Qin et al. , 2023). These components contribute to the development of higher-level cognitive abilities, such as self-awareness, adaptability, and effective collaboration. These capabilities are essential for the creation of intelligent systems that can successfully operate in complex and dynamic

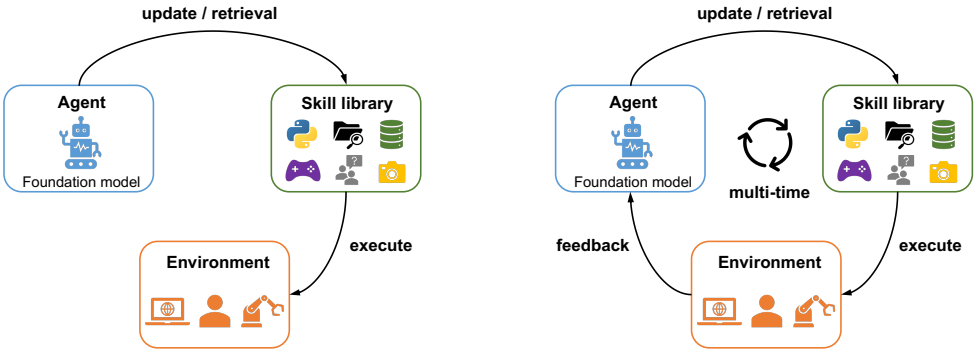


Fig. 12: Difference between introspective reasoning and extrospective reasoning. Introspective reasoning does not require interaction with the environment, while extrospective reasoning leverages observation and feedback from the external environment to adapt plans. The figure style credits from work (Qin et al. , 2023).

environments, seamlessly interact with humans, and engage in cooperative or competitive scenarios with other agents. We believe that combining foundational models with classical methods in robotics may create new opportunities, such as integrating classic approaches to perception, mapping (Pan et al. , 2020), prediction, planning (Mao et al. , 2023b), and control. Safety is a crucial aspect of embodied intelligent systems. In this context, PlanCP (Sun et al. , 2023b) suggests the application of conformal prediction to diffusion dynamic models.

3.8.1 Introspective Reasoning

Introspective reasoning, illustrated in Figure 12(a), relies solely on internal knowledge and reasoning to generate a static plan of tool use without interacting with the environment (Leake, 2012). Several related works in the field of introspective reasoning with LLMs include Program-Aided Language Models (PAL) (Gao et al. , 2023b), ProgPrompt (Singh et al. , 2023), and Code-as-Policies (Liang et al. , 2022a).

PAL (Gao et al. , 2023b) utilizes an LLM for the comprehension of natural language problems and the generation of intermediate reasoning steps in the form of executable programs. Nonetheless, the actual execution of solution steps is delegated to a programmatic runtime, such as a Python interpreter. This approach enables PAL to harness the language understanding capabilities of the LLM while making use of a distinct runtime environment for executing the generated programs. ProgPrompt (Singh et al. , 2023) presents a structured LLM prompt akin to programming, crafted to facilitate the generation of plans in diverse situational settings, encompassing different robot functionalities and tasks. This structure involves prompting the LLM with program-style descriptions of accessible actions and objects in a given environment, along with sample programs for execution. Code-as-Policies (Liang et al. , 2022a) introduces a robot-oriented framework for Language Model Generated Programs (LMPs). These LMPs are capable of depicting both reactive policies, like impedance controllers,

and waypoint-oriented strategies. The versatility of Code-as-Policies is demonstrated across multiple real robot platforms, showcasing its applicability in diverse robotic scenarios.

Introspective reasoning may have limitations in dynamic and uncertain environments where external feedback and interaction with the environment are crucial for effective planning. It may struggle to adapt plans to changing circumstances or handle unexpected events without external information.

3.8.2 Extrospective Reasoning

Introspective reasoning, despite its simplicity, lacks the ability to adjust or modify a plan based on intermediate execution results. In contrast, extrospective reasoning operates by generating plans incrementally. As shown in Figure 12(b), it accomplishes this by iteratively interacting with the environment and incorporating feedback obtained from previous executions. Extrospective reasoning actively incorporates external information gathered through interactions with the environment. This allows extrospective reasoning to adapt and refine its plans based on real-time feedback and the observed outcomes of previous actions (Acay et al. , 2007).

By actively engaging with the environment and utilizing feedback, extrospective reasoning offers a more flexible and responsive approach to generating plans, which is particularly suitable for complex and dynamic situations where the ability to adapt and learn from experience is crucial. Several related works in the field of extrospective reasoning with LLMs include Self-Ask (Press et al. , 2023), ReAct (Yao et al. , 2023c), ToolFormer (Schick et al. , 2023), and LLM-Planner (Song et al. , 2023a). Self-Ask (Press et al. , 2023) proactively generates and responds to its own follow-up queries before addressing the original question. Meanwhile, ReAct (Yao et al. , 2023c) leverages large language models to concurrently produce reasoning traces and task-specific actions. This dual approach enhances the interaction between these elements, with reasoning traces aiding in the development, monitoring, and modification of action plans, as well as managing unexpected situations. Conversely, actions facilitate the model's engagement with and acquisition of supplementary data from external entities like knowledge bases or environments. ToolFormer (Schick et al. , 2023) is designed to intelligently determine the appropriate APIs to utilize, the timing for their invocation, the specific arguments to provide, and how to effectively integrate the obtained results into subsequent token predictions. LLM-Planner (Song et al. , 2023a) utilizes the capabilities of large language models for efficient few-shot planning in the context of embodied agents.

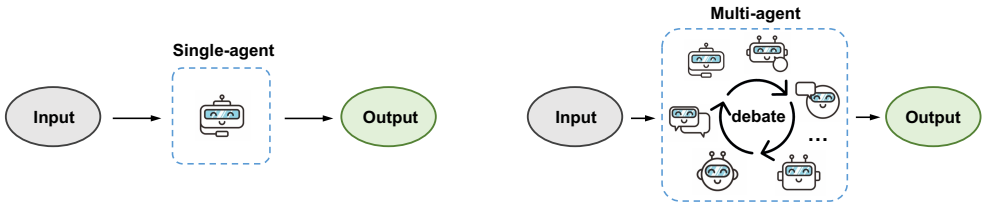
In addition to the above-mentioned research, Statler (Yoneda et al. , 2023) provides a framework equipping LLMs with a persistent, memory-like representation of the world state. It utilizes two forms of general LLMs: a world-model reader and a world-model writer, both of which interact with and update the world state. This addition of a memory-like element to the framework significantly boosts the reasoning abilities of LLMs, allowing them to process information over extended time periods, free from the constraints typically imposed by context length limitations. The explicit representation of the world state empowers LLMs to retain and access relevant information, facilitating more comprehensive and contextually aware reasoning processes. Dasgupta

et al. (2022) propose a collaborative system that combines the complementary reasoning abilities of LLMs. The system has three components: the Planner, the Actor, and the Reporter. The Planner is a pre-trained language model responsible for generating commands that guide the actions of a simple embodied agent, referred to as the Actor. The Reporter acts as a communication bridge between the Planner and the Actor, relaying relevant information to the Planner to inform its decision-making process for issuing subsequent commands. By harnessing the strengths of each component, this collaborative system aims to enhance the overall reasoning and decision-making capabilities of LLMs, allowing for more effective and context-aware interactions between language-based instructions and the embodied agent. Inner Monologue (Huang et al. , 2022c) investigates the capacity of LLMs to reason effectively in embodied contexts by leveraging natural language feedback without additional training. The authors propose that by incorporating environmental feedback, LLMs can develop an inner monologue that augments their capability to process and plan within robotic control scenarios. This development enables LLMs to gain a more comprehensive understanding of the environment and enhances their adaptability to dynamic circumstances.

The iterative nature of extrospective reasoning enables it to dynamically adjust its plan based on the evolving state of the environment and the outcomes of executed actions. This adaptive process enhances the effectiveness and efficiency of planning, as it leverages the knowledge gained from experience to continually improve future decision-making.

3.8.3 Embodied Reasoning

Recent research has highlighted the successful application of LLMs in robotics domains (Ahn et al. , 2022; Zeng et al. , 2023; Huang et al. , 2022c; Liang et al. , 2022a; Ding et al. , 2023c). Moreover, planning can be considered a form of temporal reasoning, adding to the significance of integrating LLMs into robotics. Gato (Reed et al. , 2022) functions as a multimodal, multi-task, and multi-embodiment generalist policy. It leverages supervised learning with an impressive parameter count of 1.2 billion. This technology has been acknowledged as a form of "general-purpose" artificial intelligence, representing a significant advancement towards the realization of artificial general intelligence. Robotic Transformer 1 (RT-1) (Brohan et al. , 2022) is trained on a comprehensive real-world robotics dataset consisting of over 130,000 episodes that encompass more than 700 tasks. This extensive dataset was collected over a period of 17 months using a fleet of 13 robots from Everyday Robots. RT-1 demonstrates promising properties as a scalable, pre-trained model, showcasing its ability to generalize based on factors such as data size, model size, and data diversity. The utilization of large-scale data collected from real robots engaged in real-world tasks contributes to RT-1's robustness and its potential for generalization in practical scenarios. Expanding upon the capabilities of RT-1, Robotic Transformer 2 (RT-2) (Brohan et al. , 2023) further enhances the model's understanding of the world, resulting in more efficient and accurate execution of robotic tasks. By incorporating the chain of thought reasoning, RT-2 achieves multi-stage semantic reasoning abilities. This expansion equips RT-2 with a set of emerging capabilities derived from extensive training on a vast internet-scale dataset. Prominent advancements encompass a marked improvement in



(a) Single-agent Reasoning **(b) Multi-agent Reasoning**

Fig. 13: Difference between single-agent reasoning and multi-agent reasoning.

the model's ability to generalize to unfamiliar objects, the capacity to understand commands absent from its original training data, and the capability to engage in basic reasoning when responding to user instructions. These enhancements enhance RT-2's performance and broaden its capacity to tackle a more extensive array of tasks with increased sophistication. After that, RT-X (Padalkar et al. , 2023) further extends RT-1 and RT-2 to cross-embodiment settings and shows better transferabilities and zero-shot capabilities. RoboFlamingo (Li et al. , 2023k) leverages pre-trained Vision-Language Models (VLMs) to achieve sophisticated single-step vision-language comprehension. It incorporates an explicit policy head to effectively capture sequential historical data. This design grants it the flexibility needed for implementing open-loop control strategies and is finely tuned for efficient deployment on resource-constrained platforms.

Embodied reasoning plays a vital role in the development of intelligent robots. As humans, we are educated to comprehend the world by employing numerical/physical laws and logical principles. The question arises: can we empower robots with the same capacity? Numerous everyday tasks necessitate simple reasoning based on visual perception and natural language understanding. If we aspire to have robot companions capable of collaborating with us, it is essential for them to possess the ability to understand and reason over both visual information and natural language input. The ultimate objective of creating smart robots is to enable them to act in a manner that is comparable to, or even surpasses, human capabilities (Xu et al. , 2021b). This entails embodying human-like reasoning and performance in robots, aiming to bridge the gap between humans and machines. By enabling robots to understand and reason over visual and linguistic inputs, we move closer to achieving the goal of developing robots that can effectively interact and collaborate with humans.

3.8.4 Multi-agent Reasoning

Multi-agent reasoning refers to the cognitive process by which multiple autonomous agents or entities engage in reasoning, decision-making, and communication within a shared environment or context. Compared with reasoning with a single agent, it involves the ability of individual agents to perceive, interpret, and reason about the actions, goals, beliefs, and intentions of other agents, and to adjust their own behaviors accordingly. Their differences are briefly summarized in Figure 13.

Recent studies have introduced the concept of multi-agent debate as a promising method to elevate reasoning abilities and ensure factual accuracy across diverse scenarios. In the work by Zhang et al. (2023c), they introduce a framework that leverages the capabilities of Large Language Models (LLMs) to foster cooperative interactions among multiple agents within embodied environments. This innovative approach empowers embodied agents to efficiently strategize, communicate, and collaborate with both other agents and humans, thereby enhancing their proficiency in accomplishing intricate, long-term tasks. In a similar vein, Du et al. (2023) propose a methodology that involves multiple instances of language models engaging in debates. Through iterative rounds of reasoning and response generation, these models collectively work towards reaching a common final answer. This approach has demonstrated significant improvements in mathematical and strategic reasoning across various tasks.

In contrast to the aforementioned studies, Nascimento et al. (2023) propose the integration of LLMs, such as GPT-based technologies, into multi-agent systems (MASs). They introduce the concept of incorporating LLMs into MASs to create self-adjusting agents. This integration is achieved through an LLM-based MAPE-K (Monitoring, Analyzing, Planning, Executing, and Knowledge) model (do Nascimento and de Lucena, 2017; Redbooks, 2004), which enables the agents to adapt and adjust their behaviors based on the knowledge and insights gained from LLMs.

Federated Learning (FL) has gained prominence as a technology enabling the collaborative development of communal models while safeguarding data that remains decentralized. Chen et al. (2023a) introduce the idea of a federated LLM, encompassing three crucial elements: pre-training of federated LLMs, fine-tuning of these models, and the engineering of prompts specific to federated LLMs. This approach harnesses the potential of federated learning to enhance multi-agent reasoning by leveraging LLMs.

These research efforts demonstrate the efficacy of multi-agent debate approaches in enhancing reasoning abilities and factual accuracy. By leveraging the power of large language models and enabling cooperative interactions between agents, these studies contribute to the advancement of AI systems capable of complex reasoning and improved performance across various domains.

3.8.5 Reasoning in Autonomous Driving

Reasoning within the domain of autonomous driving spans across perception (Li et al. , 2023d,j; Sun et al. , 2022b, 2023c), safety (Zhou et al. , 2023d), explainability (Echterhoff et al. , 2023; Sha et al. , 2023; Sun et al. , 2021; Huang et al. , 2021b) and system level (Chen et al. , 2023d). Chen et al. (2023d) propose the frontiers and challenges for end-to-end autonomous driving, where logical reasoning with LLMs could have substantial impacts on different driving scenarios. Zhou et al. (2023d) review some recent work on LLMs with regards to driving. It suggests that by integrating language data, vehicles as well as transportation systems can carry out reasoning and interact with real-world environments with a higher level of intelligence.

We believe that the common sense and world knowledge inherited from foundation models could unleash the substantial effectiveness of algorithms onboard to handle

corner cases and enhance explainability and safety. Below, we survey this emergent topic from two perspectives.

DriveGPT4 (Xu et al. , 2023d) represents a groundbreaking endeavor that harnesses LLMs to comprehend an interpretable end-to-end autonomous driving system. This pioneering effort not only showcases remarkable qualitative but also quantitative achievements when benchmarked against challenging standards. GPT-Driver (Mao et al. , 2023a) and Agent Driver (Mao et al. , 2023b) introduce the approach by utilizing LLMs as cognitive agents to operate a tool library. This implementation enhances driving behavior by incorporating explainability into the decision-making process. MotionLM (Seff et al. , 2023) cast multi-agent motion prediction as a language modeling task. The continuous trajectories are represented as sequences of discrete motion tokens. Among the many other attempts, one particular challenge is how to utilize logical reasoning (e.g., chain of thought) to rationalize and explain driving behaviors. Echterhoff et al. (2023) propose a new view using the concept bottlenecks for control command predictions. Tan et al. (2023) turn to language as a source of supervision to obtain dynamic traffic scenarios, surpassing prior work in terms of realism and fidelity. nuPrompt (Wu et al. , 2023b) is the first object-centric language prompt set for 3D, multi-view, and multi-frame driving scenes. It is equipped with diverse pairs of instance-prompt data and validated in the object tracking task.

3.9 Other Tasks and Applications

3.9.1 Theory of Mind (ToM)

The development of Theory of Mind (ToM)-like ability in models is speculated to have occurred naturally and independently as a consequence of their advancing language skills (Kosinski, 2023). Another explanation suggests that models were able to solve ToM tasks by uncovering and utilizing undiscovered language patterns, rather than explicitly employing ToM. While this alternative explanation may seem ordinary, it is actually remarkable as it implies the existence of undisclosed language regularities that enable the resolution of ToM tasks without the direct engagement of ToM.

3.9.2 Weather Forecasting

Weather forecasting plays a crucial role in both scientific research and societal applications. As an application of scientific reasoning, weather forecasting involves the use of reasoning skills to analyze data, identify patterns, and make predictions about future weather conditions.

MetNet-2 (Espeholt et al. , 2022) is a neural network specifically designed for high-resolution precipitation forecasting with up to a 12-hour lead time. This model excels in accurately predicting raw precipitation targets and outperforms state-of-the-art physics-based models currently used in the Continental United States. In another study, Bi et al. (2023) present Pangu-Weather, an AI-based approach designed to achieve accurate global weather forecasts in the medium range. This method utilizes 3D deep networks that incorporate earth-specific priors, allowing for the effective

handling of complex weather data patterns. To mitigate accumulation errors encountered in medium-range forecasting, a hierarchical temporal aggregation strategy is employed. By undergoing training on an extensive dataset spanning 39 years of global weather information, Pangu-Weather exhibits exceptional deterministic forecasting performance across all assessed variables when compared to the operational integrated forecasting system of the European Centre for Medium-Range Weather Forecasts (ECMWF). This underscores the remarkable effectiveness of Pangu-Weather in delivering precise global weather forecasts, offering valuable insights and advantages for a multitude of applications that heavily depend on weather-related information.

3.9.3 Medical Reasoning

Reasoning is also common in medicine. For example, clinicians reason the potential causes of a patient's symptoms and then advise which examinations to take and what treatment is the best following the diagnosis (Qiu et al. , 2023a).

With a wide medical knowledge spectrum, foundation models can conduct expert-level reasoning in the context of medicine. For example, Med PaLM 2 (Singhal et al. , 2023), a biomedical large language model (LLM), scored 86.5% in answering medical questions on the MedQA benchmark; GPT-4 passed the US Medical Licensing Exam (USMLE) with a score of 86.7%. Breakthroughs in medical reasoning brought by LLMs also inspire reasoning carried out in other medical modalities, such as medical images. For example, VisionFM (Qiu et al. , 2023b), a foundation model for ophthalmic image analysis, demonstrates impressive reasoning skills in predicting the presence of intracranial tumors from fundus photographs, surpassing both intermediate- and senior-level clinicians. RETFound (Zhou et al. , 2023e) shows remarkable performance in reasoning systemic diseases from ocular images. LLaVA-Med (Li et al. , 2023a) adapts LLaVA (Liu et al. , 2023e) to align biomedical vocabulary and learn open-ended conversational semantics, which enables the interpretation of biomedical images and achieves promising performance for biomedical visual question answering. ELIXR (Xu et al. , 2023c) incorporates a language-aligned image encoder to perform a range of vision-language reasoning tasks for chest X-ray images. Tu et al. (2023b) develop a multimodal biomedical foundation model, Med-PaLM M, to simultaneously explore clinical language, imaging, and genomics data, as well as introduce a multimodal biomedical benchmark, MultiMedBench. Given the multimodal nature of medicine, it is anticipated that medical reasoning will be further augmented by increasingly intelligent multimodal foundation models (Yang et al. , 2023f).

However, unlike in other domains, reasoning in medicine has to take more caution (Yan et al. , 2023). Rigorous verification and examination should be conducted to ensure the biomedical reasoning outcome is factually grounded, and regulations should be established and enforced to provide legitimate and safe use of foundation models for biomedical reasoning.

3.9.4 Bioinformatics Reasoning

Reasoning in bioinformatics involves analyzing and interpreting complex languages of biology and gaining insights into the processes related to life. This includes understanding genetic sequences, protein functions, and cellular mechanisms through the analysis of large-scale datasets. Foundation models are reshaping various perspectives for biological reasoning, such as predicting protein structures and designing sequences in drug discovery (Savage, 2023).

In the field of biotechnology, numerous studies highlight the efficacy of foundation models in reasoning and analyzing DNA (Nguyen et al. , 2023), RNA (Wang et al. , 2023l), and protein (Jumper et al. , 2021). A notable example is AlphaFold (Jumper et al. , 2021), which employs a transformer network architecture to precisely predict protein structures. ProGen (Madani et al. , 2023) and its subsequent ProGen2 (Nijkamp et al. , 2022b) develop a suite of large protein language models, akin to natural language models, for generating protein sequences. RFdifusion (Watson et al. , 2023) adopts a denoising diffusion approach in protein structure design, demonstrating significant advancements across various protein design tasks. In the context of protein-ligand interactions, Li et al. (2023p) train the GPT-2 model on protein-ligand binding data, utilizing language model capabilities for ligand design. Prot2Text (Abdine et al. , 2023) combines graph neural networks with LLMs to predict protein functions in a free-text format. Chen et al. (2023c) introduce a framework powered by LLMs for condition recommendation in chemical synthesis, which aids drug discovery. This framework is designed to search the most recent chemical literature, utilizing in-context learning capabilities and employing multi-LLM debate strategies to enhance effectiveness. For RNA analysis, Uni-RNA (Wang et al. , 2023l) exhibits exceptional performance in structural and functional predictions, including RNA high-order structure map prediction, by leveraging large-scale pre-training on extensive RNA sequences. Additionally, HyenaDNA (Nguyen et al. , 2023) utilizes the long-range modeling and in-context learning strengths of LLMs and is pre-trained on human reference genome data, yielding significant achievements in genomic tasks. GeneGPT (Jin et al. , 2023a) enhances LLMs by integrating the National Center for Biotechnology Information (NCBI) API, which improves answering questions related to genomics.

3.9.5 Code Generation

Code generation, also referred to as program synthesis or generating code from a natural language description (NL2Code) (Zan et al. , 2023), is the process or technology that converts inputs in natural language into computer code. NL2Code represents a significant step towards more intuitive and accessible programming, leveraging foundation model to bridge the gap between natural language and computer code.

PyMT5, delineated in the work of Clement et al. (2020), stands as a Python-based text-to-text transfer transformer, adept at translating between diverse combinations of Python method features. This singular model is capable of generating entire methods from natural language documentation strings and summarizing code into various common docstring styles. Similarly, IntelliCode Compose (Svyatkovskiy et al. , 2020)

is a versatile multilingual code completion tool, proficient in predicting code token sequences and generating syntactically correct code lines. GPT-Neo (Black et al. , 2021) exemplifies an implementation of GPT-2 and GPT-3-like models, with a focus on distributed support through Mesh Tensorflow. This approach is further extended in GPT-J and GPT-NeoX-20B, as detailed in Wang and Komatsuzaki (2021) and Black et al. (2022) respectively. PLBART (Ahmad et al. , 2021) is a model pre-trained on an extensive corpus of Java and Python functions, coupled with natural language text, utilizing a denoising autoencoding approach. CodeT5 (Wang et al. , 2021b) distinguishes itself as a unified pre-trained encoder-decoder Transformer, enhancing the semantic understanding of developer-assigned identifiers. LaMDA (Thoppilan et al. , 2022) emerges as a dialog-specialized family of Transformer-based models, pre-trained on a substantial volume of dialog data and web text.

CodeParrot (Tunstall et al. , 2022) is a GPT-2 based model trained for Python code generation, while Codex (Chen et al. , 2021c) showcases a GPT language model fine-tuned on a vast array of public code from GitHub. Chandel et al. (2022) delve into the practicality of a Data Science assistant empowered by a transformer model, JuPyT5, trained on public Jupyter Notebook repositories, and introduce a new evaluation metric, DSP. PolyCode (Xu et al. , 2022) is a GPT-2 based model with substantial coding proficiency across multiple languages, trained on a large code dataset. AlphaCode (Li et al. , 2022f) stands out as a code generation system, demonstrating notable performance in programming competitions. CodeRL (Le et al. , 2022) merges pre-trained language models with reinforcement learning for program synthesis. ERNIE-Code (Chai et al. , 2022) employs unique pre-training methods, focusing on both monolingual and cross-lingual learning. Pangu-Coder (Christopoulou et al. , 2022) adopts a two-stage training strategy, initially focusing on raw programming language data and subsequently on text-to-code generation. FIM (Bavarian et al. , 2022) demonstrates the efficacy of autoregressive language models in text infilling. Zan et al. (2022) introduce CERT, a model comprising a sketcher and generator for detailed code creation, trained on unlabelled data. InCoder (Fried et al. , 2022) focuses on code file generation from a large, permissively licensed code corpus, enabling code infilling with bidirectional context. Nijkamp et al. (2022a) present CodeGen, a family of large language models for both natural language and programming, accompanied by the JAXFORMER training library.

CodeGeeX (Zheng et al. , 2023c) is a multilingual model for code generation, trained on a vast dataset of programming languages. SantaCoder (Allal et al. , 2023) is a model with 1.1 billion parameters, trained on Java, JavaScript, and Python subsets from The Stack (Kocetkov et al. , 2022), and assessed using the MultiPL-E text-to-code benchmark. This research revealed that intensifying the filtering of near-duplicates enhances performance, and interestingly, choosing files from repositories with more than five GitHub stars tends to reduce performance significantly. In contrast, StarCoder (Li et al. , 2023i) is a more robust model with 15.5 billion parameters and an 8K context length. It boasts infilling capabilities and rapid large-batch inference, enabled by multi-query attention, and is trained on a vast dataset of one trillion tokens from The Stack (Kocetkov et al. , 2022). WizardCoder (Luo et al. , 2023f) enhances Code LLMs with intricate instruction fine-tuning, adapting the Evol-Instruct method for

the code domain. AceCoder (Li et al. , 2023g) incorporates two innovative solutions to address coding challenges: firstly, it employs guided code generation, prompting LLMs to initially analyze requirements and produce preliminary outputs like test cases; secondly, it features example retrieval, selecting similar programs as prompt examples to provide relevant content such as algorithms and APIs. CodeGen2 (Nijkamp et al. , 2023) aims to make the training of LLMs for program synthesis more efficient by integrating four essential elements: model architectures, learning methods, infill sampling, and data distributions. CodeT5+ (Wang et al. , 2023o) forms a family of encoder-decoder LLMs for code, characterized by flexible module combinations to address a broad spectrum of downstream code tasks. CodeTF (Bui et al. , 2023) is an open-source Transformer-based library dedicated to cutting-edge Code LLMs and code intelligence applications. Code Llama (Roziere et al. , 2023) represents a family of large language models for code, based on Llama 2, and offers top-tier performance among open models, along with infilling capabilities, support for large input contexts, and the ability to follow instructions in a zero-shot manner for programming tasks. CodeFuse (Di et al. , 2023) is tailor-made for code-related tasks and is unique in its support for both English and Chinese prompts, accommodating over 40 programming languages.

3.9.6 Long-Chain Reasoning

Long-chain reasoning refers to the ability to connect and reason about a series of multiple, often complex pieces of information or events in a long sequential, and extended manner. Long-chain reasoning is often required in complex problem-solving, decision-making, and understanding of intricate systems.

Ho et al. (2022) introduce Fine-tune-CoT, a method that leverages very large teacher models to generate reasoning samples for fine-tuning smaller models. By employing Fine-tune-CoT, smaller models acquire significant reasoning capabilities, surpassing prompt-based baselines and even outperforming the teacher model in numerous tasks.

Before the emergence of foundational models, the reasoning capabilities of earlier models were notably limited (Sun et al. , 2022a). This limitation primarily stemmed from the tendency of learning-based models to rapidly forget previous information. Long-chain reasoning has great potential for application in AI Agent Reasoning or Embodied Reasoning, enabling them to handle more intricate and nuanced tasks. Despite the emergence of foundational models like GPT-4, mastering long-chain reasoning continues to be a significant challenge. We emphasize the immense utility of long-chain reasoning in applications such as decision-making, planning, and question-answering. With this in mind, we aim to draw attention to this area, encouraging researchers in foundational models to further investigate and advance in this field.

3.9.7 Abstract Reasoning

Abstract reasoning refers to the cognitive ability to analyze and manipulate abstract concepts, ideas, or symbols without relying on specific contexts or concrete examples. It involves transcending immediate sensory input and specific instances to identify underlying patterns, relationships, and fundamental principles. Abstract reasoning requires the identification and application of general patterns based on limited data.

Gendron et al. (2023) extensively evaluate state-of-the-art LLMs in abstract reasoning tasks. Their research reveals that these models demonstrate notably limited performance compared to their performance on other natural language tasks. The findings suggest that LLMs face challenges when it comes to effectively tackling abstract reasoning, highlighting the need for further advancements in this area.

3.9.8 Defeasible Reasoning

Defeasible reasoning refers to a mode of reasoning in which conclusions can be overturned or revised based on new evidence or information (Madaan et al. , 2021). CURIOUS (Madaan et al. , 2021) is a framework that supports defeasible reasoning for humans, utilizing an inference graph (Pollock, 2009). In the context of defeasible inference, Rudinger et al. (2020) have provided three noteworthy datasets: δ-ATOMIC, δ-SNLI, and δ-SOCIAL. These datasets exhibit diversity by covering different domains, offering unique challenges for studying defeasible inference. δ-ATOMIC pertains to commonsense reasoning, presenting scenarios that require drawing defeasible inferences based on background knowledge and understanding of everyday situations. δ-SNLI focuses on natural language inference, requiring reasoning about the relationships between premises and hypotheses. δ-SOCIAL involves reasoning about social norms and conventions, providing a platform for investigating the application of defeasible reasoning in understanding and interpreting social behavior. Zhou et al. (2020) introduce a testbed aimed at evaluating models' abilities to simulate human cognitive processes such as knowledge abstraction, concretization, and completion (KACC). These cognitive abilities play a crucial role in understanding the world and effectively managing acquired knowledge. The testbed includes new datasets characterized by larger concept graphs, ample cross-view links, and dense entity graphs, providing a more comprehensive representation of knowledge. Within this experimental framework, the authors introduce innovative challenges, specifically multi-hop knowledge abstraction (MKA) and multi-hop knowledge concretization (MKC). These tasks necessitate intricate reasoning capabilities from models, involving the abstraction or concretization of knowledge across multiple sequential steps. Kazemi et al. (2023) frame the problem of *reasoning with contradictory information*, guided by source preferences, as a classical problem of *defeasible reasoning*. This formulation allows for a comprehensive exploration of models' abilities to handle conflicting information and prioritize different sources in the reasoning process. BoardgameQA (Kazemi et al. , 2023) is a dataset designed to assess the defeasible reasoning capabilities of models. The dataset consists of 1000 training examples, 500 validation examples, and 1000 testing examples for each variation.

Each of these datasets presents distinct challenges and opportunities for studying and advancing defeasible inference within various domains. Researchers can leverage these datasets to explore the capabilities and limitations of defeasible reasoning models in different contexts, contributing to the development of robust and adaptable reasoning systems with foundation model technologies.

Dataset	Choices	Knowledge Types	Questions
Swag Zellers et al. (2018)	4	Temporal, Physical	113,000
PHYRE Bakhtin et al. (2019)	/	Physical	25
HellaSwag Zellers et al. (2019)	4	Temporal, Physical	70,000
WinoGrande Sakaguchi et al. (2021)	2	Social, Physical	44,000
Social IQA Sap et al. (2019)	3	Social	35,350
PIQA Bisk et al. (2020)	2	Physical	21,020
SummEdits Laban et al. (2023)	2	Social	6,348
CConS Kondo et al. (2023)	/	Physical	1,112

Table 5: Commonsense Reasoning Benchmark Statistics. Choices: the number of choices for each question; Questions: the number of questions.

3.10 Benchmarks, Datasets, and Metrics

Benchmarks, datasets, and metrics play a crucial role in evaluating and advancing reasoning capabilities in various domains, driving innovation, and fostering the development of more capable and reliable reasoning systems. These resources provide standardized frameworks and tasks that enable researchers and developers to objectively assess the performance of reasoning models and compare different approaches. Representative datasets are summarized in Table 8 and 9.

3.10.1 Commensense Reasoning

In addition to CQA Talmor et al. (2019) and CoS-E Rajani et al. (2019), there are several other benchmarks available for evaluating commonsense reasoning (Table 5): PHYRE (PHYsical REasoning) benchmark Bakhtin et al. (2019) consists of 25 task templates that focus on physical reasoning. CConS (Counter-commonsense Contextual Size comparison) dataset Kondo et al. (2023) investigates the impact of physical commonsense on the contextualized size comparison task. It includes both contexts that align with physical commonsense and those that deviate from it. The dataset comprises 139 templates and automatically generates 1,112 examples. SummEdits Laban et al. (2023) is a benchmark spanning 10 domains. It is designed to be more cost-effective per sample compared to previous benchmarks, offering a 20-fold improvement in efficiency. The benchmark is highly reproducible and aims to evaluate the performance of Language Model-based Systems (LLMs) on complex tasks, addressing issues with existing evaluation benchmarks.

Furthermore, commonsense knowledge encompasses various categories, including physical commonsense, social commonsense, and temporal commonsense. Benchmarks in this domain generally fall into two tasks: multiple-choice evaluation and generative evaluation. Multiple-choice benchmarks, such as SWAG Zellers et al. (2018), HellaSWAG Zellers et al. (2019), Social IQA Sap et al. (2019), and PIQA Bisk et al. (2020), require models to select the correct answer from a set of options. Generative evaluation Lin et al. (2020a), as seen in benchmarks like ProtoQA Boratko et al. (2020) and CommonGen Lin et al. (2020b), involves generating answers based on provided questions and context. Rainbow Lourie et al. (2021) is a universal commonsense

reasoning benchmark that integrates six existing tasks: 1) αNLI Bhagavatula et al. (2019); 2) Cosmos QA Huang et al. (2019); 3) HellaSWAG Zellers et al. (2019); 4) PIQA Bisk et al. (2020); 5) Social IQA Sap et al. (2019); and 6) WinoGrande Sakaguchi et al. (2021). It covers both social and physical commonsense reasoning and provides a comprehensive evaluation platform.

Metrics

In multiple-choice benchmarks, accuracy is the primary metric used to evaluate a model's ability to select the correct answer. However, in language generation evaluations, automated metrics like BLEU Papineni et al. (2002) may not always align perfectly with human judgment, so they should be used with caution.

In the case of the PHYRE benchmark Bakhtin et al. (2019), a metric measuring performance called AUCCESS is computed. AUCCESS aggregates the success percentages across different attempts by using a weighted average. The formula for AUCCESS is AUCCESS $= \sum_k w_k \cdot s_k / \sum_k w_k$. Here, w_k represents weights that place more emphasis on tasks with fewer attempts, and it is calculated as $w_k = \log(k+1) - \log(k)$. The variable s_k denotes the success percentage at the k-th attempt. AUCCESS takes into account the performance across multiple attempts and provides a more comprehensive evaluation that rewards models for solving tasks with fewer attempts.

3.10.2 Mathematical Reasoning

Math Word Problems (MWPs)

There have been several benchmark datasets introduced for math word problem-solving. One such dataset is Alg514 Kushman et al. (2014), which is also used by Zhou et al. (2015) for evaluation. Alg514 consists of 514 algebra word problems sourced from online platforms. Each problem in the dataset is annotated with linear equations, and the template of each problem must appear at least six times within the entire set. Another dataset, Verb395 Hosseini et al. (2014) is a collection of addition and subtraction problems. The DRAW dataset Upadhyay and Chang (2015) features 1,000 algebraic word problems, each accompanied by linear equation annotations, collected from `algebra.com`. Meanwhile, SingleEQ Koncel-Kedziorski et al. (2015) is comprised of 508 problems, with each problem corresponding to a single equation. MaWPS Koncel-Kedziorski et al. (2016) repository provides interfaces for adding new word problems, which allows for the further extension of the dataset. These benchmark datasets cover various levels of difficulty and are useful for evaluating math word problem-solving approaches. Dolphin18K Huang et al. (2016) consists of over 18,000 annotated math word problems in the field of elementary mathematics. The dataset includes both the unedited text of the problem and either a single or multiple pieces of response text supplied by the individuals who answered the problems. MATH Hendrycks et al. (2021b) comprises 12,500 challenging competition mathematics problems. Each problem in this dataset is accompanied by a full step-by-step solution. This rich annotated information allows models to be trained to generate detailed answer derivations and explanations. TabMWP Lu et al. (2022b) features a collection of 38,431 grade-level, open-domain problems that necessitate mathematical reasoning through both text and tables. This dataset is divided into training, dev, and

testing subsets, following a 6:2:2 distribution. In TabMWP, every query is associated with a tabular context displayed as an image, semi-structured text, and a structured table. The average length of these questions is 22.1 words, with the solutions averaging 49.5 words. The problems in TabMWP can be of two types: free-text questions and multiple-choice questions. Every problem comes with annotated gold-standard solutions that illustrate the multi-step reasoning involved.

GSM8K Cobbe et al. (2021) is a math word problem dataset that consists of 8.5K grade school math word problems. These problems exhibit varying levels of linguistic complexity and difficulty. With problem lengths ranging from 2 to 8 steps, they require a diverse set of mathematical skills and strategies to solve effectively. Multilingual Grade School Math (MGSM) benchmark Shi et al. (2023) consists of 250 grade-school math problems that have been manually translated from the GSM8K dataset Cobbe et al. (2021) into ten languages with diverse linguistic typologies. The MGSM benchmark serves as an evaluation tool to assess the reasoning abilities of language models across multiple languages. It helps to identify areas where models may face challenges, such as cross-lingual reasoning and handling linguistic variations between languages. By incorporating typologically diverse languages, the benchmark ensures its relevance and applicability to real-world multilingual scenarios.

There are two Chinese datasets, Math23K Wang et al. (2017) and HMWP Qin et al. (2020), specifically designed for math word problems at the elementary school level. Math23K Wang et al. (2017) consists of 23,161 problems that are annotated with structured equations and corresponding answers. The Hybrid Math Word Problems dataset (HMWP) Qin et al. (2020) includes three types of math word problems extracted from a Chinese K12 math word problem bank. The dataset comprises 5,491 math word problems, categorized as follows: 2,955 one-unknown-variable linear MWPs, 1,636 two-unknown-variable linear MWPs, and 900 one-unknown-variable non-linear MWPs. Additionally, there is the DRAW1K dataset Upadhyay and Chang (2017), which contains 1,000 general algebra word problems. This dataset includes human-annotated derivations, which serve as information structures for problem-solving. The authors have also provided derivation annotations for over 2,300 algebraic word problems to facilitate future evaluations. They suggest evaluating solvers based on "derivation accuracy". Math23K-F and MAWPS-F Liu et al. (2023f) are datasets that provide high-quality, precise annotations of formula usage in each reasoning step for Math Word Problems. These datasets aim to enhance the understanding of how formulas are utilized throughout the problem-solving process. In conjunction with these datasets, the authors propose the Formulamastered Solver (FOMAS) system Liu et al. (2023f), which incorporates insights from the dual process theory and consists of two components: the Knowledge System and the Reasoning System. The former is responsible for learning and acquiring formula knowledge, while the latter leverages this knowledge to solve math word problems. This dual-component architecture enables FOMAS to leverage formula knowledge in the reasoning process effectively. The Academia Sinica Diverse MWP Dataset (ASDiv) Miao et al. (2020) consists of 2,305 English math word problems (MWPs). The Math Word Problems (MWPs) in this dataset display a diverse range of textual patterns, encompassing the majority of problem types that are typically introduced in elementary education. Additionally,

every problem within the collection is meticulously categorized based on its type and educational grade level, providing a clear indication of its difficulty level.

Existing MWP corpora can be categorized into four main groups: (1) Number Word Problem corpora, which contain problems related to numbers exclusively; (2) Arithmetic Word Problem corpora, which involve the four basic arithmetic operations and can be either single-step or multi-step problems; (3) Algebraic Word Problem corpora, which focus on algebraic MWPs; and (4) Mixed-type MWP corpora, which are large-scale collections of MWPs from daily algebra or GRE/GMAT examinations. SVAMP Patel et al. (2021c), is a collection of 1,000 math word problems (MWPs), created by introducing variations to initial examples from the ASDiv-A dataset. This compilation features 26 distinct equation models, with each problem incorporating an average of 1.24 operations. Although SVAMP's Corpus Lexicon Diversity (CLD) Miao et al. (2020), falls short when compared to ASDiv-A, it presents a higher level of difficulty. The creators of SVAMP challenge the notion that lexical diversity is a definitive measure of quality in MWP datasets. SVAMP's target audience is students at the elementary school level.

Geometry Problem Solving

GeoS Seo et al. (2015) comprises 186 shaded area problems in geometry. This dataset combines text understanding and diagram interpretation. In contrast, GeoShader Alvin et al. (2017) is a smaller dataset containing 102 shaded area problems. These problems are sourced from standard mathematics textbooks from the United States and released exams from the Indian Class X examination. Another benchmark, GEOS++ Sachan et al. (2017) includes 1,406 questions mirroring the style of SAT exams, covering content from grades 6 through 10. This dataset is segmented into training (350 questions), development (150 questions), and testing (906 questions) subsets, ensuring a balanced representation of questions from each grade level. The authors provide ground-truth logical forms for the 500 annotated questions in the training and development sets. Similarly, GEOS-OS Sachan and Xing (2017) comprises 2,235 geometry problems with demonstrations sourced from a set of grade 6-10 Indian high school math textbooks. As a numerical reasoning benchmark that incorporates multi-modality, GeoQA Chen et al. (2021a) stands out. GeoQA includes 4,998 geometric problems, each accompanied by annotated programs. Notably, GeoQA surpasses previous benchmarks such as GeoS Seo et al. (2015) and GEOS++ Sachan et al. (2017) in terms of size and diversity.

UniGeo Chen et al. (2022a) is a comprehensive and large-scale benchmark for geometry problems. It includes 4,998 calculation problems sourced from GeoQA Chen et al. (2021a), along with an additional 9,543 proving problems. The proving problems are split into train, validation, and test sets in the proportion of 7.0:1.5:1.5. Each problem is labeled with reasons and mathematical expressions in a way of constituting a multi-step proof. To evaluate model performance, the authors define five sub-tasks: Parallel, Triangle, Quadrangle, Congruent, and Similarity, providing detailed insights into model capabilities. Geometry3K Lu et al. (2021a), on the other hand, consists of 3,002 multiple-choice geometry problems with dense annotations in formal language.

This dataset encompasses a wide range of geometric shapes and objectives, featuring SAT-like problems sourced from two high-school textbooks.

Math Question Answering Datasets

The AQuA dataset Ling et al. (2017b) comprises 100,000 samples of questions, answers, and rationales, with a focus on program induction through rationale generation. Each question in the dataset is divided into four parts: the problem description (the "question"), the answer "options" (in a multiple-choice format), the "rationale" description used to arrive at the correct answer, and the label indicating the "correct option". MathQA Amini et al. (2019), an extension of the AQuA dataset Ling et al. (2017b), enhances it by including fully-specified operational programs. MathQA contains 37,000 English multiple-choice math word problems spanning various mathematical domains. The dataset is randomly divided into training, development, and test sets using an 80/12/8% split ratio. Specifically tailored for LLMs, Advanced Reasoning Benchmark (ARB) Sawada et al. (2023) is designed to provide more challenging problems in advanced reasoning across multiple fields. ARB includes problems from diverse domains such as mathematics, physics, biology, chemistry, and law. Its purpose is to serve as a benchmark that surpasses the difficulty levels of previous benchmarks, pushing the boundaries of advanced reasoning tasks.

IconQA Lu et al. (2021b), is an expansive question-answering dataset featuring 107,439 questions. It includes three distinct sub-tasks: choosing from multiple images, selecting from various text options, and completing blank spaces in sentences. The dataset is split into train, validation, and test sets at the proportion of 6:2:2. Icon645 Lu et al. (2021b) contains 645,687 colored icons belonging to 377 classes. These icon-based question-answering pairs enable the evaluation of various reasoning skills, including visual reasoning and commonsense reasoning. MultiHiertt Zhao et al. (2022b) is a dataset expertly annotated with 10,440 QA pairs, centered on questions and answers pertaining to Multi Hierarchical Tabular and Textual data. This dataset is derived from a variety of financial reports. Documents within MultiHiertt include several tables, mostly hierarchical, accompanied by substantial unstructured text. The complexity and challenge of the reasoning required for each question in MultiHiertt surpass those in existing benchmarks. Detailed annotations of the reasoning steps and supporting facts are included to highlight complex numerical reasoning.

In the realm of Textual QA datasets, DROP (Discrete Reasoning Over the content of Paragraphs) Dua et al. (2019) comprises 96,567 questions curated from a diverse range of Wikipedia categories, with a particular focus on sports game summaries and historical narratives. It aims to support methods that combine distributed representations with symbolic and discrete reasoning techniques.

Moving on to Tabular QA datasets, WTQ (Wiki Table Questions) or WikiTableQA Pasupat and Liang (2015) consists of 22,033 complex question-answer pairs derived from 2,108 HTML tables extracted from Wikipedia. WTQ is tailored to support question-answering tasks specifically focused on semi-structured tables. Additionally, WikiSQL Zhong et al. (2018) concentrates on simple SQL queries and single tables. This dataset includes 80,654 meticulously annotated examples of questions and corresponding SQL queries. Spanning 24,241 tables from Wikipedia, WikiSQL

stands out for its substantial scale, surpassing comparable datasets in size. Spider Yu et al. (2018) is tailored for complex, cross-domain semantic parsing and text-to-SQL challenges. This dataset consists of 10,181 questions and 5,693 unique, intricate SQL queries spread across 200 databases. These databases consist of multiple tables covering 138 different domains. Furthermore, Yu et al. (2018) propose a novel task for the text-to-SQL problem using the Spider dataset. AIT-QA (Airline Industry Table QA) Katsis et al. (2022) is tailored for complex and domain-specific Table QA tasks, with a specific focus on the airline industry. The resulting test dataset comprises 515 questions generated from 116 tables. These tables are selected from the 10-K forms of 13 airlines, covering the years between 2017 and 2019. HiTab Cheng et al. (2022) focuses on question-answering (QA) and natural language generation (NLG) tasks specifically tailored for hierarchical tables. This cross-domain dataset is constructed from a rich collection of statistical reports and Wikipedia pages, exhibiting unique characteristics: Firstly, the majority of tables in HiTab are hierarchical, adding complexity to the dataset. Secondly, the questions in HiTab are not generated from scratch by annotators but rather revised from real and meaningful sentences authored by analysts. Lastly, to uncover intricate numerical reasoning in data analysis, fine-grained annotations of quantity and entity alignment are provided. The dataset consists of 3,597 tables, divided into train (70%), dev (15%), and test (15%) sets with no overlap.

For Hybrid QA Dataset, HybridQA Chen et al. (2020a) is a comprehensive and extensive question-answering dataset designed to challenge reasoning abilities on heterogeneous information sources. This dataset encompasses approximately 70,000 question-answering pairs that are aligned with 13,000 Wikipedia tables. It aims to evaluate the ability to reason and extract information from diverse and varied data sources. Free917 Cai and Yates (2013) comprises 917 questions sourced from 81 domains within the Freebase database. Freebase is an online, user-contributed, relational database that covers a wide range of knowledge domains. The dataset involves 635 Freebase relations, which have been annotated with lambda calculus forms. However, due to the requirement for logical forms, scaling up the Free917 dataset becomes challenging as it necessitates expertise in annotating logical forms. In contrast, WebQuestions Berant et al. (2013) contains question-answer pairs gathered from non-experts. It presents a higher number of word types compared to datasets like ATIS Hemphill et al. (1990), posing greater difficulties in lexical mapping. Nevertheless, WebQuestions exhibits simpler structural complexity, with many questions consisting of a unary, a binary, and an entity. This dataset comprises 5,810 question-answer pairs. The questions were collected using the Google Suggest API, while the answers were curated from Freebase with the assistance of Amazon MTurk. WebQuestionsSP (WebQSP) Yih et al. (2016) is derived from WebQuestions Berant et al. (2013) that includes semantic parses for questions answerable using Freebase. It provides SPARQL queries for 4,737 questions, making it possible to directly execute them on Freebase. WebQSP is larger in size compared to Free917 Cai and Yates (2013) and offers semantic parses in SPARQL format with standard Freebase entity identifiers. For evaluating reading comprehension (RC) and question-answering (QA) tasks, the dataset WebQComplex or ComplexWebQuestions Talmor and Berant (2018) proves valuable. It consists of 34,689 complex examples of broad and intricate questions, accompanied by answers, web

snippets, and SPARQL queries. The dataset automatically generates more complex queries involving function composition, conjunctions, superlatives, and comparatives. MetaQA (MoviE Text Audio QA) Zhang et al. (2017) is a comprehensive dataset comprising over 400,000 questions designed for both single and multi-hop reasoning. It also provides more realistic versions in text and audio formats. MetaQA expands upon WikiMovies Miller et al. (2016) and serves as a comprehensive extension to it. These datasets, WebQuestionsSP Yih et al. (2016), WebQComplex Talmor and Berant (2018), and MetaQA Zhang et al. (2017), offer valuable resources for various question answering and reasoning tasks, catering to different complexities and domains.

In the realm of text-centric datasets featuring singular passages, the Stanford Question Answering Dataset (SQuAD) Rajpurkar et al. (2016) stands out as a significant reading comprehension collection. It contains over 100,000 questions, all crafted by crowdworkers using a range of Wikipedia articles. SQuAD uniquely pairs each question with a specific reading passage, and the answer to each question is a segment extracted directly from that passage. The dataset encompasses a total of 107,785 question-answer pairs across 536 articles. A distinctive feature of SQuAD is its absence of pre-defined answer choices for the questions, unlike some other datasets in this category. Instead, systems are required to select the answer from all possible spans within the passage, posing the challenge of dealing with a relatively large number of candidate answers.

For open-domain text-only datasets, TriviaQA Joshi et al. (2017) is another reading comprehension dataset that includes over 650,000 question-answer-evidence triples. TriviaQA consists of 95,000 question-answer pairs contributed by trivia enthusiasts. Furthermore, for each question in the dataset, an average of six independent evidence documents are compiled, offering robust distant supervision that enhances the quality of question-answering support. This makes TriviaQA an invaluable tool for assessing the capabilities of systems to understand and respond to questions within an open-domain context. HotpotQA Yang et al. (2018) is a dataset comprising 113,000 question-answer pairs sourced from Wikipedia. It exhibits four key features: The questions in HotpotQA necessitate the identification and reasoning based on various supporting documents to deduce answers. The dataset includes a broad spectrum of questions that are not confined to established knowledge bases or specific knowledge frameworks. HotpotQA offers crucial sentence-level supporting facts necessary for the reasoning process. This level of granularity enables QA systems to reason with robust supervision and provide explanations for their predictions. Additionally, HotpotQA introduces a new type of factoid comparison question. These questions evaluate the ability of QA systems to extract relevant facts and effectively perform necessary comparisons. By incorporating these four key features, HotpotQA offers a comprehensive and challenging dataset for evaluating QA systems' capabilities in multi-document reasoning, generalization, explanation generation, and factoid comparison. Natural-QA Kwiatkowski et al. (2019) is a question-answering dataset comprising real anonymized queries that were aggregated from interactions with the Google search engine. Natural-QA includes a total of 307,373 training examples that are publicly available. For the development data, there are 7,830 examples that have been annotated with 5 possible answers. Additionally, the test data consists of another 7,842 examples, also annotated with 5-way annotations.

MultiModalQA (MMQA) Talmor et al. (2021) is an intricate question-answering dataset designed to challenge models in joint reasoning across multiple modalities, including text, tables, and images. The dataset comprises 29,918 questions, and a notable 35.7% of these questions require cross-modality reasoning. GeoTSQA Li et al. (2021b) is a dataset that focuses on the tabular scenario-based question answering (TSQA) task within the domain of geography. It consists of 556 scenarios accompanied by 1,012 real multiple-choice questions that are contextualized within these tabular scenarios.

TheoremQA Chen et al. (2023f) is a question-answering dataset that revolves around the concept of theorems. It consists of 800 high-quality questions, which cover 350 theorems spanning various disciplines such as Mathematics, Physics, Electrical Engineering and Computer Science (EE&CS), and Finance. TAT-QA Zhu et al. (2021) serves as a question-answering benchmark specifically focused on the domain of finance. The dataset evaluates the ability of models to answer questions based on both tables and text. The questions in TAT-QA often require numerical reasoning skills, such as performing arithmetic operations, counting, comparing or sorting values, and combining multiple reasoning steps. The benchmark encompasses 16,552 questions associated with 2,757 hybrid contexts derived from real-world financial reports. It covers a wide range of finance-related topics and scenarios, including stock prices, financial reports, bank transactions, and currency exchange rates. FinQA Chen et al. (2021d) is a dataset specifically designed to facilitate numerical reasoning tasks with financial data. The dataset comprises 8,281 question-answer pairs that revolve around financial calculations. Importantly, each pair is accompanied by detailed reasoning steps that provide insights into the process of arriving at the answer.

Metrics

GeoS++ Sachan et al. (2017) employs Normalized Mutual Information (NMI) Strehl and Ghosh (2002) to assess the quality of axiom mention clustering. To measure the lexicon usage diversity of a given MWP corpus, Miao et al. (2020) introduced the use of BLEU Papineni et al. (2002). They also proposed the Corpus Lexicon Diversity (CLD) metric to assess the lexical diversity of a given corpus Miao et al. (2020). Cheng et al. (2022) adopt Execution Accuracy (EA) as their evaluation metric. This approach follows the methodology proposed by Pasupat and Liang (2015), which measures the percentage of samples with correct answers. To evaluate the performance of TSQA models on GeoTSQA, Li et al. (2021b) employ two standard information retrieval evaluation metrics: Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR). These metrics provide quantitative measures of the models' retrieval effectiveness and ranking accuracy when answering questions based on the tabular scenarios.

3.10.3 Logical Reasoning

There are four notable logical reasoning datasets: ProofWriter Tafjord et al. (2021), PrOntoQA Saparov and He (2023), FOLIO Han et al. (2022), and LogicalDeduction from BIG-Bench Srivastava et al. (2023).

Dataset	Train Size	Dev size	Test size	Task Type	Synthetic	Type
αNLI Bhagavatula et al. (2019)	169,654	-	1532	NLI	✗	Abductive
ProofWriter Tafjord et al. (2021)	69,814	10,158	20,058	FV	✓	Deductive
FOLIO Han et al. (2022)	1,004	204	227	FV	✗	Deductive
LogicalDeduction Srivastava et al. (2023)	-	-	1300	FV	✗	Deductive
PrOntoQA Saparov and He (2023)	-	-	200	MCQA	✗	Deductive

Table 6: Logical Reasoning Benchmarks Luo et al. (2023d). There are three types of tasks: multiple choice question answer (MCQA); natural language inference (NLI); and fact verification (FV).

ProofWriter Tafjord et al. (2021) builds upon the original RuleTaker D* datasets Clark et al. (2020) and introduces two additional variants. The closed-world assumption (CWA) variant addresses minor inconsistencies related to negation, while the open-world assumption (OWA) variant incorporates an open-world assumption during reasoning. The RuleTaker D* datasets Clark et al. (2020) consist of five subsets (D0, D1, D2, D3, and D5), each containing 100k questions. PrOntoQA (Proof and Ontology-Generated Question-Answering) Saparov and He (2023) contains examples generated from synthetic world models represented in first-order logic. FOLIO Han et al. (2022) is a human-annotated, open-domain dataset that covers a wide range of logical complexities and diversities. It provides first-order logic (FOL) annotations and consists of 1,435 examples. Additionally, the dataset includes 487 sets of premises that serve as rules for deductive reasoning to evaluate the validity of conclusions. The LogicalDeduction task Srivastava et al. (2023) within BIG-Bench serves as an evaluation benchmark for assessing the ability to perform multi-step logical reasoning. This task involves deducing the order of a sequence of objects based on a minimal set of given conditions. Each instance includes a naturally ordered context with three to seven similar objects, such as differently colored books on a shelf. Alongside the context, a set of simple clues is provided. More relevant datasets and their corresponding statistics are presented in Table 6.

These logical reasoning datasets contribute to the development and evaluation of models and systems that aim to enhance logical reasoning capabilities. They provide diverse scenarios and challenges, ranging from synthetic world models to real-world contexts, enabling researchers to explore and advance logical reasoning in various domains.

3.10.4 Causal Reasoning

The Tübingen cause-effect pairs dataset Mooij et al. (2016) encompasses 108 cause-effect pairs obtained from 37 datasets spanning diverse domains, including meteorology, biology, medicine, engineering, and economics. This dataset serves as a

benchmark for evaluating causal reasoning abilities. In contrast, the Neuropathic Pain dataset Tu et al. (2019) focuses on the relationships between nerves and the corresponding symptoms observed in patients. Due to its specialized medical terminology and domain-specific knowledge, interpreting the variable names within this dataset requires expertise in the field. The Arctic sea ice dataset Huang et al. (2021c) presents a graph derived from domain knowledge, featuring 12 variables and 48 edges. It offers valuable insights into the dynamics of Arctic sea ice.

Counterfactual reasoning, even in the absence of actual causality, is a valuable capability for language models. It aids in decision-making, planning, and uncovering hidden insights that may not be immediately apparent in the original context. CRASS (Counterfactual Reasoning Assessment) Frohberg and Binder (2021) is a benchmark specifically developed to evaluate the proficiency of language models in dealing with counterfactual queries. This benchmark comprises 275 instances in which the language model is presented with counterfactual conditional questions. In each instance, the model is tasked with selecting the most suitable response from a provided set of multiple-choice options.

In evaluating the performance of language models on causal reasoning benchmarks and datasets like CRASS, the commonly used metric is top-k accuracy Frohberg and Binder (2021). This metric quantifies the model's ability to make correct predictions by considering the top k ranked choices. It serves as a quantitative measure of the model's proficiency in causal reasoning tasks. Percentage of Preference Li et al. (2023f) is a metric used to evaluate logical completions in both counterfactual and factual scenarios. This metric provides a quantitative measure of the extent to which a language model's generated completions align with human preferences and judgments in terms of logical consistency.

3.10.5 Visual Reasoning

In order to establish a benchmark for grounded grammar induction, researchers have curated a large-scale dataset known as PARTIT Hong et al. (2021a). This dataset consists of human-written sentences that provide detailed descriptions of part-level semantics for 3D objects. PTR Hong et al. (2021b) is an extensively curated dataset tailored for visual reasoning analysis. It includes around 70,000 synthetic RGB-D images, each accompanied by detailed ground truth data on objects and part-level annotations. These annotations cover a range of aspects such as spatial and geometric relationships, semantic instance segmentation, color attributes, and key physical properties like stability. PTR is designed to facilitate research on part-based conceptual, relational, and physical reasoning. Compositional Language and Elementary Visual Reasoning (CLEVR) Johnson et al. (2017) is a widely used diagnostic benchmark that evaluates a wide array of visual reasoning abilities. It consists of 100,000 rendered images and around one million automatically generated questions, with 853,000 unique questions. CLEVR offers a challenging set of images and questions designed to assess various aspects of visual reasoning, including tasks like counting, comparing, logical reasoning, and memory retention. This dataset provides a robust platform for testing and advancing visual reasoning algorithms and models, enabling researchers to

Tasks	Cat.	Evaluation Metric
phone recognition	discr.	phone error rate (PER)
automatic speech recognition	discr.	word error rate (WER)
keyword spotting	discr.	accuracy (ACC)
query by example spoken term detection	discr.	maximum term weighted value (MTWV)
speaker identification	discr.	accuracy (ACC)
automatic speaker verification	discr.	equal error rate (EER)
speaker diarization	discr.	diarization error rate (DER)
intent classification	discr.	accuracy (ACC)
slot filling	discr.	F1-score and character error rate (CER)
emotion recognition	discr.	accuracy (ACC)
voice conversion	gen.	mel-cepstrum distortion (MCD)
speech separation	gen.	scale-invariant signal-to-distortion ratio improvement (SI-SDRi)
speech enhancement	gen.	perceptual evaluation of speech quality (PESQ) short time objective intelligibility (STOI)

Table 7: Metrics of Audio Reasoning Tasks. Here "Cat." denotes the category of the tasks. "discr." and "gen." stand for discriminative and generative tasks.

explore and enhance their capabilities in this field. Outside Knowledge Visual Question Answering (OK-VQA) Marino et al. (2019) is a dataset specifically designed for visual question-answering tasks that necessitate the utilization of external knowledge to generate accurate answers. The dataset comprises 14,055 open-ended questions, each associated with five ground truth answers. During the annotation process, the questions were carefully filtered to ensure that they all required external knowledge, such as information from sources like Wikipedia. Additionally, efforts were made to mitigate dataset bias by reducing questions with frequently occurring answers. This dataset serves as a valuable resource for developing and evaluating methods that can effectively leverage external knowledge for visual question-answering tasks.

3.10.6 Audio Reasoning

The most widely adopted benchmark datasets for different aspects of audio reasoning, i.e., the Speech processing Universal PERformance Benchmark (SUPERB) Yang et al. (2021) for discriminative tasks and the enhanced Speech processing Universal PERformance Benchmark (SUPERB-SG) Tsai et al. (2022) for generative tasks, have been introduced in Section 3.6. The evaluation metrics for these tasks are listed in Table 7.

The availability of datasets in a wide variety of languages contributes to the success of self-supervised learning (SSL) of speech representations, which lays the crucial foundation for audio reasoning. One of the largest and most widely utilized speech corpora for foundation model pre-training is the Libri-light Kahn et al. (2020) dataset, which contains approximately 60,000 hours of speech in English originating from audiobooks. Didi Dictation and Didi Callcenter Jiang et al. (2021b) are large-scale corpora in Chinese, each containing roughly 10,000 hours of data collected from mobile dictation applications or phone calls. Apart from English and Chinese, multilingual corpora of substantial sizes are available as well, including VoxPopuli Babu et al. (2022) (400,000 hours, 23 languages), Multilingual LibriSpeech Pratap et al. (2020) (50,000 hours, 8 languages) and Common Voice Ardila et al. (2020) (11,000 hours, 76 languages).

Since there is typically no ground-truth transcription available, it is impractical to utilize these datasets to train the conventional hidden Markov models (HMMs) and the supervised deep neural network (DNN) or end-to-end (E2E) models. The advance of SSL-based models leverages the power of data and provides a good and universal starting point for further fine-tuning using labeled data for the aforementioned downstream tasks.

3.10.7 Multimodal Reasoning

In their comprehensive study, Liu et al. (2023m) conducted an in-depth evaluation of publicly accessible multimodal models, focusing on their efficacy in a range of text-centric tasks. These tasks include text recognition, encompassing scene text, artistic text, and handwritten text; text-based visual question answering, which involves document text, scene text, and bilingual text; key information extraction from various sources such as receipts, documents, and nutrition facts labels; and the recognition of handwritten mathematical expressions. The study identified both strengths and weaknesses in these models. While they excel in word recognition through semantic understanding, they struggle when it comes to perceiving combinations of characters lacking semantic meaning. Additionally, the models exhibit consistent performance regardless of text length and have limited capabilities in detecting intricate image details. Overall, the study concludes that even the most powerful existing multimodal models fall short compared to domain-specific methods in traditional text tasks. These findings underscore the necessity for innovative strategies to enhance zero-shot multimodal techniques and improve model performance in complex tasks.

Regarding evaluation benchmarks, Vedantam et al. (2015) contribute to the field by introducing two datasets: PASCAL-50S and ABSTRACT-50S. These datasets are designed for evaluating image caption generation methods. They serve as valuable resources that enable researchers to assess the performance and quality of image captioning models. By utilizing these datasets, researchers can advance image caption generation techniques, fostering progress and innovation in this area of research. LVLM-eHub Xu et al. (2023a) serves as a comprehensive and extensive evaluation benchmark for publicly available large multimodal models. It rigorously assesses the performance of eight LVLMs across six categories of multimodal capabilities. The evaluation process involves the utilization of 47 datasets and 1 arena online platform, providing a robust and standardized framework for evaluating LVLMs. Odouard and Mitchell (2022) present a concept-based approach to systematic evaluations, with a focus on assessing the proficiency of AI systems in utilizing a given concept across different instances. Their evaluation approach, as described in the cited work, entails conducting case studies within two specific domains: RAVEN, which is influenced by the Raven's Progressive Matrices Raven and Court (1938), and the Abstraction and Reasoning Corpus (ARC) Acquaviva et al. (2021). These are frequently utilized for assessing and advancing the capacity for abstraction in AI systems. This methodology provides valuable information about the AI systems' understanding and application of abstract reasoning abilities, shedding light on their ability to grasp and utilize concepts effectively. In a related study, Yin et al. (2023e) expanded the research on Multimodal Large Language Models (MLLMs) by incorporating point clouds. They

introduced the LAMM-Dataset and LAMM-Benchmark, which specifically focus on improving 2D image and 3D point cloud understanding.

Hallucination is a well-known issue and has long been present in multimodal foundation models as well. Recent studies Dai et al. (2023); Li et al. (2023n) have investigated the performance of Visual Language Pretraining (VLP) models and Vision and Language Models (VLMs) in terms of object hallucination. Dai et al. (2023) discovered that despite advancements in VLP models, hallucinations remain a common issue. Interestingly, the study revealed that models with higher scores on conventional metrics like CIDEr Vedantam et al. (2015) tended to exhibit more unfaithful results. The authors also found that patch-based features yielded the best results, with smaller patch resolutions reducing object hallucination. To tackle this issue, They proposed a straightforward yet effective VLP loss called Object Masked Language Modeling (ObjMLM) Dai et al. (2023), which further mitigates object hallucination. By decoupling various VLP objectives, the authors demonstrated the importance of token-level image-text alignment and controlled generation in reducing hallucination. Similarly, Li et al. (2023n) conducted evaluation experiments on representative VLMs and discovered widespread object hallucination issues. They also found that visual instructions can influence hallucination, with objects that frequently appear in the instructions or co-occur with image objects being more susceptible to hallucination by VLMs. Furthermore, the authors observed that existing evaluation methods may be influenced by the input instructions and generation styles of VLMs. To address this concern, they proposed an improved evaluation method called POPE (Polling-based Object Probing Evaluation) to assess object hallucination more effectively. Zhao et al. (2023e) put forth a methodology for examining the robustness of open-source large VLMs in realistic and high-risk scenarios, where adversaries have limited black-box system access and aim to deceive the model into producing targeted responses. The authors begin by crafting targeted adversarial examples against pre-trained models like CLIP Radford et al. (2021) and BLIP Li et al. (2022b). They later transfer these adversarial examples to other VLMs, including MiniGPT-4 Zhu et al. (2023b), LLaVA Liu et al. (2023e), UniDiffuser Bao et al. (2023), BLIP-2 Li et al. (2023e), and Img2Prompt Guo et al. (2023a). The authors discovered that the effectiveness of targeted evasion in large VLMs can be significantly enhanced by employing black-box queries. This approach yields a surprisingly high success rate in generating targeted responses, thereby highlighting the vulnerability of these models to adversarial attacks. In addition, Huang et al. (2023c) introduce T2I-CompBench, a comprehensive benchmark for assessing text-to-image generation models' capabilities in processing compositional prompts. It evaluates how these T2I models interpret and represent compositional concepts, such as attribute binding, object relationships, and complex compositions. T2I-CompBench also delves into effectively utilizing multimodal LLMs for evaluation in this context.

Regarding the metrics, DePlot Liu et al. (2023b) introduces a metric called Relative Number Set Similarity (RNSS) for comparing table similarity. RNSS takes into account the table's structure and numeric values while remaining unaffected by column/row permutations. CIDEr, as introduced by Vedantam et al. (2015), is presented as an automated metric for the evaluation of image captioning.

Dataset	Tasks	Size
Swag Zellers et al. (2018)	Commonsense	113,000
PHYRE Bakhtin et al. (2019)	Commonsense	25
HellaSwag Zellers et al. (2019)	Commonsense	70,000
WinoGrande Sakaguchi et al. (2021)	Commonsense	44,000
Social IQA Sap et al. (2019)	Commonsense	35,350
PIQA Bisk et al. (2020)	Commonsense	21,020
SummEdits Laban et al. (2023)	Commonsense	6,348
CConS Kondo et al. (2023)	Commonsense	1,112
Alg514 Kushman et al. (2014)	Math	514
Verb395 Hosseini et al. (2014)	Math	395
Dolphin1878 Shi et al. (2015)	Math	1878
DRAW Upadhyay and Chang (2015)	Math	1000
SingleEQ Koncel-Kedziorski et al. (2015)	Math	508
Dolphin18K Huang et al. (2016)	Math	18,000
MATH Hendrycks et al. (2021b)	Math	12,500
TabMWP Lu et al. (2022b)	Math	38,431
GSM8K Cobbe et al. (2021)	Math	8,500
MGSM Shi et al. (2023)	Math	250
Math23K Wang et al. (2017)	Math	23,161
HMWP Qin et al. (2020)	Math	5,491
ASDiv Miao et al. (2020)	Math	2,305
SVAMP Patel et al. (2021c)	Math	1,000
GeoS Seo et al. (2015)	Math	186
GeoShader Alvin et al. (2017)	Math	102
GEOS++ Sachan et al. (2017)	Math	1,406
GEOS-OS Sachan and Xing (2017)	Math	2,235
GeoQA Chen et al. (2021a)	Math	4,998
UniGeo Chen et al. (2022a)	Math	4,998
Geometry3K Lu et al. (2021a)	Math	3,002
AQuA dataset Ling et al. (2017b)	Math	100,000
MathQA Amini et al. (2019)	Math	37,000
ARB Sawada et al. (2023)	Math (physics, biology, chemistry, law)	1,207
IconQA Lu et al. (2021b)	Math	107,439
MultiHiertt Zhao et al. (2022b)	Math	10,440
DROP Dua et al. (2019)	Textual QA	96,567
WTQ Pasupat and Liang (2015)	Textual QA	22,033
WikiSQL Zhong et al. (2018)	Textual QA	80,654
Spider Yu et al. (2018)	Textual QA	10,181
HybridQA Chen et al. (2020a)	Hybrid QA	$\sim$ 70,000
MetaQA Zhang et al. (2017)	Hybrid QA	$\sim$ 400,000
SQuAD Rajpurkar et al. (2016)	Hybrid QA	$\sim$ 100,000
TriviaQA Joshi et al. (2017)	Hybrid QA	95,000
HotpotQA Yang et al. (2018)	Hybrid QA	113,000
MMQA Talmor et al. (2021)	Hybrid QA	29,918
TheoremQA Chen et al. (2023f)	Hybrid QA	800
TAT-QA Zhu et al. (2021)	Hybrid QA	16,552
FinQA Chen et al. (2021d)	Hybrid QA	8,281

Table 8: Summary of Some Reasoning Datasets 1

In summary, these studies contribute to the development of innovative models, benchmarks, and evaluation metrics in the field of multimodal understanding. They explore various aspects of multimodal reasoning, including visual instruction-following, reasoning on charts, object detection, and image-related tasks. The proposed approaches and evaluations shed light on the strengths and weaknesses of existing models and highlight the need for further advancements in multimodal techniques.

Dataset	Tasks	Size
APPS Hendrycks et al. (2021a)	Program Synthesis	10,000
HumanEval Chen et al. (2021b)	Program Synthesis	164
MathQA-Python Austin et al. (2021)	Program Synthesis	23,914
MBPP Austin et al. (2021)	Program Synthesis	974
αNLI Bhagavatula et al. (2019)	Logical	171,186
ProofWriter Tafjord et al. (2021)	Logical	100,030
FOLIO Han et al. (2022)	Logical	1,435
PrOntoQA Saparov and He (2023) Srivastava et al. (2023)	Logical	200
LogicalDeduction Srivastava et al. (2023)	Logical	1,300
Tübingen cause-effect pairs dataset Mooij et al. (2016)	Causal	108
Neuropathic Pain dataset Tu et al. (2019)	Causal	N/A
Arctic sea ice dataset Huang et al. (2021c)	Causal	N/A
CRASS Frohberg and Binder (2021)	Causal	275
PARTIT Hong et al. (2021a)	Visual	$\sim$ 10,000
PTR Hong et al. (2021b)	Visual	70,000
CLEVR Johnson et al. (2017)	Visual	100,000
OK-VQA Marino et al. (2019)	Visual	14,055
VoxPopuli Babu et al. (2022)	Audio	400,000 hours
Libri-light Kahn et al. (2020)	Audio	60,000 hours
Multilingual LibriSpeech Pratap et al. (2020)	Audio	50,000 hours
Common Voice Ardila et al. (2020)	Audio	11,000 hours
Didi Dictation Jiang et al. (2021b)	Audio	10,000 hours
Didi Callcenter Jiang et al. (2021b)	Audio	10,000 hours
PASCAL-50S Vedantam et al. (2015)	Multimodal	1,000
ABSTRACT-50S Vedantam et al. (2015)	Multimodal	500
LVLM-eHub Xu et al. (2023a)	Multimodal	47 sub-dataset
LAMM-Dataset Yin et al. (2023e)	Multimodal	25 sub-datasets
RoboTHOR Deitke et al. (2020)	Embodied	/
VirtualHome Puig et al. (2018)	Embodied	/
Gibson Xia et al. (2018)	Embodied	/
BEHAVIOR-1K Li et al. (2022a)	Embodied	/
Habitat Manolis Savva* et al. (2019)	Embodied	/
DriveLM DriveLM Contributors (2023)	Driving	360,000
nuScenes QA Qian et al. (2023b)	Driving	460,000
HAD Kim et al. (2019)	Driving	5,675

Table 9: Summary of Some Reasoning Datasets 2

3.10.8 Embodied Reasoning

RoboTHOR Deitke et al. (2020) is a platform designed to develop and test embodied AI agents in both simulated and physical environments. VirtualHome Puig et al. (2018) is another platform that focuses on modeling complex activities occurring in typical household settings. It offers support for program descriptions that cover a wide variety of activities found in people's homes. Gibson Xia et al. (2018) places emphasis on real-world perception for embodied agents. To bridge the gap between simulation and reality, iGibson Li et al. (2021a) and BEHAVIOR-1K Li et al. (2022a) extend the simulation capabilities to encompass a more diverse range of household

tasks and achieve high levels of realism. These platforms provide researchers with tools to explore and evaluate embodied AI approaches in realistic simulated environments. Habitat Manolis Savva* et al. (2019) boasts high performance, reaching several thousand frames per second (fps) even when running single-threaded. Habitat-Lab Szot et al. (2021) is a high-level, modular library that supports comprehensive development in the realm of embodied AI. It enables the specification of a range of embodied AI tasks, including navigation, interaction, following instructions, and answering questions. This platform allows researchers to tailor embodied agents with particular physical attributes, sensors, and functionalities, and to evaluate their performance on these tasks using established metrics.

These simulation platforms hold great potential for evaluating LLMs on robotics tasks. By leveraging these simulators, researchers can assess the performance and capabilities of LLMs in the context of real-world scenarios, further advancing the field of embodied AI.

3.10.9 Autonomous Driving

DriveLM DriveLM Contributors (2023) is a comprehensive driving benchmark to investigate the role of LLMs in various aspects. It introduces the reasoning ability of Large Language Models in autonomous driving to guarantee explainable planning and thus make safe decisions. The questions and answers (QA) in perception, prediction, and planning modules are connected in a graph-style structure, with QA pairs as nodes, and objects' relationships as edges. Compared to predecessors, such as nuScenes QA Qian et al. (2023b) and HAD Kim et al. (2019), DriveLM draws many merits from them and improves the logical reasoning in more tasks and a wide diversity of scenarios.

3.10.10 Code Generation

Code generation Sun et al. (2021) encompasses several datasets and benchmarks that contribute to the advancement of code generation and evaluation. The APPS dataset Hendrycks et al. (2021a), consists of 10,000 problems derived from coding competitions. It functions as a standard for assessing code generation tasks that are guided by natural language descriptions. This dataset provides a platform for researchers to measure and contrast the effectiveness of models in producing code in response to natural language prompts. Additionally, the study highlights concerns with using BLEU Papineni et al. (2002) as a metric for code generation, suggesting that it may not be reliable in this context. Similarly, HumanEval Chen et al. (2021b) consists of 164 handwritten programming problems and serves as a benchmark for evaluating the performance of Codex Chen et al. (2021b). Each problem in HumanEval includes a function signature, docstring, body, multiple unit tests, and each problem has 7.7 tests on average. MathQA-Python Austin et al. (2021) is a Python version of the MathQA Amini et al. (2019) benchmark. It contains 23,914 problems that evaluate models' ability to synthesize code from complex textual descriptions. Notably, the study found that providing natural language feedback from humans resulted in a significant reduction in error rates compared to the models' initial predictions. The Mostly Basic Programming Problems (MBPP) dataset Austin et al. (2021) consists

of 974 programming tasks specifically designed to be solvable by entry-level programmers. In MBPP, there is a greater emphasis on the usage of imperative control flow structures like loops and conditionals. On the other hand, MathQA-Python Austin et al. (2021) contains more intricate natural language descriptions, offering a higher level of complexity in the problem statements.

4 Foundation Model Techniques

In this section, we provide a concise overview of various foundation model techniques. Here, we present distinct categories of reasoning techniques:

- Pre-Training (Section 4.1): Exploring data and architecture of reasoning foundation models.
- Fine-tuning (Section 4.2): Focusing on reasoning foundation models' fine-tuning data and techniques.
- Alignment Training (Section 4.3): Examining the alignment techniques employed by reasoning foundation models.
- In-Context Learning (Section 4.4): Introducing in-context learning in reasoning foundation models.
- LLM with Agent (Section 4.5): Focusing on the reasoning foundation model as an agent for multiple tasks.

4.1 Pre-Training

In the pre-training part, LLMs can acquire essential language understanding and generation skills. Here, the data and architecture are critical for the foundation model. Therefore, we will discuss them in the following sections.

4.1.1 Data Source

Foundation models are data-driven, and both quality and quantity of data lie at the core of foundation model development. Figure 14 presents three broad types of data sources for foundation model pre-training.

Text Data

The realm of publicly accessible large-scale text datasets has seen a considerable expansion, presenting a rich variety of resources for myriad applications. A prime example is the Pile (Gao et al. , 2020), an extensive English text corpus, notable for its impressive volume of 825 GB, and specifically curated to facilitate the training of large-scale language models. This corpus is comprised of 22 diverse subsets, recognized for their variety and quality, amalgamating both existing and newly created content, with a significant portion sourced from scholarly and professional domains. A considerable amount of this data is amassed through web crawling initiatives, akin to the CommonCrawl project. It is crucial to acknowledge that such web crawling produces a spectrum of content, from high-caliber material like Wikipedia entries to lower-tier content such as spam emails, necessitating rigorous filtering and processing to elevate data quality. Another notable dataset in this field is the C4 dataset (Raffel et al. , 2019), representing

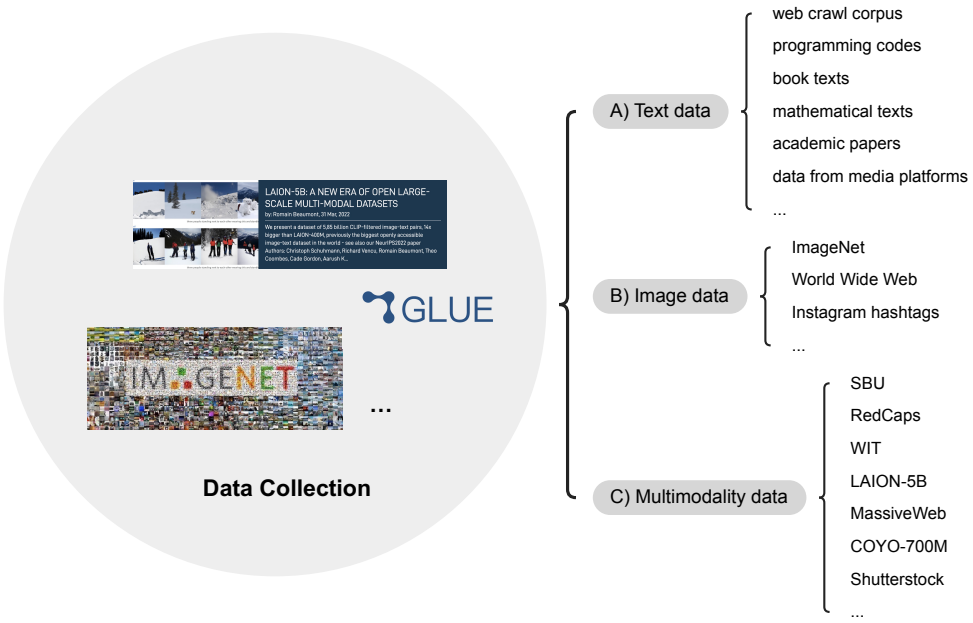


Fig. 14: A diverse suite of data sources and datasets for pre-training foundation models, mainly including text data, image data, and multimodality data.

an expansive and refined version of the Common Crawl web corpus, extensively utilized in various sectors. In contrast, the ROOTS dataset (Laurençon et al. , 2022) emerges as an immense resource, encompassing 1.6TB and spanning 59 languages across 46 natural languages, derived from three macro regions and nine language families. It also includes material in 13 programming languages, with Java, PHP, and C++ comprising the majority of its content. The Gutenberg project (Lahiri, 2014) offers a selection of 3,036 English books by 142 authors. This collection, a subset of the larger Project Gutenberg corpus, has been diligently cleaned to remove metadata, licensing details, and transcribers' notes to the fullest extent. The CLUECorpus (Xu et al. , 2020) stands out in the Chinese text domain as a substantial 100GB resource. This community-led project integrates nine varied tasks, ranging from single-sentence/sentence-pair classifications to machine reading comprehension, all rooted in authentic Chinese text. Additionally, the Proof-Pile dataset (Azerbayev et al. , 2023), with its impressive 8 billion tokens, is notable in the mathematical text sphere. It is distinguished for being among the few open-source language models specifically tuned for the general mathematics field. The peS2o dataset (Soldaini and Lo, 2023), consisting of around 40 million open-access academic papers, is an invaluable asset. It has undergone thorough cleaning, filtering, and formatting, making it ideal for pre-training language models. Originating from the Semantic Scholar Open Research Corpus (S2ORC), it expands the availability of academic text resources. Furthermore, researchers have access to various public conversation datasets, like the Reddit corpus (Roller et al. , 2020).

Data from online social media platforms also offers a wealth of conversational content. Scientific text collections typically focus on aggregating materials such as arXiv papers, scientific textbooks, mathematical websites, and related scientific materials. The complex nature of scientific data, often laden with mathematical symbols and protein sequences, requires specialized tokenization and preprocessing methods for standardization and uniform processing by language models. Recent research (Austin et al. , 2021) highlights the benefits of training Large Language Models (LLMs) on extensive code corpora, leading to marked enhancements in generated program quality. These corpora are often sourced from platforms like StackOverflow and GitHub. Lastly, the RedPajama project (Computer, 2023) deserves mention for its remarkable feat in reproducing LLaMA's training dataset, encompassing an impressive 1.2 trillion tokens. This dataset includes a vast array of tokens from CommonCrawl, C4, GitHub, Books, ArXiv, Wikipedia, and StackExchange, presenting a comprehensive and diverse resource for the development and refinement of language models.

Image Data

The methodology of supervised pre-training using extensive, human-curated datasets like ImageNet (Deng et al. , 2009) and ImageNet21K (Ridnik et al. , 2021) has become a prevalent approach in developing transferable visual representations. This process is structured to create a linkage between an input image and a distinct label, each corresponding to a specific visual concept. With the growing need for large-scale pre-training, the generation of copious amounts of noisy labels from image-text pairings sourced from the World Wide Web has become increasingly relevant. Leveraging these noisy labels, numerous leading industrial research labs have skillfully assembled vast classification datasets using semi-automatic data pipelines. Notable examples of such endeavors include JFT (Sun et al. , 2017) and I2E (Wu et al. , 2023e). Additionally, they have utilized proprietary data sources, such as Instagram hashtags (Singh et al. , 2022), to enrich their datasets further and augment the precision of their pre-trained models. This strategy has significantly contributed to the advancement of sophisticated visual recognition systems, equipping them with the ability to effectively identify and categorize a wide spectrum of visual concepts and objects.

Multimodality Data

The domain of large-scale datasets features several notable examples. SBU (Ordonez et al. , 2011), for instance, executes an extensive number of Flickr queries and then rigorously filters the results to produce 1 million images, each paired with a caption that is visually pertinent. Conversely, RedCaps (Desai et al. , 2021) is a substantial dataset encompassing 12 million image-text pairs, sourced from Reddit. The WIT dataset (Srinivasan et al. , 2021) is distinguished by its curated compilation of 37.6 million image-text instances, enhanced with entity information, covering 108 Wikipedia languages, and incorporating 11.5 million unique images. Other relatively large datasets in this field include Shutterstock (Nguyen et al. , 2022), LAION-400M (Schuhmann et al. , 2021), and COYO-700M (Byeon et al. , 2022). OpenAI's CLIP (Radford et al. , 2021) was refined through an impressive collection of 400 million image-text pairs, meticulously sourced from the web. Recently, the emergence of

datasets at the billion-scale level has been observed. The LAION-5B dataset (Schuhmann et al. , 2022), for instance, comprises 5.85 billion CLIP-filtered image-text pairs, of which 2.32 billion are in the English language. DataComp (Gadre et al. , 2023) functions as a platform for dataset experiments, focusing on a new pool of 12.8 billion image-text pairs collected from Common Crawl. Flamingo (Alayrac et al. , 2022) introduces the MultiModal MassiveWeb (M3W) dataset, aggregating text and images from about 43 million web pages, and aligning images with text according to the Document Object Model (DOM). A noteworthy project in this context is ImageBind (Girdhar et al. , 2023), which aims to develop a joint embedding covering six distinct modalities, including images, text, audio, depth, thermal, and IMU data, with potential extension to other modalities such as point clouds (Guo et al. , 2023b). This ambitious endeavor signifies a major step forward in fostering a deeper comprehension of multimodal data by establishing meaningful links across diverse data types. As multimodal learning advances, these developments in dataset creation and application are crucial to the ongoing innovation in the field.

Data for Reasoning

The significance of code and paper data in enhancing the reasoning abilities of foundation models is paramount. Discussing code data first, research by CoCoGen (Madaan et al. , 2022b) indicates that when structured commonsense reasoning tasks are approached as code generation problems, pre-trained language models (LMs) for code exhibit superior reasoning capabilities compared to those trained on natural language. This holds true even for tasks that do not involve source code. Such code data are readily accessible on GitHub and through various filtered datasets available to the public. Highlighting this, StarCoder (Li et al. , 2023i) has released an extensive pretraining dataset (783GB) to further refine LMs' proficiency in coding. In terms of paper data, Galactica (Taylor et al. , 2022) stands out, having been trained on a vast corpus of scientific papers, reference material, knowledge bases, and other diverse sources. This model demonstrates superior performance across a spectrum of scientific tasks compared to existing models. Paper data primarily originates from academic platforms like Arxiv, with a notable emphasis on mathematics papers. Additionally, the peS2o (Soldaini and Lo, 2023) dataset, encompassing over 40 million open-access academic papers from the Semantic Scholar Open Research Corpus (S2ORC), provides a substantial resource for the pretraining of models.

4.1.2 Network Architecture

The foundation model architecture is essential. We discuss different network architectures in the following and show them in Figure 15.

Encoder-decoder Architecture

The seminal Transformer model, as delineated by Vaswani et al. (2017), is founded on the encoder-decoder framework. This paradigm employs dual stacks of Transformer blocks, wherein one functions as the encoder and the other as the decoder. The encoder phase involves utilizing multi-head self-attention layers in a stacked arrangement to

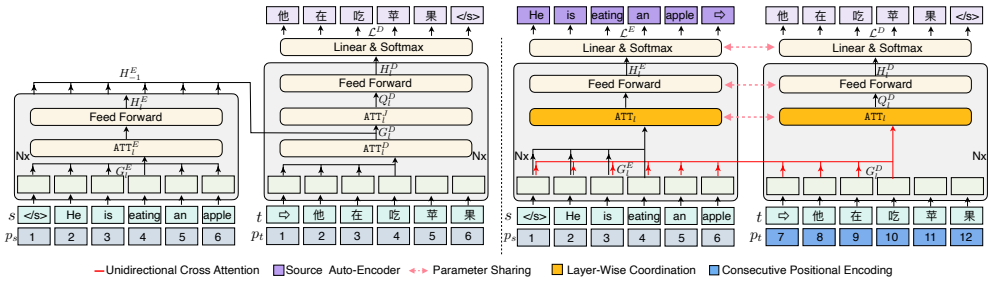


Fig. 15: Illustration of Encoder-Decoder framework (left) and Decoder-Only framework (right). The figure credits from (Fu et al. , 2023c).

decode the intrinsic information within the input sequence, thereby yielding latent representations. In the subsequent phase, the decoder applies cross-attention mechanisms to these representations, facilitating the generation of the target sequence. This innovative architecture is extensively applied in sequence-to-sequence modeling tasks, such as neural machine translation. In a distinctive approach, BERT (Kenton and Toutanova, 2019) is engineered for the pretraining of deep bidirectional representations from the unlabeled text. It uniquely processes both left and right contexts concurrently across all layers, rendering it exceptionally versatile for a plethora of NLP tasks. Conversely, BART (Lewis et al. , 2020) incorporates a conventional Transformer-based architecture for neural machine translation. While its structure may appear straightforward, BART can be regarded as an evolution of BERT, amalgamating the bidirectional encoder of BERT and the unidirectional, left-to-right decoder of GPT, along with other advanced pretraining methodologies. Furthermore, Pre-trained Language Models (PLMs) following the encoder-decoder architecture paradigm, exemplified by T5 (Raffel et al. , 2019), have consistently showcased remarkable performance across a wide array of NLP tasks.

Decoder-only Architecture

The decoder-only architecture is characterized by its strategic use of an attention mask, a pivotal element that ensures each input token is exclusively attentive to preceding tokens, including itself. This unique configuration facilitates a unidirectional flow of information from antecedent tokens to the current token within the decoder, thereby streamlining the processing of input and output tokens. This approach not only simplifies the learning mechanism but also bolsters the model's coherence and consistency. In the domain of language modeling, the GPT (Generative Pre-trained Transformer) series epitomizes the decoder-only architecture. This series encompasses GPT-1 (Radford et al. , 2018), GPT-2 (Radford et al. , 2019), and the notably advanced GPT-3 (Brown et al. , 2020). GPT-3, in particular, serves as a quintessential model within this paradigm, exemplifying the architectural efficacy, especially in in-context learning, a distinguishing feature of Large Language Models (LLMs). The decoder-only architecture's influence transcends the GPT lineage, significantly impacting the broader field of LLMs. Numerous cutting-edge language models have

adopted this architectural framework as their foundational structure. For instance, OPT (Zhang et al. , 2022a) employs the decoder-only architecture to achieve commendable natural language understanding capabilities. Gopher (Rae et al. , 2021) also leverages this unidirectional flow to escalate the complexity and scale of language modeling tasks. Moreover, the decoder-only architecture has been instrumental in the evolution of models like BLOOM (Scao et al. , 2022), which utilize its unidirectional information flow for tasks necessitating contextual comprehension. LLaMA (Touvron et al. , 2023a) and its successor, LLaMA-2 (Touvron et al. , 2023b), have integrated this architectural style to propel advancements in language modeling, achieving remarkable performances across various NLP benchmarks. GLM (Zeng et al. , 2022) further underscores the decoder-only architecture's efficacy in a range of language understanding tasks, underscoring its vital role in the contemporary landscape of language modeling.

CLIP Variants

CLIP (Radford et al. , 2021) employs an innovative approach by simultaneously training an image encoder and a text encoder to infer the correct pairings among a set of <image, text> pairs. This strategy forms the bedrock of its learning process. In contrast, FILIP (Yao et al. , 2021) enhances alignment at a finer granularity by incorporating a cross-modal late interaction mechanism. This mechanism employs token-wise maximum similarity measurements between visual and textual tokens to provide guidance for the contrastive objective, resulting in more precise alignments. FLIP (Li et al. , 2023o) introduces a groundbreaking training technique that involves randomly masking and removing a significant portion of image patches. This approach increases the number of image-text pairs that can be learned within the same wall-clock time, enabling more samples to be contrasted per iteration without significantly increasing memory usage. On the language encoder side, K-Lite (Shen et al. , 2022) suggests incorporating external knowledge in the form of Wiki definitions for entities in combination with their original alt-text for contrastive pre-training. Empirical evidence indicates that enriching text descriptions in this manner leads to improved CLIP performance. LaCLIP (Fan et al. , 2023) leverages the in-context learning ability of large language models to rewrite text descriptions for their associated images, further enhancing the model's performance by aligning descriptions more effectively with the visual content. DetCLIP, as introduced in Yao et al. (2022a), represents a pioneering approach in parallel visual-concept pre-training for open-world detection. It leverages knowledge enrichment from a meticulously crafted concept dictionary. Meanwhile, its successor, DetCLIPv2 (Yao et al. , 2023a) capitalizes on the maximum word-region similarity between region proposals and textual words to steer the contrastive objective.

Other Architectures

Traditional Transformer architectures are often limited by their quadratic computational complexity. To address this, recent research has focused on developing more efficient language modeling architectures. The S4 model (Gu et al. , 2021) offers an innovative solution by applying a low-rank correction to condition the state matrix, thus stabilizing its diagonalization and reducing the complexity of the state space

model (SSM) to operations akin to a Cauchy kernel. Similarly, GSS (Mehta et al. , 2022) emerges as a compelling alternative to the S4 and DSS (Gupta et al. , 2022) models, with the advantage of markedly faster training times. In contrast, H3 (Dao et al. , 2022) is designed to excel in specific functions like recalling earlier tokens in the sequence and comparing tokens across the sequence, further enhancing its efficiency through the integration of FlashCov. For those exploring subquadratic alternatives to attention mechanisms, Hyenra (Poli et al. , 2023) offers a notable solution. This model is crafted by combining implicitly parametrized long convolutions with data-controlled gating, significantly diminishing computational requirements. RWKV (Peng et al. , 2023a) utilizes a linear attention mechanism, allowing the model to function as either a Transformer or an RNN. This approach not only facilitates parallelized computations during training but also ensures constant computational and memory complexity during inference, marking it as the first non-transformer architecture scalable to tens of billions of parameters. RetNet (Sun et al. , 2023e) represents another significant contribution, striking an optimal balance between training parallelism, cost-effective inference, and robust performance. LongNet (Ding et al. , 2023a) introduces dilated attention, a technique that significantly widens the attention field as the distance between tokens increases, thereby enabling effective scaling of sequence length to over a billion tokens. Lastly, Streaming-LLM (Xiao et al. , 2023b) presents an efficient framework that allows Language Models (LLMs) trained with a finite-length attention window to adapt to infinite sequence lengths without additional fine-tuning. This breakthrough has extended the sequence length capability of these models to 4 million tokens.

4.2 Fine-Tuning

A fundamental strategy employed by Large Language Models (LLMs) revolves around the concept of pre-training on extensive general domain data, followed by customizing the model to suit particular tasks or domains. This approach endows LLMs with a comprehensive understanding of language patterns, enabling them to subsequently fine-tune their performance across a broad spectrum of downstream tasks, including natural language understanding, generation, and translation. The process of adaptation assumes paramount significance in achieving exceptional results in these specific tasks, as it empowers the LLM to leverage its previously acquired knowledge and apply it to new instances. The adaptation process encompasses a variety of techniques, ranging from thorough fine-tuning of the pre-trained model to the incorporation of task-specific layers or modules, as well as the utilization of transfer learning methods like knowledge distillation.

4.2.1 Data Source

Benchmark Data

A natural step in the process of data collection entails the adaptation of pre-existing NLP benchmarks. Given that these benchmarks are open-source, researchers find it both more convenient and cost-effective to utilize reasoning benchmarks to bolster

the model's reasoning capabilities. However, challenges arise concerning the availability of benchmarks in terms of quantity and scale, and the manual creation of new benchmarks proves to be a resource-intensive task. To tackle this issue, researchers are devising strategies to generate fine-tuning data for reasoning synthesis using an advanced language model.

Synthesis Data

This section delves into the synthesis of reasoning data utilizing advanced Large Language Models (LLMs) and subsequently harnesses the generated data for fine-tuning. The core of this research revolves around the application of Chain-of-Thought (CoT) techniques to LLMs, leading to the creation of reasoning paths. Subsequently, the generated data is leveraged for the model fine-tuning (Fu et al. , 2023b; Hsieh et al. , 2023; Huang et al. , 2022a; Li et al. , 2022c; Magister et al. , 2022). Additionally, the Finetune-CoT method, as introduced by Ho et al. (2022), involves the sampling of multiple reasoning paths from LLMs, which are then used for fine-tuning student models with the correct ones. The Distilling step-by-step approach, proposed by Hsieh et al. (2023), introduces a novel mechanism with two primary objectives: (a) training smaller models surpassing LLMs and (b) achieving this feat with reduced training data requirements for fine-tuning or distillation. Furthermore, the Self-Improve approach, as detailed in Huang et al. (2022a), includes the selection of rationale-augmented answers with the highest confidence for unlabeled questions using Chain-of-Thought prompting and self-consistency. Subsequently, the LLM is fine-tuned using these self-generated solutions as target outputs, with the additional step of feeding the question and ground-truth label to LLMs to prompt their reasoning path. An alternative approach involves leveraging several examples with human-written explanations as demonstrations of LLMs, followed by the generation of explanations for the training set (Li et al. , 2022c). Notably, this research provides evidence supporting the feasibility of fine-tuning a student model based on the chain of thought outputs generated by a larger teacher model, resulting in improved task performance across various types of reasoning datasets, including arithmetic, commonsense, and symbolic reasoning (Magister et al. , 2022). In the domain of mathematics, the WizardMath framework (Luo et al. , 2023c), introduces a novel method termed Reinforcement Learning from Evol-Instruct Feedback (RLEIF). This approach initially generates diverse math instruction data using math-specific Evol-Instruct. Subsequently, it involves the training of an instruction reward model (IRM) and a process-supervised reward model (PRM) (Yuan et al. , 2023a; Lightman et al. , 2023). The IRM assesses the quality of the evolved instruction, while the PRM receives feedback for each step in the solution. Furthermore, MetaMath (Yu et al. , 2023c) introduces an innovative question bootstrapping method to augment the training dataset, resulting in MetaMathQ. This method entails the rewriting of questions with both forward and backward reasoning paths and utilizes LLMs to rephrase the question text. Lastly, MAmmoTH, as introduced by Yue et al. (2023), presents a new math hybrid instruction-tuning dataset named MathInstruct. This dataset boasts two significant characteristics: extensive coverage of diverse math fields and complexity levels, as well as the incorporation of hybrid Chain-of-Thought (CoT) and Process-of-Thought (PoT) rationales. The paper "Orca" by Mukherjee

et al. (2023) introduces a method called explanation tuning. This approach involves fine-tuning a model using pairs of queries and responses. The responses are augmented with detailed explanations from GPT-4, clarifying the teacher model's reasoning process as it generates each response. In a subsequent work, "Orca2" by Mitra et al. (2023), a technique named Prompt Erasing is proposed. This method involves modifying the training process by replacing the specific instructions provided to the student system with generic ones, omitting specific details about how to execute the task.

4.2.2 Parameter-Efficient Fine-tuning

One of the fundamental paradigms in building foundation models entails thorough pre-training on general domain data, succeeded by customization for specific tasks or domains. As model sizes continue to grow, conducting comprehensive fine-tuning that alters all model parameters becomes progressively unfeasible. Hence, the importance of parameter-efficient fine-tuning in efficiently refining foundation models cannot be emphasized enough. Some representative approaches of different types of techniques are shown in Figure 16.

Adapter Tuning

Adapter tuning, a technique that employs specialized neural network modules called "adapters" within Transformer models, is discussed in Houlsby et al. (2019). An innovative adaptation method, LLaMA-Adapter (Zhang et al. , 2023h), has been developed to effectively fine-tune LLaMA models for instruction-following tasks. LLaMA-Adapter showcases its efficiency by introducing only 1.2 million learnable parameters into the pre-trained LLaMA 7B model, utilizing 52,000 self-instruct demonstrations, and completing the fine-tuning process in under an hour using 8 A100 GPUs. MAD-X (Pfeiffer et al. , 2020) is an adapter-based framework which is designed to learn modular language and task representations that can be adapted to various tasks and languages with high portability and transfer of high parameter-efficiency. On the other hand, AdaMix (Wang et al. , 2022d) fine-tunes a mixture of adaptation modules in each transformer layer while keeping the majority of PLM weights frozen. Compacter (Karimi Mahabadi et al. , 2021) integrates task-specific weight matrices into the weights of a pre-trained model, which can be efficiently obtained as a sum of Kronecker products between shared "slow" weights and "fast" rank-one matrices as defined in each Compacter layer. Lastly, He et al. (2021) introduce a unified framework that establishes connections between these approaches.

Low-Rank Adaptation

Low-Rank Adaptation (LoRA) (Hu et al. , 2022), shown in Figure 16(a), offers a distinctive approach aimed at reducing the number of trainable parameters in pre-trained Transformer models when applied to downstream tasks. This technique involves the freezing of pre-trained model weights and the introduction of trainable rank decomposition matrices into each layer of the Transformer architecture. While low-rank decomposition has limitations in terms of representation power, KronA (Edalati et al. , 2022) opts for the Kronecker product as an alternative to low-rank representation. AdaLoRA (Zhang et al. , 2023g) parametrizes incremental updates through singular

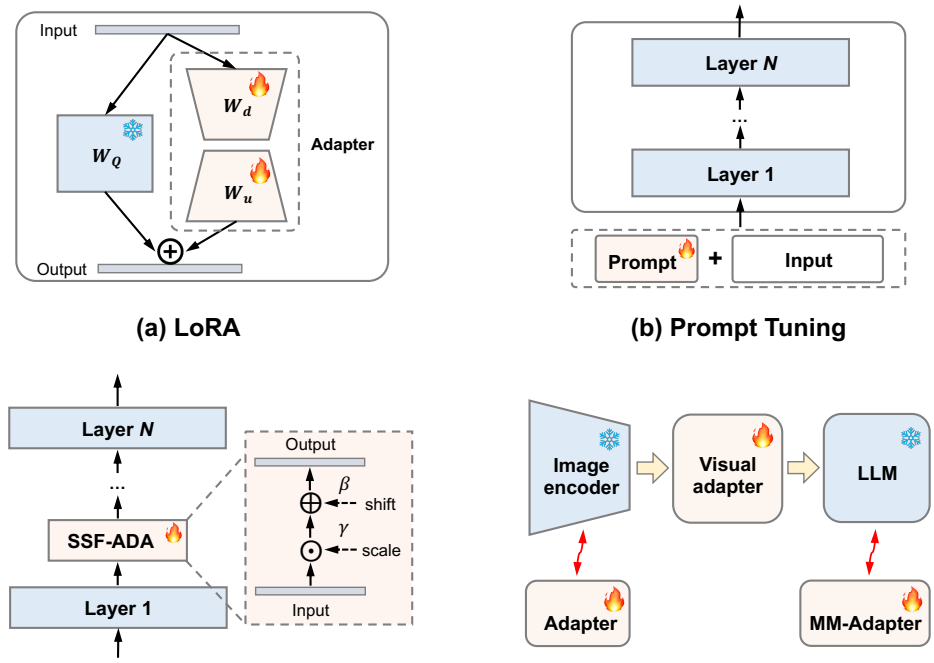


Fig. 16: Illustrations of different parameter-efficient training approaches. (a) Low-Rank Adaptation (LoRA) maintains the original weights of the pre-trained model unchanged, while integrating trainable matrices based on rank decomposition into every layer of the network for adjusting parameters. The figure credits from LoRA (Hu et al. , 2022). (b) Prompt tuning incorporates trainable prompt vectors at the input layer and uses the prompt-augmented input to tackle specific downstream issues. (c) SSF only needs to scale and shift the deep features extracted by a pre-trained network for parameter-efficient fine-tuning. The figure credits from SSF (Lian et al. , 2022). (d) MMA trains lightweight adapters to bridge the gap between large language models and vision-language tasks to enable the joint optimization of the vision and language models. The figure credits from MMA (Luo et al. , 2023b).

value decomposition, allowing for the effective pruning of unimportant singular values. DyLoRA (Valipour et al. , 2022) adopts an alternative method, focusing on training LoRA blocks across a spectrum of ranks instead of just one. This is done by organizing the representations acquired by the adapter module at different ranks throughout the training process. For those in search of an efficient fine-tuning solution, "Efficient Fine-tuning of Quantized LLMs" (QLoRA) (Dettmers et al. , 2023) provides an appealing option. QLoRA enables the fine-tuning of models with up to 65 billion parameters on a single 48GB GPU, making it a practical and accessible choice for both researchers and practitioners.

Prompt Tuning

Prefix tuning, initially introduced in Li and Liang (2021), extends Transformer-based language models by appending a sequence of trainable continuous vectors, known as "prefixes" to each layer. It lays the foundation for prompt tuning, a concept akin to "prefix tuning" (Lester et al. , 2021), with a primary focus on integrating trainable prompt vectors exclusively at the input layer. Prompt tuning represents a simple yet highly effective approach for obtaining "soft prompts" that empower fine-tuned language models to excel in specific downstream tasks, e.g., classification (Yang et al. , 2022a), which has been illustrated in Figure 16(b). In a similar context, OptiPrompt (Zhong et al. , 2021) operates within the continuous embedding space to optimize performance. On the other hand, P-tuning (Liu et al. , 2023k) leverages trainable continuous prompt embeddings alongside discrete prompts, demonstrating effectiveness across both pre-trained and fine-tuned language models, whether in fully supervised or few-shot settings. An evolution of this concept, P-tuning V2 (Liu et al. , 2021a), proposes the integration of continuous prompts into every layer of the pre-trained model, not restricting itself solely to the input layer. This extension provides a comprehensive approach to harnessing continuous prompts throughout the model's architecture.

Partial Parameter Tuning

In contrast to the aforementioned approaches that emphasize parameter efficiency, partial parameter tuning distinguishes itself by not introducing any additional components but rather by selectively fine-tuning specific parameters within the original model. Bitfit (Zaken et al. , 2021) exemplifies this concept as a method for sparse fine-tuning, concentrating solely on adjusting the bias terms of the model. Child-Tuning (Xu et al. , 2021a) adopts a strategic approach to parameter adaptation. It targets a subset of parameters known as the "child network" within large pre-trained models while carefully masking out gradients from the non-child network during the backward pass. In the case of SSF (Lian et al. , 2022), corresponding to Figure 16(c), the method introduces learnable parameters in training. These extra parameters can be seamlessly integrated into the original pre-trained model weights through re-parameterization at inference, with modifications applied to either the complete set or a subset of these parameters. DiffFit (Xie et al. , 2023b), on the other hand, presents a parameter-efficient fine-tuning strategy tailored for large pre-trained diffusion models. This method enables rapid adaptation to new domains by fine-tuning bias terms and incorporating newly introduced scaling factors into specific layers of the model. Fu et al. (2023d) theoretically analyze the parameter sparsity in fine-tuning approaches and design SAM to optimize the selection of suitable parameters.

Mixture-of-Modality Adaption

Luo et al. (2023a,b) have developed a pioneering method for fine-tuning vision-language models, termed Mixture-of-Modality Adaptation (MMA). Illustrated in Figure 16(d), MMA serves as a comprehensive optimization framework that unifies the image encoder with Large Language Models (LLMs) via efficient adapters. This work also introduces a cutting-edge routing algorithm in MMA, enabling the

model to dynamically modify its reasoning pathways for both single- and multimodal instructions. Utilizing MMA, the authors have created LaVIN (Luo et al. , 2023b), a significant vision-language instructed model that exhibits enhanced training efficiency and improved reasoning abilities across a range of instruction-following tasks. LaVIN demonstrates superior performance compared to existing multimodal LLMs. The MMA methodology and LaVIN model hold considerable potential in augmenting the utility of vision-language models, particularly in applied fields like robotics and autonomous systems. In a similar context, LLaMA-Adapter V2 (Gao et al. , 2023c) represents a visual instruction model that focuses on parameter efficiency and the seamless integration of visual information. This model incorporates several strategies to boost its performance, including expanding its learnable parameter set, adopting an early fusion approach to integrate visual tokens into the initial layers of LLMs, and applying a joint training approach for both image-text pairings and instruction-following datasets. Alternatively, LLaVA (Liu et al. , 2023e) presents itself as an integrated multimodal model that undergoes an end-to-end training process. LLaVA links a vision encoder and an LLM to process a wide spectrum of tasks involving both vision and language comprehension. LLaVA-1.5 (Liu et al. , 2023d) introduces relatively straightforward adjustments, like utilizing CLIP-ViT-L-336px with an MLP projection and integrating task-specific VQA data with basic response formatting prompts. These modifications enable LLaVA-1.5 to set a robust baseline performance, achieving top-tier results across 11 benchmark tasks.

4.3 Alignment Training

The methodology of alignment training introduces an innovative approach that employs learning techniques to optimize language models using human feedback directly. This concept has initiated a new paradigm in which language models are fine-tuned to correspond with intricate human values more closely. While Large Language Models (LMs) can be prompted to execute a variety of natural language processing (NLP) tasks based on given examples, they often manifest unintended behaviors. These include generating fictitious information, creating biased or offensive text, or failing to comply with user directives. Such discrepancies stem from the divergence between the traditional language modeling objective—predicting the next token from the web-based text—and the goal of "following user instructions in a manner that is both helpful and safe." This incongruity suggests a misalignment in the language modeling objective. Rectifying these unintentional behaviors is critically important, especially given the widespread application of language models in numerous domains.

4.3.1 Data Source

We define the data as $d_k = (i_k, y_k)$, where i_k represents the instruction and y_k denotes the corresponding response.

Human Data

Databricks has curated a comprehensive crowd-sourced instruction dataset known as "databricks-dolly-15k" (Conover et al. , 2023), containing a total of 15,000 instructions. In addition to this, the OpenAssistant corpus (Köpf et al. , 2023) consists of more than 10,000 dialogues, involving the participation of over 13,000 international annotators. UnifiedQA (Khashabi et al. , 2020) has undergone evaluation across 20 diverse datasets, covering various linguistic phenomena. CrossFit (Ye et al. , 2021) has been established as an NLP benchmark, encompassing 160 tasks converted from publicly available NLP datasets into a unified text-to-text format. P3 (Sanh et al. , 2021) has collected over 2,000 English prompts from more than 270 datasets, while MetaICL (Min et al. , 2022) has conducted experiments across 142 NLP datasets with seven different meta-training and target splits. ExMix (Aribandi et al. , 2022) offers a diverse set of 107 supervised NLP tasks. The Natural Instructions dataset (Mishra et al. , 2022) comprises 61 tasks, and Super-NaturalInstructions (Wang et al. , 2022c) expands upon it with over 1.5k tasks. Flan 2022 (Longpre et al. , 2023) combines various sources for instruction tuning, adapting templates to achieve strong evaluation performance. xP3 (Crosslingual Public Pool of Prompts) (Muennighoff et al. , 2022) is a collection of prompts and datasets spanning 46 languages and 16 NLP tasks, which aids multilingual models BLOOMZ and mT0 in zero-shot instruction-following. LongForm (Köksal et al. , 2023) selects 15,000 target text examples from the C4 and English Wikipedia corpus. Furthermore, ShareGPT, a website, actively encourages users to share their engaging ChatGPT/GPT4 conversations, resulting in a wealth of diverse, human-authored instructions capable of eliciting high-quality ChatGPT/GPT4 responses. To create non-English datasets, the Open Instruction Generalist (COIG) (Zhang et al. , 2023b) translates English instructions into Chinese and utilizes annotators to rectify and reorganize the instructions.

Synthesis Data

Gathering data from human sources can be a resource-intensive and time-consuming process. Given the remarkable success of Large Language Models (LLMs) like GPT-4, utilizing LLM responses to formulate instructions for training other LLMs in Reinforcement Learning from Human Feedback (RLHF) has become increasingly viable.

Pioneering work in this area, as demonstrated by Self-Instruct (Wang et al. , 2022b), harnesses the in-context learning capability of ChatGPT to generate a substantial volume of instructions. These instructions are drawn from a predefined set of human-annotated examples, spanning a wide range of topics and task types. Building upon this approach, Aplaca (Taori et al. , 2023) and its various iterations (Peng et al. , 2023c; Chiang et al. , 2023) employ LLMs to generate numerous training pairs for RLHF. Instruction Backtranslation (Li et al. , 2023m) leverages Self-augmentation to create responses along with instructions and utilizes Self-curation to generate instructions based on responses. Unnatural Instructions (Honovich et al. , 2022) stands out as a substantial dataset of innovative instructions, comprising 64,000 examples generated by LLMs through seed examples and rephrasing, resulting in a dataset of approximately 240,000 instances. The OPT-IML Bench (Iyer et al. , 2022) serves as a

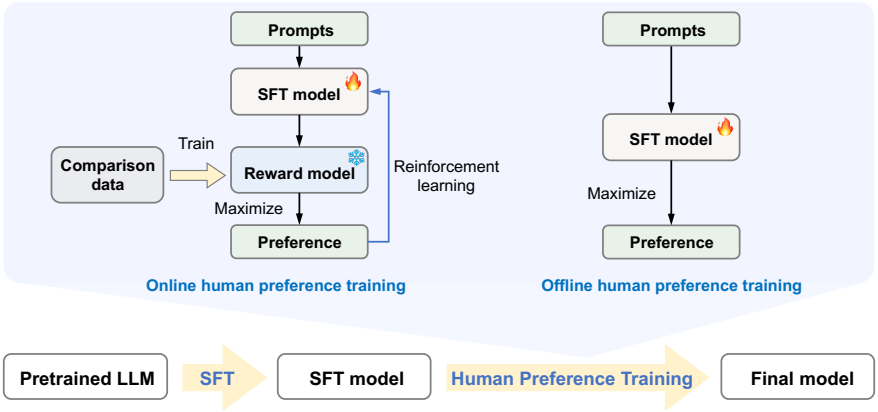


Fig. 17: The development process for large language model's (LLM's) alignment training. First, LLM is conventionally optimized via Supervised Fine-Tuning (SFT) using high-quality instruction data. Then, it may be further adjusted through Human Preference Training. The related techniques include online human preference training (left) that needs reinforcement learning and offline ones (right) that directly optimizes the policy to satisfy the preferences best.

benchmark for Instruction Meta-Learning (IML), featuring 2,000 tasks derived from eight existing benchmarks. It evaluates model generalizations using the vanilla GPT-3's Self-Instruct approach, yielding over 52,000 instructions and 82,000 instances. Koala (Geng et al. , 2023) is a small yet high-quality dataset curated from various sources, including ChatGPT Distillation Data, resulting in a comprehensive and diverse dataset. GPT4All (Anand et al. , 2023) comprises approximately one million prompt-response pairs from the GPT-3.5-Turbo OpenAI API, spanning the period from March 20, 2023, to March 26, 2023. Alpaca-GPT4 (Peng et al. , 2023c) includes 52,000 examples of instruction-following in both English and Chinese. It incorporates feedback data from GPT-4 to enhance zero-shot performance. LaMini-LM (Wu et al. , 2023d) contains a vast dataset of 2.58 million instruction-response pairs generated by the GPT-3.5-Turbo model. These pairs are drawn from various prompt sources to ensure diversity. CoEdIT (Raheja et al. , 2023) is a system that offers an 82,000 dataset of <instruction: source, target> pairs for text editing model training and evaluation. UltraChat (Ding et al. , 2023d) is an open-source collection of multi-round dialogues, including a million-scale multi-turn instructional conversation data. CoT-Collection (Kim et al. , 2023) augments Chain-of-Thought (CoT) rationales with 1.88 million instances from the FLAN Collection (Longpre et al. , 2023). Dynosaur (Yin et al. , 2023a) is a dynamic paradigm for data curation in instruction tuning, continuously expanding by incorporating new datasets from the Huggingface Datasets Platform.

4.3.2 Training Pipeline

A common method for enhancing Large Language Models (LLMs) to more accurately interpret and respond to human intentions through specific guidance is known as Supervised Fine-Tuning (SFT). This technique involves processing an instructional input, labeled as x, and then calculating the cross-entropy loss in relation to the actual correct response, denoted as y. The main role of SFT is to assist LLMs in understanding the deeper meanings within text prompts and to produce appropriate replies. However, a significant drawback of SFT is its lack of capacity to make detailed distinctions between the best and less ideal responses. Overcoming this challenge necessitates additional training strategies, such as incorporating human preference training.The overall training pipeline is presented in Figure 17.

Online Human Preference Training

Reinforcement Learning from Human Feedback (RLHF)(Ouyang et al. , 2022) represents a strategy developed to interpret human preferences by incorporating additional reward models within the framework of Proximal Policy Optimization (PPO)(Schulman et al. , 2017). RLHF is divided into three primary phases: 1) The initial stage includes the creation of a comprehensive set of guidelines and the application of Supervised Fine-Tuning (SFT) on pre-existing Large Language Models (LLMs); 2) The next phase involves human evaluators who manually grade pairs of responses, aiding in the development of a reward model that evaluates the effectiveness of the responses generated; 3) Lastly, the SFT model (policy) undergoes refinement through PPO, leveraging the rewards determined by the reward model.

While the PPO framework is known for its effectiveness in learning human preferences, it can present challenges and exhibit less stability during training. An alternative approach, Reward Ranked Fine-Tuning (RAFT) (Dong et al. , 2023a), initially involves sampling a substantial batch of instructions. Subsequently, responses are generated by the current LLMs, and the resulting data is ranked using a reward model. Only the top instances, as determined by the reward model, are then used for SFT. Additionally, Advantage-Induced Policy Alignment (APA) (Zhu et al. , 2023a) employs a squared error loss function based on estimated advantages, offering an alternative perspective on policy alignment within the RLHF framework.

Offline Human Preference Training

The implementation of those online algorithms can often be challenging due to the intricate interactions required between policy, behavior policy, reward, and value models. This complexity necessitates the adjustment of numerous hyperparameters to strengthen performance. To mitigate this problem, offline learning of human preferences has been studied.

One such approach is Direct Preference Optimization (DPO) (Rafailov et al. , 2023), which aims to implicitly optimize the same objective as existing Reinforcement Learning from Human Feedback (RLHF) algorithms. Preference Ranking Optimization (PRO) (Song et al. , 2023b) takes this further by fine-tuning Large Language Models (LLMs) to better align with human preferences and introduces Supervised

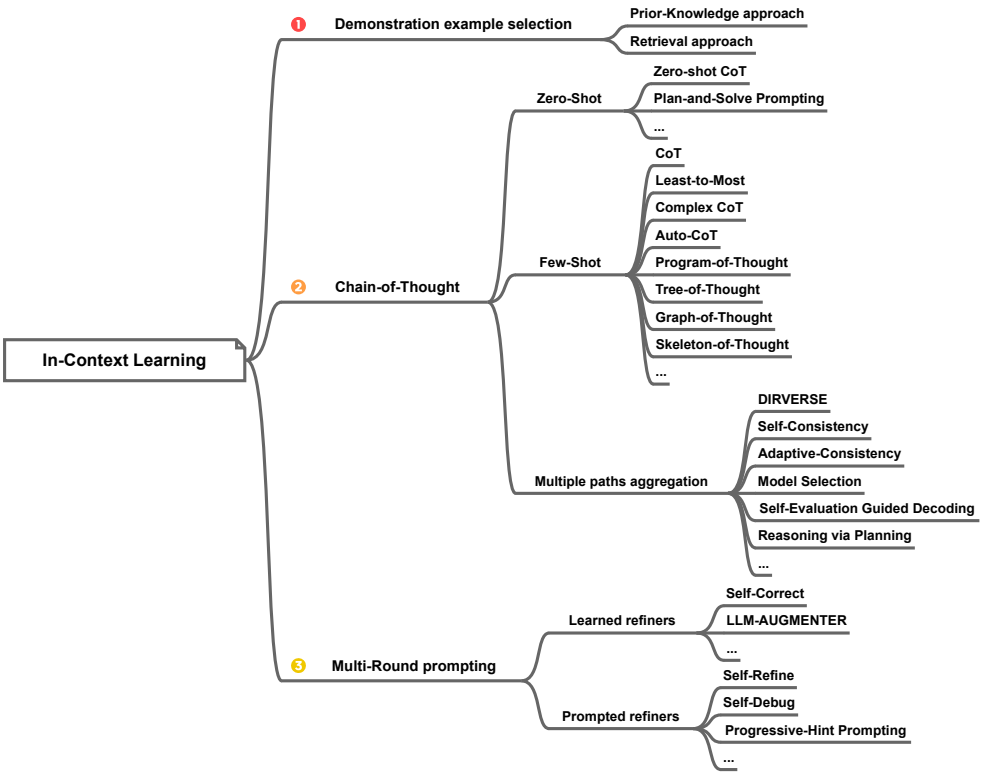


Fig. 18: Common techniques for In-Context Learning.

Fine-Tuning (SFT) training objectives for regularization. Sequence Likelihood Calibration (SLiC) (Zhao et al. , 2022a) focuses on adjusting the probability of sequences created by the model to more closely match those of reference sequences within the model's latent space. In contrast, Rank Responses to align Human Feedback (RRHF) (Yuan et al. , 2023b) aligns model probabilities of multiple responses with human preferences using ranking loss, providing a simpler yet effective alternative that retains the performance of the Proximal Policy Optimization (PPO) algorithm.

4.4 In-Context Learning

As described in Brown et al. (2020), In-Context Learning (ICL) is a method that utilizes a meticulously crafted natural language prompt encompassing both the task description and a subset of task examples to provide demonstrations. This process commences with the task description, followed by the careful selection of a few examples from the task dataset to serve as demonstrations. These chosen instances are then intricately arranged into natural language prompts using thoughtfully designed templates. Subsequently, the test instance is combined with these demonstrations as input for Language Models (LLMs) to generate the desired output. Leveraging the provided task demonstrations, LLMs can effectively discern and execute novel tasks

without the need for explicit gradient updates. It is important to note that ICL shares a fundamental connection with instruction tuning, as both methods use natural language to structure tasks or instances. Nevertheless, instruction tuning necessitates the fine-tuning of LLMs to adapt models, whereas ICL relies purely on prompting LLMs for their applications. Moreover, it's important to note that instruction tuning can improve the ICL abilities of LLMs for executing specific tasks, particularly in zero-shot situations where only task descriptions are provided (Chung et al. , 2022). A diverse set of common techniques are introduced next and listed in Figure 18.

4.4.1 Demonstration Example Selection

The effectiveness of In-Context Learning (ICL) often exhibits considerable variability based on the choice of demonstration examples. Therefore, it becomes crucial to carefully select a subset of examples that can truly harness the ICL capacity of Language Models (LLMs). Two primary methods for demonstration selection are prevalent: heuristic approaches and LLM-based approaches, as explored in the works of Liu et al. (2022a) and Lee et al. (2022).

Prior-Knowledge Approach Due to their cost-effectiveness and simplicity, heuristic techniques have been widely adopted in previous research for the selection of demonstrations. Many studies have integrated k-NN-based retrievers to identify semantically relevant examples for specific queries, as evidenced by Liu et al. (2022a) and Lee et al. (2022). However, it is important to note that these approaches typically operate on a per-example basis, lacking a holistic evaluation of the entire example set. To overcome this limitation, diversity-centric selection strategies have been introduced to curate a subset of examples that collectively represent the spectrum of specific tasks, as explored in the works of Levy et al. (2022) and Hongjin et al. (2022). Moreover, research conducted by Ye et al. (2022) takes into account both relevance and diversity in the demonstration selection process. Intriguingly, Complex CoT (Fu et al. , 2022) advocates the inclusion of intricate examples that involve extensive reasoning steps, while Auto-CoT (Zhang et al. , 2022b) suggests the sampling of a more diverse set of examples for demonstration.

Retrieval Approach Another area of research is dedicated to harnessing the capabilities of Language Models (LLMs) in selecting demonstrations. For instance, LLMs can be employed to directly assess the informativeness of each example by quantifying the performance improvement resulting from its inclusion, as demonstrated by Li and Qiu (2023a). In a related vein, Rubin et al. (2022) introduce an approach called EPR, which involves a two-stage retrieval process. Initially, EPR recalls similar examples through an unsupervised method and subsequently ranks them using a dense retriever. Building upon this, Dr.ICL (Luo et al. , 2023e) applies the EPR approach to a broader spectrum of evaluation tasks, encompassing QA, NLI, MathR, and BC. Within the context of in-context learning, Compositional Exemplars for In-context Learning (CEIL) (Ye et al. , 2023a) utilizes Determinantal Point Processes (DPPs) to learn the interaction between input and in-context examples. This model is optimized using a well-crafted contrastive learning objective. Additionally, LLM-R (Wang et al. , 2023h) adopts a ranking method for retrieved candidates, relying on

the conditional LLM log probabilities of the ground-truth outputs. It employs a cross-encoder-based reward model for capturing fine-grained ranking signals from LLMs, and a bi-encoder-based dense retriever trained through knowledge distillation. The Unified Demonstration Retriever (UDR) (Li et al. , 2023l) utilizes a shared demonstration retrieval model to overcome the issue of non-transferability among retrievers across different tasks. UDR ranks candidate examples based on LLM's feedback. With trained retrievers, DQ-LoRe (Xiong et al. , 2023a) utilize Dual Queries and Low-rank approximation Re-ranking to automatically select exemplars for in-context learning.

4.4.2 Chain-of-Thought

Zero-Shot CoT

Zero-shot CoT (Kojima et al. , 2022) introduces a novel approach to enhance model reasoning abilities by incorporating additional sentences. For instance, empirical evidence has demonstrated that including the phrase "Let's think step by step" can significantly boost the model's reasoning skills. In a similar vein, Plan-and-Solve (PS) Prompting (Wang et al. , 2023g) presents a two-fold strategy. First, it involves formulating a plan to break down the overall task into smaller, manageable subtasks. Subsequently, these subtasks are executed according to the devised plan. More precisely, PS prompting replaces the original "Let's think step by step" from Zero-shot CoT with a new prompt that encourages a more detailed approach: "Let's first understand the problem and devise a plan to solve it. Then, let's proceed to execute the plan and solve the problem step by step."

Few-Shot CoT

Chain-of-Thought (CoT) (Wei et al. , 2022b) has charted a significant course for enhancing the reasoning capabilities of Language Models (LLMs) by employing detailed reasoning paths as prompts. This directional trend has given rise to various CoT variants, such as least-to-most (Zhou et al. , 2023a), complex CoT (Fu et al. , 2022), program-of-thought (Chen et al. , 2022c), equation-of-thought (Liu et al. , 2023i), program-aid-language (Gao et al. , 2023b), mathprompter (Imani et al. , 2023), and code prompting (Hu et al. , 2023b). However, it is worth noting that all these methods require annotations, which impose practical limitations on their application. To address this constraint, Auto-CoT (Zhang et al. , 2022b) proposes a novel approach that utilizes Zero-Shot-CoT (Kojima et al. , 2022) to generate CoT reasoning paths. Furthermore, Auto-CoT divides these reasoning paths into different clusters and selects questions that are most closely aligned with the centroid of each cluster. Memory-of-Thought (Li and Qiu, 2023b) selects relevant, high-quality thoughts from external memory during the reasoning process. Taking a step further, Tree-of-Thought (Yao et al. , 2023b) models the human thought process not only as a chain but also as a tree, whereas Graph-of-Thought (Yao et al. , 2023d) extends this concept to represent human thought processes as both chains and graphs. Additionally, Skeleton-of-Thought (Ning et al. , 2023) guides LLMs to first create the basic structure of the answer and then uses batched decoding to simultaneously fill in the details of each skeleton.

Multiple Paths Aggregation

The DIVERSE approach (Li et al. , 2022g) employs a voting verifier to consolidate final answers derived from multiple reasoning paths. In a similar vein, the Self-Consistency method (Wang et al. , 2023n) suggests sampling multiple reasoning paths and making a majority vote to determine the ultimate results. Building on this direction, the concept of complexity-based voting has been introduced, retaining reasoning paths with high complexity for majority voting (Fu et al. , 2022). Furthermore, Model Selection (Zhao et al. , 2023d) takes a different approach by sampling two answers via Chain-of-Thought (CoT) and Plan-of-Thought (PoT) and then employing a Language Model (LLM) to select the correct one. Instead of generating complete reasoning paths, Self-Evaluation Guided Decoding (Xie et al. , 2023c) samples various reasoning steps at the step level and utilizes beam search to complete the search tree. One notable limitation of Self-Consistency is its relatively high cost. To mitigate this drawback, Adaptive-Consistency (Aggarwal et al. , 2023) progressively samples reasoning paths until predefined criteria are met. Two concurrent approaches related to Tree-of-Thought (Yao et al. , 2023b; Long, 2023) gradually sample reasoning steps rather than complete reasoning paths. Additionally, Reasoning via Planning (RAP) (Hao et al. , 2023a) repurposes the LLM as both a world model and a reasoning agent. It incorporates a principled planning algorithm, based on Monte Carlo Tree Search, to facilitate strategic exploration within the extensive reasoning space. Exchange-of-Thought (Yin et al. , 2023c) and X-of-Thoughts (Liu et al. , 2023i) introduce a variety of external reasoning insights and reasoning methods to enhance reasoning performance.

4.4.3 Multi-Round Prompting

Multi-round prompting enhances the response through iterative refinement, unlike single-round prompting methods such as Chain of Thought or self-consistency, which do not employ this process of progressive improvement.

Learned Refiners

The Learned Refiner necessitates a training process, and the acquisition of supervised refinement typically involves pairs of feedback and refinement (Schick et al. , 2022; Du et al. , 2022; Yasunaga and Liang, 2020; Madaan et al. , 2021). CURIOUS (Madaan et al. , 2021) initially constructs a graph that represents relevant influences. This graph is then integrated as an additional input for responding to the question. PEER (Schick et al. , 2022) is an advanced collaborative language model that replicates the entire writing process, encompassing drafting, suggesting modifications, proposing edits, and providing explanations for its actions. In contrast, Read, Revise, Repeat (R3) (Du et al. , 2022) aims to achieve superior text revisions with minimal human intervention. It achieves this by analyzing model-generated revisions and user feedback, making document revisions, and engaging in repeated human-machine interactions. DrRepair (Yasunaga and Liang, 2020) introduces a program feedback graph that connects symbols relevant to repairing source code with diagnostic feedback. It then employs a graph neural network to model the reasoning process. Self-Correction (Welleck et al. , 2022) takes an innovative approach by decoupling an imperfect base generator, such

as a standard language model or supervised sequence-to-sequence model, from a separate corrector. This corrector learns to refine outputs progressively. Furthermore, LLM-Augmenter (Peng et al. , 2023b) continuously enhances LLM prompts to improve model responses by incorporating feedback generated by utility functions, such as the factuality score of an LLM-generated response.

Prompted Refiners

The REFINER framework (Paul et al. , 2023) is a comprehensive system designed to fine-tune Language Models (LMs) with the specific goal of generating intermediate reasoning steps, aided by a critic model that automates feedback on the reasoning process. The Self-Refine framework (Madaan et al. , 2023) comprises two vital components: it first generates an output using an LLM and then employs the same LLM to provide feedback on its output through an iterative self-refinement process. Self-Debugging (Chen et al. , 2023h) integrates both LLM and tool feedback to enhance performance. Similarly, Progressive-Hint Prompting (PHP) (Zheng et al. , 2023a) utilizes previous answers as references for generating subsequent responses. Furthermore, employing distinct prompts for LLMs allows for the assignment of different roles in handling various aspects (Dong et al. , 2023b; Fu et al. , 2023a; Du et al. , 2023). Du et al. (2023) introduce a complementary approach to enhance language responses, involving multiple instances of language models engaging in discussions about their individual responses and reasoning processes over multiple rounds to arrive at a shared final answer. Self-collaboration (Dong et al. , 2023b) utilizes multiple LLMs as distinct "experts", each responsible for specific subtasks within complex assignments, defining strategies for collaboration and interaction. Fu et al. (2023a) observe that only a subset of the considered language models demonstrate proficiency in self-improvement through AI feedback, as weaker models may struggle with understanding game rules or incorporating AI feedback for further enhancements. In conclusion, models exhibit varying abilities to learn from feedback based on their roles, and interactions between LLMs and tools can further enhance reasoning performance (Chen et al. , 2023h; Gou et al. , 2023; Zhang et al. , 2023e; Yang et al. , 2023b; Olausson et al. , 2023).

4.5 Autonomous Agent

Agents that operate autonomously have often been considered a key route to achieving Artificial General Intelligence (AGI). These agents are adept at performing tasks by independently formulating plans and following instructions. At present, these autonomous entities primarily rely on Large Language Models (LLM) to control and orchestrate various tools (Xi et al. , 2023; Wang et al. , 2023s), including web browsers and code interpreters, to complete their designated tasks, as shown in Figure 19.

VISPROG (Gupta and Kembhavi, 2022) is a neuro-symbolic approach for complex visual tasks, using large language models to generate Python-like modular programs without task-specific training. It provides a comprehensive and interpretable rationale. ToolFormer (Schick et al. , 2023) is a self-supervised model that decides which APIs to call, when, and with what arguments to incorporate the results into token prediction based on demonstrations. ART (Paranjape et al. , 2023) introduces a framework for automatic reasoning and tool use, utilizing frozen LLMs to generate intermediate

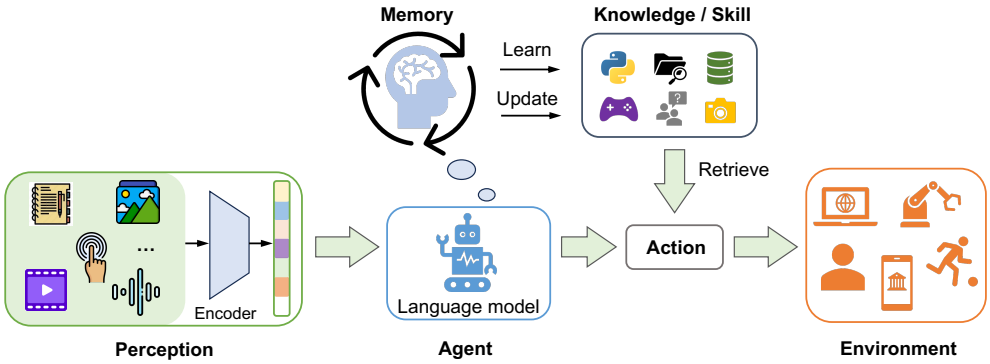


Fig. 19: The general pipeline for LLM with autonomous agent. An LLM agent leverages an LLM as its digital brain mastering multiple abilities and possessing a high-level of intelligence. The agent can receive a diverse set of encoded data as input and correspondingly construct or have access to knowledge bases and skill libraries. With sufficient knowledge and prompts, the agent can work semi-autonomously to operate a spectrum of tasks.

reasoning steps and seamlessly integrating external tools. CAMEL (Li et al. , 2023b) presents a pioneering communicative agent framework known as "role-playing", which employs inception prompting to direct chat agents towards achieving tasks while upholding alignment with human intentions. HuggingGPT (Shen et al. , 2023) connects AI models for solving tasks, using ChatGPT for task planning and selecting models based on function descriptions in Hugging Face. Chameleon (Lu et al. , 2023) augments LLMs with plug-and-play modules for complex reasoning, synthesizing programs by composing various tools for tasks. Wang et al. (2023k) propose to learn "planning tokens", a soft prompt. TRICE (Qiao et al. , 2023) addresses the challenge of teaching language models when and how to use tools, proposing a two-stage framework for learning through feedback from tool execution. ChatCoT (Chen et al. , 2023i) presents a tool-augmented chain-of-thought reasoning framework for chat-based LLMs, using multi-turn conversations to integrate thought chains and tool usage naturally. MultiTool-CoT (Inaba et al. , 2023) leverages chain-of-thought prompting to incorporate multiple external tools during the reasoning process. AssistGPT (Gao et al. , 2023a) introduces a multimodal AI assistant with an interleaved code and language reasoning approach, including planning, execution, inspection, and autonomous learning. OpenAGI (Ge et al. , 2023) is an open-source AGI research platform for real-world tasks, using natural language queries to select and execute appropriate models and proposing a Reinforcement Learning from Task Feedback mechanism. ToolkenGPT (Hao et al. , 2023b) combines the benefits of finetuning LLMs with tool demonstration data and in-context learning, representing tools as tokens ("toolkens") for flexible tool calls. AutoGPT (gravitas/auto gpt, 2023) decomposes problems into subproblems and employs tools to solve them. ReAct (Yao et al. , 2023c) explores the interleaved generation of reasoning traces and task-specific actions for enhanced

synergy in language tasks, improving interpretability and trustworthiness. Reflexion (Shinn et al. , 2023) reinforces language agents through linguistic feedback and episodic memory, improving decision-making in subsequent trials. CREATOR (Qian et al. , 2023a) endows LLMs to create their own tools through documentation and code realization, addressing limitations in tool-using ability. Voyager (Wang et al. , 2023b) is an LLM-powered agent in Minecraft for lifelong learning, incorporating automatic curriculum, skill library, and iterative prompting mechanisms. AutoAgents (Chen et al. , 2023b) can adaptively generate specialized agents to build a team of agents based on task definitions. SwiftSage (Lin et al. , 2023) is an agent framework inspired by human cognition's dual-process theory, integrating behavior cloning and LLMs for complex reasoning tasks, enhancing problem-solving efficiency. These references cover a wide range of approaches and frameworks for enhancing the capabilities of large language models across various domains.

5 Discussion: Challenges, Limitations, and Risks

Foundation models have shown promising capabilities in reasoning tasks, opening up new possibilities for the field. It is also essential to acknowledge the challenges, limitations, and risks associated with their use.

Hallucinations

Despite the promising progress made in foundation models, it is important to acknowledge that these models still face challenges, specifically in relation to the issue of hallucinations (Li et al. , 2023n; Mündler et al. , 2023). Hallucination refers to the generation of outputs by foundation models that contain fabricated or incorrect information, deviating from the intended or expected outputs. These hallucinations can be problematic, as they undermine the reliability and accuracy of the model's generated content.

The hallucination problem in foundation models arises due to various factors. One key factor is the reliance on large-scale pre-training data, which can contain biased or erroneous information. This can lead to the model learning and propagating false patterns or generating unrealistic outputs. Another significant factor contributing to the hallucination issue in foundation models is the models' lack of ability to acknowledge their own knowledge limitations. When confronted with questions beyond their understanding, these models tend to fabricate seemingly plausible answers instead of admitting their lack of knowledge (Yin et al. , 2023d).

Addressing the hallucination problem in foundation models is an ongoing area of research. Techniques such as fine-tuning task-specific data, incorporating external knowledge sources, and developing advanced evaluation metrics have been explored to mitigate hallucinations. Researchers are also exploring methods to enhance reasoning capabilities in foundation models, enabling them to make more informed and accurate predictions.

It is worth noting that while progress has been made in reducing hallucinations, completely eliminating them remains a challenge due to the inherent complexity of language understanding and generation.

Context Length

Another limitation is to optimize context length and context construction. For example, GPT models start with 2K window size (GPT-3 (Brown et al. , 2020)) and go all the way to 32K (GPT-4 (OpenAI, 2023a)). A longer context window is useful for working with long sequence data, such as gene sequences. By having a larger context window, LLM is capable of handling more lengthy inputs such as entire documents, or comprehending the full scope of an article. This ability enables LLM to produce more contextually relevant responses by leveraging a more comprehensive understanding of the input.

Increasing the context window size in foundation models can bring several benefits, such as capturing longer-range dependencies and improving the model's understanding of context. However, it also comes with certain challenges and costs. In earlier studies, it was observed that the costs associated with larger context window sizes exhibited a quadratic increase as the number of tokens grew (Aryan et al. , 2023). This means that the computational resources required to process and train the models become significantly higher as the window size grows. LongNet (Ding et al. , 2023a) represents a modified version of the Transformer model, capable of handling sequences exceeding 1 billion tokens in length, while still maintaining its effectiveness on shorter sequences. LongNet also has a linear computation complexity. Position Interpolation (Chen et al. , 2023e) implements a linear downscaling of input position indices to align with the initial context window size during inference. This approach prevents extending beyond the context length trained for, which might otherwise result in abnormally high attention scores and interfere with the self-attention mechanism.

Indeed, while increasing the context window size in language models offers benefits, it is important to consider the tradeoff between window size and generalization ability. Researchers have highlighted that there can be a tradeoff between them (Liu et al. , 2023g). One challenge worth exploring is how to increase the context window length without sacrificing the model's performance and generalization capabilities. It is crucial to find strategies that allow models to capture longer-range dependencies and context while maintaining their ability to generalize well to new or unseen inputs.

Multimodal Learning

Multimodal learning is an incredibly powerful but often underappreciated aspect of reasoning. It finds applications in numerous fields where multimodal data is essential, including healthcare (such as CT, X-ray, MRI scans, and gene sequences), robotics, e-commerce, retail, gaming, and entertainment. The integration of different modalities in these domains enables a more comprehensive understanding of the data and facilitates more sophisticated reasoning processes.

One of the key advantages of multimodal reasoning is its potential to significantly improve model performance. While some prior works have delved into multimodal reasoning, such as the multimodal language Model for embodied reasoning proposed

by PaLM-E (Driess et al. , 2023) and the visual language model for fear-shot learning known as Flamingo (Alayrac et al. , 2022), there is still ample room for exploring additional data modalities. Incorporating modalities like video, audio, 3D data, and multiple images not only enriches the information available to the models but also opens up exciting possibilities for a more nuanced and comprehensive understanding of the world. Other potential applications of foundation model reasoning lie in the domain of Electronic Design Automation (EDA) for program design (Huang et al. , 2021a) and Formal Methods (Woodcock et al. , 2009).

Formal methods, intrinsically linked to logical reasoning, are mathematical strategies employed in the realm of computer science for the design, specification, verification, and analysis of both software and hardware. These techniques are anchored in structured logic, the theory of automata, and other comprehensive mathematical frameworks. They are used to meticulously examine the behavior, accuracy, and dependability of systems. The utilization of formal methods empowers researchers and professionals to guarantee the integrity and precision of intricate systems, establishing them as indispensable in the creation and assessment of software and hardware. The amalgamation of formal methods with foundational models opens doors to augmenting reasoning abilities in the design of software and hardware systems. Formal methods bring to the table precise mathematical methods for defining and confirming system characteristics, whereas foundational models contribute potent language comprehension and reasoning skills. The synthesis of these methodologies can foster the development of more dependable and resilient software and hardware systems.

By leveraging multimodality reasoning and further expanding the exploration of various data modalities, we can unlock new insights and capabilities in reasoning systems. It is crucial to recognize and harness the power of multimodal reasoning to fully exploit the potential of reasoning in diverse domains.

Efficiency and Cost

Efficiency and cost are significant challenges for foundation models for reasoning. Foundation models, especially those with large architectures and extensive training data, can be computationally expensive to train and deploy. The large number of parameters requires more memory and computational resources for processing. This poses challenges in terms of scalability, accessibility, and cost-effectiveness. Efficient reasoning models should be capable of performing fast and real-time inference to meet the demands of interactive applications. However, the complex computations involved in reasoning tasks can lead to slower inference times, hindering real-time performance and user experience. Therefore, it is crucial to enhance the speed and cost-effectiveness of foundation models, making them cheaper and faster.

There are several techniques that can be employed to improve the efficiency of foundation models, including:

- Model Pruning (Sun et al. , 2023d; Wang et al. , 2020): removing unnecessary connections, parameters, or layers from the model. This results in a more compact architecture, reducing computational requirements.

- Compression (Zhu et al. , 2023c) and Quantization (Tao et al. , 2022): reduce the size of the model or reduce the precision of model parameters, using fewer bits to represent them. This reduces memory usage and computational complexity.
- Knowledge Distillation (Gu et al. , 2023b): training a smaller model (student) to mimic the behavior and predictions of a larger model or ensemble of models (teacher). This transfer of knowledge allows for efficient inference with reduced computational resources.
- Low-Rank Factorization (Ren and Zhu, 2023; Hsu et al. , 2022): replace high-dimensional tensors with lower-dimensional tensors. By reducing the number of parameters, these methods improve efficiency without significant loss in performance.

By employing these techniques, it is possible to enhance the efficiency of foundation models, making them faster and more cost-effective for various reasoning tasks and applications.

Human Preference

Addressing the risks and potential harms associated with foundation models, such as bias, unfairness, manipulation, and misinformation, requires careful consideration and proactive measures. One approach is to focus on improving learning from human preference and feedback to ensure more responsible and accurate model behavior.

To mitigate these risks, we can explore several strategies. First, we need mechanisms to incorporate diverse perspectives and mitigate bias during the training and fine-tuning phases of foundation models. This can involve diverse data collection, representative sampling, and careful annotation processes that involve input from a wide range of human perspectives. Continual learning and adaptation, informed by human feedback, can also play a crucial role. By enabling ongoing interactions between models and human annotators or users, we can gather feedback and iteratively refine the models' behavior. This iterative process helps identify and rectify potential biases, unfairness, or misinformation, allowing the models to improve over time. Furthermore, ensuring that the outputs of foundation models align with real-world evidence, experimental findings, and explicit knowledge is essential. This requires incorporating robust fact-checking mechanisms and validation processes into the model training pipeline. Additionally, leveraging external sources of information, such as trusted databases or expert knowledge, can help verify and validate the outputs generated by the models.

Constitutional AI, as proposed by Bai et al. (2022), offers an approach that involves both supervised learning and reinforcement learning phases like "RL from AI Feedback" (RLAIF). Similarly, Bakker et al. (2022) explore the use of fine-tuning a large language model (LLM) with 70 billion parameters to generate statements that maximize the expected approval for people with different and diverse perspectives. This approach emphasizes the importance of incorporating human preferences and diverse viewpoints during the model training process.

By integrating these techniques and approaches, we can work towards mitigating the risks and potential harms associated with foundation models. Improving learning from human preference, continual learning informed by feedback, and ensuring

fidelity to real-world evidence are challenging steps in building more responsible and trustworthy reasoning systems.

Multilingual Support

While reasoning itself is a language-agnostic process, the availability of comprehensive knowledge sources is often limited to a few languages, primarily English. Historically, language foundation models have demonstrated exceptional reasoning performance, primarily in English, with relatively limited support for other languages such as Chinese and Japanese. Currently, there is a lack of robust multilingual reasoning language foundation models that excel across various languages.

Fang et al. (2022) propose utilizing English as a pivot language in their commonsense reasoning framework. They employ a translate-retrieve-translate (TRT) strategy, leveraging English knowledge sources to enhance their reasoning capabilities. Furthermore, Huang et al. (2023a) introduce cross-lingual thought prompting (XLT) as a systematic approach to improving the multilingual capabilities of Language and Reasoning Models (LLMs).

Given these advancements, there is a growing interest in developing foundation models dedicated to multilingual reasoning. Building robust models that excel in multilingualism presents an intriguing avenue for future research and development.

In summary, to address these challenges, ongoing research and development efforts are necessary. This includes advancing the deployment of reasoning models.

6 Future Directions

Further research and development in this area have the potential to unlock even more advanced reasoning abilities in foundation models.

6.1 Safety and Privacy

The rise of foundation models and their application to reasoning tasks has highlighted the critical need to ensure their safety and trustworthiness (Huang et al. , 2023e).

Various intended attacks have been identified, including the robustness gap (Shreya and Khapra, 2022), backdoor attacks (Shen et al. , 2021b; Kurita et al. , 2020), poisoning (Carlini et al. , 2023), disinformation (Nelson et al. , 2008), privacy leakage (Li et al. , 2023c), and unauthorized disclosure of information (Perez and Ribeiro, 2022). Specifically, backdoor attacks involve the injection of malicious knowledge into foundation models through techniques such as poisoning the training data (Shen et al. , 2021b) or modifying model parameters (Kurita et al. , 2020).

As one of the most principled techniques for training machine learning models with privacy, differential privacy allows for training on datasets without revealing any details of individual training examples, providing enhanced privacy protection (Shi et al. , 2022; Behnia et al. , 2022). Another effective way of defensing adversarial attacks is by adversarial training, which can provide another layer of security when facing malicious yet human invisible perturbations added in model inputs (Li et al. , 2023h; Li and Spratling, 2023).

In response to some copyright concerns, Kirchenbauer et al. (2023) introduce a watermarking framework specifically designed for proprietary language models. This framework enables the embedding of watermarks with minimal impact on text quality and facilitates their detection using an efficient open-source algorithm, eliminating the need for accessing the language model API or parameters.

6.2 Interpretability and Transparency

Additionally, there is a need for increased transparency and interpretability of foundation models (Liao and Vaughan, 2023). As these models become more complex and sophisticated, understanding their reasoning processes and the factors influencing their outputs becomes increasingly important.

Sometimes, foundation models generate toxic content, which may incite violence and cause infordemic (Bender et al. , 2021; Weidinger et al. , 2021). They can inadvertently disclose sensitive information, thereby jeopardizing privacy and security. Additionally, LLMs can contribute to the dissemination of misinformation, both intentionally and unintentionally (Pan et al. , 2023b; Buchanan et al. , 2021; Kreps et al. , 2022; Zhou et al. , 2023c). The complex and uncertain nature of foundation models further compounds these challenges. These models exhibit a remarkable capacity to perform a wide range of tasks across diverse contexts (Bommasani et al. , 2021). However, their massive and opaque architectures hinder a comprehensive understanding of their capabilities and behaviors, making it difficult to ascertain their decision-making processes and potential biases. This lack of transparency raises concerns regarding model interpretability, control, and accountability.

Developing techniques and frameworks for model interpretability can help address concerns regarding transparency and accountability.

6.3 Autonomous Language Agents

The capacity for logical reasoning is crucial in achieving complex tasks in embodied environments, and it plays a significant role in embodied intelligence (Dasgupta et al. , 2022). Foundation Models have exhibited powerful capabilities for reasoning and flexibility through the process of in-context learning (Yang et al. , 2023d). Recent studies, such as Voyager and DEPS, have explored the use of LLMs for planning in Minecraft (Wang et al. , 2023b,r). DEPS specifically proposes an interactive planning approach based on LLMs (Wang et al. , 2023r). LLMs have also shown promise in generating action sequences directly based on natural language instructions without requiring extra domain knowledge (Li et al. , 2022d). Equipping embodied agents with commonsense knowledge is crucial for their successful completion of complex human instructions in diverse environments (Wu et al. , 2023g).

In the context of reasoning for autonomous agents, there are key characteristics:

- Infinite Task Capability: Foundation models empower agents with the capacity to handle an extensive range of tasks, even those that are not pre-defined or anticipated in advance. This flexibility allows agents to dynamically generate tasks based on their understanding of the context and the specific needs of the users.

- Autonomous Task Generation: Foundation model reasoning enables agents to autonomously generate new tasks within a given context. This capability empowers agents to take initiative, identify opportunities, and propose relevant tasks to users. They can adapt and respond to changing circumstances, making them more versatile, proactive, and efficient in fulfilling user requirements.
- Value System: Autonomous agents are driven by a value system empowered by a pre-trained foundation model, which serves as the foundation for task generation. This value system guides the agent's decision-making process, taking into account factors such as priorities, preferences, and ethical considerations. By leveraging the capabilities of the foundation model, agents can make informed decisions aligned with human values, ensuring responsible and ethical behavior.
- World Model: The foundation model can also be utilized as a world model that represents the real world and serves as the basis for agents' interactions and reasoning. This comprehensive model enables agents to understand the context, interpret natural language inputs, and generate appropriate responses or actions. With the foundation model as their world model, agents can effectively navigate and operate within their environment, enhancing their ability to interact intelligently and respond to user needs.

By leveraging foundation models, autonomous agents can facilitate more meaningful and effective interactions with users, better understand their intents and needs, and generate relevant tasks accordingly. This approach opens up promising avenues for research in areas such as contextual understanding, human-like reasoning, and personalized assistance. Ultimately, it enhances the overall user experience and enables the development of more sophisticated and intelligent AI systems.

Given their reasoning capabilities, Foundation Models hold significant potential for applications in human-computer interaction and embodied intelligence. can be leveraged to create interactive and adaptive systems that can dynamically respond to user input and adapt their behavior accordingly. This involves developing models that can learn from user interactions and update their knowledge and behavior over time. By enabling Foundation Models to actively engage with users and adapt to their preferences and needs, we can create more personalized and user-centric human-computer interaction experiences.

6.4 Reasoning for Science

Future work can also build upon the research on temporal reasoning in multimodal question-answering tasks or sound reasoning (Brandt and McClure, 2011), as demonstrated by Audio Question Answering (AQA) (Fayek and Johnson, 2019). Researchers can delve deeper into understanding and developing foundation models that can reason and make inferences based on auditory information. This can have implications in areas such as audio-based decision-making systems, environmental monitoring, and audio scene understanding.

Furthermore, the application of multimodal reasoning can be extended to domains like medical reasoning and diagnosis, particularly in the context of gene sequence analysis. This can aid in the identification of genetic disorders, personalized medicine, and the exploration of potential treatments.

Overall, future work can focus on advancing multimodal reasoning abilities in foundation models. These endeavors can contribute to the development of more intelligent and context-aware systems in various fields.

6.5 Super Alignment

Superintelligence alignment, according to OpenAI[§], is the next machine-learning question of utmost importance. However, ensuring the control and alignment of potentially superintelligent AI systems poses significant challenges. Current techniques, such as Reinforcement Learning from Human Feedback (RLHF), heavily rely on human supervision and reasoning. As AI systems surpass human intelligence, human supervision becomes inadequate, necessitating new scientific and technical breakthroughs in alignment research. Existing alignment techniques will not scale to superintelligence due to the limitations of human reasoning and supervision. The prospect of controlling and steering highly intelligent AI systems to prevent them from going rogue remains an unsolved challenge. Without reliable means of supervising these reasoning systems surpassing human capabilities, ensuring their alignment with human intent becomes increasingly difficult.

One approach to address the challenge of ensuring that reasoning systems surpassing human intelligence adhere to human intent is to develop a roughly human-level automated alignment researcher. By creating such a system, it becomes possible to leverage extensive computational resources to scale alignment efforts and iteratively align superintelligence.

7 Conclusion

This survey illuminates the evolutionary path of foundation models in the field of reasoning, showcasing a discernible progression in complexity and efficacy from their initial stages to current advancements. While we acknowledge the remarkable strides made in data-driven thinking, it is crucial for us to objectively recognize both the strengths and limitations of large models. Emphasizing the importance of enhancing their interpretability and security becomes imperative in this context. We also note that with all the papers surveyed in this work, a consensus is yet to be reached on how to push forward the reasoning ability of foundation models to a consistently superhuman level (which can for instance win an IMO medal or even solve open mathematical problems).

In conclusion, while foundation models offer exciting possibilities in reasoning tasks, it is essential to approach their development and application with a critical perspective. It is crucial to acknowledge the challenges, limitations, and risks associated with LLM-based reasoning. By doing so, we can foster responsible and thoughtful advancements in this field, ensuring the development of robust and reliable reasoning systems.

[§] https://openai.com/blog/introducing-superalignment

References

(2022) Thor: Wielding hammers to integrate language models and automated theorem provers. Advances in Neural Information Processing Systems 35:8360–8373

Abdine H, Chatzianastasis M, Bouyioukos C, et al. (2023) Prot2text: Multimodal protein's function generation with gnns and transformers. arXiv preprint arXiv:230714367

Acay DL, Pasquier P, Sonenberg L (2007) Extrospection: Agents reasoning about the environment. 3rd IET International Conference on Intelligent Environments

Acquaviva S, Pu Y, Nye M, et al. (2021) Larc: Language annotated abstraction and reasoning corpus. In: Proceedings of the Annual Meeting of the Cognitive Science Society

Aggarwal P, Madaan A, Yang Y, et al. (2023) Let's sample step by step: Adaptive-consistency for efficient reasoning with llms. 2305.11860

Ahmad WU, Chakraborty S, Ray B, et al. (2021) Unified pre-training for program understanding and generation. arXiv preprint arXiv:210306333

Ahn M, Brohan A, Brown N, et al. (2022) Do as i can, not as i say: Grounding language in robotic affordances. In: Conference on Robot Learning

Ai Q, Bai T, Cao Z, et al. (2023) Information retrieval meets large language models: A strategic report from chinese ir community. 2307.09751

Al-Ajlan A (2015) The comparison between forward and backward chaining. International Journal of Machine Learning and Computing 5(2):106

Alayrac JB, Donahue J, Luc P, et al. (2022) Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems 35:23716–23736

Alibali MW, Boncoddo R, Hostetter AB (2014) Gesture in reasoning: An embodied perspective. In: The Routledge handbook of embodied cognition. Routledge, p 150–159

Allal LB, Li R, Kocetkov D, et al. (2023) Santacoder: don't reach for the stars! arXiv preprint arXiv:230103988

Alvin C, Gulwani S, Majumdar R, et al. (2017) Synthesis of solutions for shaded area geometry problems. In: The Thirtieth International Flairs Conference

Amini A, Gabriel S, Lin P, et al. (2019) Mathqa: Towards interpretable math word problem solving with operation-based formalisms. arXiv preprint arXiv:190513319

Anand Y, Nussbaum Z, Duderstadt B, et al. (2023) Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all

Anil R, Dai AM, Firat O, et al. (2023) Palm 2 technical report. 2305.10403

Anthropic (2023) Introducing claude

Araci D (2019) Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:190810063

Ardila R, Branson M, Davis K, et al. (2020) Common Voice: A Massively-Multilingual Speech Corpus. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp 4218–4222

Aribandi V, Tay Y, Schuster T, et al. (2022) Ext5: Towards extreme multi-task scaling for transfer learning. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=Vzh1BFUCiIX

Aroca-Ouellette S, Paik C, Roncone A, et al. (2021) Prost: Physical reasoning of objects through space and time. 2106.03634

Aryan A, Nain AK, McMahon A, et al. (2023) The costly dilemma: Generalization, evaluation and cost-optimal deployment of large language models. 2308.08061

Austin J, Odena A, Nye M, et al. (2021) Program synthesis with large language models. arXiv preprint arXiv:210807732

Azerbayev Z, Schoelkopf H, Paster K, et al. (2023) Llemma: An open language model for mathematics. 2310.10631

Babu A, Wang C, Tjandra A, et al. (2022) XLS-R: Self-supervised cross-lingual speech representation learning at scale. In: INTERSPEECH, pp 2278–2282

Baevski A, Zhou Y, Mohamed A, et al. (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems 33:12449–12460

Baevski A, Hsu WN, Xu Q, et al. (2022) Data2vec: A general framework for self-supervised learning in speech, vision and language. In: International Conference on Machine Learning, PMLR, pp 1298–1312

Bai Y, Kadavath S, Kundu S, et al. (2022) Constitutional ai: Harmlessness from ai feedback. 2212.08073

Bakhtin A, van der Maaten L, Johnson J, et al. (2019) Phyre: A new benchmark for physical reasoning. Advances in Neural Information Processing Systems 32

Bakker MA, Chadwick MJ, Sheahan HR, et al. (2022) Fine-tuning language models to find agreement among humans with diverse preferences. 2211.15006

Balashankar A, Subramanian L (2021) Learning faithful representations of causal graphs. 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). Association for Computational Linguistics, Online, pp 839–850, https://doi.org/10.18653/v1/2021.acl-long.69, URL https://aclanthology.org/2021.acl-long.69

Ban T, Chen L, Wang X, et al. (2023) From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. arXiv preprint arXiv:230616902

Bao F, Nie S, Xue K, et al. (2023) One transformer fits all distributions in multi-modal diffusion at scale. 2303.06555

Bao H, Dong L, Piao S, et al. (2021) Beit: Bert pre-training of image transformers. arXiv preprint arXiv:210608254

Barbieri F, Anke LE, Camacho-Collados J (2021) Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. arXiv preprint arXiv:210412250

Barras B, Boutin S, Cornes C, et al. (1997) The coq proof assistant reference manual: Version 6.1. PhD thesis, Inria

Bavarian M, Jun H, Tezak N, et al. (2022) Efficient training of language models to fill in the middle. arXiv preprint arXiv:220714255

Bear DM, Wang E, Mrowca D, et al. (2021) Physion: Evaluating physical prediction from vision in humans and machines. arXiv preprint arXiv:210608261

Behnia R, Ebrahimi MR, Pacheco J, et al. (2022) Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, pp 560–566

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '21, p 610–623, https://doi.org/10.1145/3442188.3445922, URL https://doi.org/10.1145/3442188.3445922

Bengio Y (2017) The consciousness prior. arXiv preprint arXiv:170908568

Berant J, Chou A, Frostig R, et al. (2013) Semantic parsing on freebase from question-answer pairs. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1533–1544

Berghofer S, Strecker M (2004) Extracting a formally verified, fully executable compiler from a proof assistant. Electronic Notes in Theoretical Computer Science 82(2):377–394

Berglund L, Tong M, Kaufmann M, et al. (2023) The reversal curse: Llms trained on" a is b" fail to learn" b is a". arXiv preprint arXiv:230912288

Berzonsky MD (1978) Formal reasoning in adolescence: An alternative view. Adolescence 13(50):279

Bhagavatula C, Bras RL, Malaviya C, et al. (2019) Abductive commonsense reasoning. arXiv preprint arXiv:190805739

Bi K, Xie L, Zhang H, et al. (2023) Accurate medium-range global weather forecasting with 3d neural networks. Nature pp 1–6

Bisk Y, Zellers R, Gao J, et al. (2020) Piqa: Reasoning about physical commonsense in natural language. In: Proceedings of the AAAI conference on artificial intelligence, pp 7432–7439

Black S, Gao L, Wang P, et al. (2021) GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. https://doi.org/10.5281/zenodo.5297715, URL https://doi.org/10.5281/zenodo.5297715, If you use this software, please cite it using these metadata.

Black S, Biderman S, Hallahan E, et al. (2022) GPT-NeoX-20B: An open-source autoregressive language model. In: Fan A, Ilic S, Wolf T, et al. (eds) Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models. Association for Computational Linguistics, virtual+Dublin, pp 95–136, https://doi.org/10.18653/v1/2022.bigscience-1.9, URL https://aclanthology.org/2022.bigscience-1.9

Bommasani R, Hudson DA, Adeli E, et al. (2021) On the opportunities and risks of foundation models. 2108.07258

Boratko M, Li XL, Das R, et al. (2020) Protoqa: A question answering dataset for prototypical common-sense reasoning. arXiv preprint arXiv:200500771

Bottou L, Peters J, Quiñonero-Candela J, et al. (2013) Counterfactual reasoning and learning systems: The example of computational advertising. Journal of Machine Learning Research 14(11)

Brandt A, McClure R (2011) Sound reasoning

Brewka G (2012) Default Reasoning, Springer US, Boston, MA, pp 915–917. https://doi.org/10.1007/978-1-4419-1428-6_634, URL https://doi.org/10.1007/978-1-4419-1428-6_634

Brohan A, Brown N, Carbajal J, et al. (2022) Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:221206817

Brohan A, Brown N, Carbajal J, et al. (2023) Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: TODO

Brown T, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. Advances in neural information processing systems 33:1877–1901

Bubeck S, Chandrasekaran V, Eldan R, et al. (2023) Sparks of artificial general intelligence: Early experiments with gpt-4. 2303.12712

Buchanan B, Lohn A, Musser M, et al. (2021) Truth, lies, and automation: How language models could change disinformation. URL: https://cset georgetown edu/publication/truth-lies-and-automation/(visited on 10/13/2021)

Bui ND, Le H, Wang Y, et al. (2023) Codetf: One-stop transformer library for state-of-the-art code llm. arXiv preprint arXiv:230600029

Burgess CP, Matthey L, Watters N, et al. (2019) Monet: Unsupervised scene decomposition and representation. arXiv preprint arXiv:190111390

Byeon M, Park B, Kim H, et al. (2022) Coyo-700m: Image-text pair dataset

Byrne RM (2007) The rational imagination: How people create alternatives to reality. MIT press

Byrne RM, Tasso A (1999) Deductive reasoning with factual, possible, and counterfactual conditionals. Memory & cognition 27:726–740

Cai LW, Dai WZ, Huang YX, et al. (2021) Abductive learning with ground knowledge base. In: IJCAI, pp 1815–1821

Cai Q, Yates A (2013) Large-scale semantic parsing via schema matching and lexicon extension. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Sofia, Bulgaria, pp 423–433, URL https://aclanthology.org/P13-1042

Cao Y, Xu X, Sun C, et al. (2023) Segment any anomaly without training via hybrid prompt regularization. arXiv preprint arXiv:230510724

Carlini N, Jagielski M, Choquette-Choo CA, et al. (2023) Poisoning web-scale training datasets is practical. arXiv preprint arXiv:230210149

Chai Y, Wang S, Pang C, et al. (2022) Ernie-code: Beyond english-centric cross-lingual pretraining for programming languages. arXiv preprint arXiv:221206742

Chandel S, Clement CB, Serrato G, et al. (2022) Training and evaluating a jupyter notebook data science assistant. arXiv preprint arXiv:220112901

Chang TY, Liu Y, Gopalakrishnan K, et al. (2021) Go beyond plain fine-tuning: Improving pretrained models for social commonsense. 2105.05913

Charalambous Y, Tihanyi N, Jain R, et al. (2023) A new era in software security: Towards self-healing software via large language models and formal verification. arXiv preprint arXiv:230514752

Chen C, Feng X, Zhou J, et al. (2023a) Federated large language model: A position paper. 2307.08925

Chen G, Dong S, Shu Y, et al. (2023b) Autoagents: A framework for automatic agent generation. arXiv preprint arXiv:230917288

Chen J, Tang J, Qin J, et al. (2021a) Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. arXiv preprint arXiv:210514517

Chen J, Li T, Qin J, et al. (2022a) UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 3313–3323, URL https://aclanthology.org/2022.emnlp-main.218

Chen K, Li J, Wang K, et al. (2023c) Towards an automatic ai agent for reaction condition recommendation in chemical synthesis. arXiv preprint arXiv:231110776

Chen L, Wu P, Chitta K, et al. (2023d) End-to-end autonomous driving: Challenges and frontiers. arXiv preprint arXiv:230616927

Chen M, Tworek J, Jun H, et al. (2021b) Evaluating large language models trained on code. arXiv arXiv:2107.03374 [cs.LG]

Chen M, Tworek J, Jun H, et al. (2021c) Evaluating large language models trained on code. arXiv preprint arXiv:210703374

Chen S, Wang C, Chen Z, et al. (2022b) WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing 16(6):1505–1518

Chen S, Wong S, Chen L, et al. (2023e) Extending context window of large language models via positional interpolation. 2306.15595

Chen W, Zha H, Chen Z, et al. (2020a) HybridQA: A dataset of multi-hop question answering over tabular and textual data. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 1026–1036, https://doi.org/10.18653/v1/2020.findings-emnlp.91, URL

https://aclanthology.org/2020.findings-emnlp.91

Chen W, Ma X, Wang X, et al. (2022c) Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. arXiv preprint arXiv:221112588

Chen W, Yin M, Ku M, et al. (2023f) Theoremqa: A theorem-driven question answering dataset. arXiv preprint arXiv:230512524

Chen X, Ding M, Wang X, et al. (2023g) Context autoencoder for self-supervised representation learning. International Journal of Computer Vision pp 1–16

Chen X, Lin M, Schärli N, et al. (2023h) Teaching large language models to self-debug. 2304.05128

Chen Y, Li L, Yu L, et al. (2020b) UNITER: universal image-text representation learning. In: ECCV, pp 104–120

Chen Z, Chen W, Smiley C, et al. (2021d) Finqa: A dataset of numerical reasoning over financial data. Proceedings of EMNLP 2021

Chen Z, Yi K, Li Y, et al. (2022d) Comphy: Compositional physical reasoning of objects and events from videos. arXiv preprint arXiv:220501089

Chen Z, Zhou K, Zhang B, et al. (2023i) Chatcot: Tool-augmented chain-of-thought reasoning on chat-based large language models. 2305.14323

Cheng Y, Li L, Xu Y, et al. (2023) Segment and track anything. arXiv preprint arXiv:230506558

Cheng Z, Dong H, Wang Z, et al. (2022) HiTab: A hierarchical table dataset for question answering and natural language generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 1094–1110, https://doi.org/10.18653/v1/2022.acl-long.78, URL https://aclanthology.org/2022.acl-long.78

Chiang WL, Li Z, Lin Z, et al. (2023) Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. URL https://lmsys.org/blog/2023-03-30-vicuna/

Chowdhery A, Narang S, Devlin J, et al. (2022) Palm: Scaling language modeling with pathways. 2204.02311

Christopoulou F, Lampouras G, Gritta M, et al. (2022) Pangu-coder: Program synthesis with function-level language modeling. arXiv preprint arXiv:220711280

Chung HW, Hou L, Longpre S, et al. (2022) Scaling instruction-finetuned language models. arXiv preprint arXiv:221011416

Chung YA, Hsu WN, Tang H, et al. (2019) An unsupervised autoregressive model for speech representation learning. In: INTERSPEECH, pp 146–150

Clark P, Tafjord O, Richardson K (2020) Transformers as soft reasoners over language. arXiv preprint arXiv:200205867

Clement CB, Drain D, Timcheck J, et al. (2020) Pymt5: multi-mode translation of natural language and python code with transformers. arXiv preprint arXiv:201003150

Cobbe K, Kosaraju V, Bavarian M, et al. (2021) Training verifiers to solve math word problems. arXiv preprint arXiv:211014168

Collins A, Michalski R (1989) The logic of plausible reasoning: A core theory. cognitive science 13(1):1–49

Computer T (2023) Redpajama: An open source recipe to reproduce llama training dataset. URL https://github.com/togethercomputer/RedPajama-Data

Conover M, Hayes M, Mathur A, et al. (2023) Free dolly: Introducing the world's first truly open instruction-tuned llm

Creswell A, Shanahan M, Higgins I (2023) Selection-inference: Exploiting large language models for interpretable logical reasoning. In: The Eleventh International Conference on Learning Representations, URL https://openreview.net/forum?id=3Pf3Wg6o-A4

Cropper A, Dumančić S, Evans R, et al. (2022) Inductive logic programming at 30. Machine Learning pp 1–26

Dai W, Liu Z, Ji Z, et al. (2023) Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Dubrovnik, Croatia, pp 2136–2148, URL https://aclanthology.org/2023.eacl-main.156

Daniel K (2017) Thinking, fast and slow

Dao T, Fu DY, Saab KK, et al. (2022) Hungry hungry hippos: Towards language modeling with state space models. arXiv preprint arXiv:221214052

Das P, Sercu T, Wadhawan K, et al. (2021) Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. Nature Biomedical Engineering 5(6):613–623

Dasgupta I, Kaeser-Chen C, Marino K, et al. (2022) Collaborating with language models for embodied reasoning. In: Second Workshop on Language and Reinforcement Learning, URL https://openreview.net/forum?id=YoS-abmWjJc

De Raedt L, Kersting K (2010) Statistical relational learning. Encyclopedia of Machine Learning

Deitke M, Han W, Herrasti A, et al. (2020) Robothor: An open simulation-to-real embodied ai platform. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3164–3174

Deng J, Dong W, Socher R, et al. (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp 248–255

Desai K, Kaul G, Aysola Z, et al. (2021) Redcaps: Web-curated image-text data created by the people, for the people. arXiv preprint arXiv:211111431

Dettmers T, Pagnoni A, Holtzman A, et al. (2023) Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:230514314

Di P, Li J, Yu H, et al. (2023) Codefuse-13b: A pretrained multi-lingual code large language model. arXiv preprint arXiv:231006266

Ding D, Hill F, Santoro A, et al. (2020) Object-based attention for spatio-temporal reasoning: Outperforming neuro-symbolic models with flexible distributed architectures. arXiv preprint arXiv:201208508 1

Ding D, Hill F, Santoro A, et al. (2021a) Attention over learned object embeddings enables complex visual reasoning. In: Beygelzimer A, Dauphin Y, Liang P, et al. (eds) Advances in Neural Information Processing Systems, URL https://openreview.net/forum?id=lHmhW2zmVN

Ding J, Ma S, Dong L, et al. (2023a) Longnet: Scaling transformers to 1,000,000,000 tokens. 2307.02486

Ding M, Chen Z, Du T, et al. (2021b) Dynamic visual reasoning by learning differentiable physics models from video and language. Advances in Neural Information Processing Systems 34:887–899

Ding M, Xiao B, Codella N, et al. (2022) Davit: Dual attention vision transformers. In: European Conference on Computer Vision, Springer, pp 74–92

Ding M, Shen Y, Fan L, et al. (2023b) Visual dependency transformers: Dependency tree emerges from reversed attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14528–14539

Ding M, Xu Y, Chen Z, et al. (2023c) Embodied concept learner: Self-supervised learning of concepts and mapping through instruction following. In: Conference on Robot Learning, PMLR, pp 1743–1754

Ding N, Chen Y, Xu B, et al. (2023d) Enhancing chat language models by scaling high-quality instructional conversations. arXiv preprint arXiv:230514233

do Nascimento NM, de Lucena CJP (2017) Fiot: An agent-based framework for self-adaptive and self-organizing applications based on the internet of things. Information Sciences 378:161–176. https://doi.org/https://doi.org/10.1016/j.ins.2016.10.031, URL https://www.sciencedirect.com/science/article/pii/S0020025516313664

Dong H, Xiong W, Goyal D, et al. (2023a) Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:230406767

Dong L, Xu S, Xu B (2018) Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: ICASSP, pp 5884–5888

Dong Y, Jiang X, Jin Z, et al. (2023b) Self-collaboration code generation via chatgpt. 2304.07590

Dosovitskiy A, Beyer L, Kolesnikov A, et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Machine Learning

Driess D, Xia F, Sajjadi MSM, et al. (2023) Palm-e: An embodied multimodal language model. In: arXiv preprint arXiv:2303.03378

DriveLM Contributors (2023) Drive on Language. URL https://github.com/OpenDriveLab/DriveLM/

Du W, Kim ZM, Raheja V, et al. (2022) Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. arXiv preprint arXiv:220403685

Du Y, Li S, Torralba A, et al. (2023) Improving factuality and reasoning in language models through multiagent debate. 2305.14325

Dua D, Wang Y, Dasigi P, et al. (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). Association for Computational Linguistics, Minneapolis, Minnesota, pp 2368–2378, https://doi.org/10.18653/v1/N19-1246, URL https://aclanthology.org/N19-1246

Echterhoff J, Yan A, Han K, et al. (2023) Driving through the concept gridlock: Unraveling explainability bottlenecks. arXiv preprint arXiv:231016639

Edalati A, Tahaei M, Kobyzev I, et al. (2022) Krona: Parameter efficient tuning with kronecker adapter. arXiv preprint arXiv:221210650

Eichenberg C, Black S, Weinbach S, et al. (2022) MAGMA – multimodal augmentation of generative models through adapter-based finetuning. In: Findings of the

Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 2416–2428, URL https://aclanthology.org/2022.findings-emnlp.179

Espeholt L, Agrawal S, Sønderby C, et al. (2022) Deep learning for twelve hour precipitation forecasts. Nature communications 13(1):1–10

Evans JSB, Thompson VA (2004) Informal reasoning: Theory and method. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale 58(2):69

Fan L, Krishnan D, Isola P, et al. (2023) Improving clip training with language rewrites. arXiv preprint arXiv:230520088

Fang Y, Wang S, Xu Y, et al. (2022) Leveraging knowledge in multilingual commonsense reasoning. In: Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, Dublin, Ireland, pp 3237–3246, https://doi.org/10.18653/v1/2022.findings-acl.255, URL https://aclanthology.org/2022.findings-acl.255

Fayek HM, Johnson J (2019) Temporal reasoning via audio question answering. 1911.09655

Feder A, Oved N, Shalit U, et al. (2021) CausaLM: Causal Model Explanation Through Counterfactual Language Models. Computational Linguistics 47(2):333–386. https://doi.org/10.1162/coli_a_00404, URL https://doi.org/10.1162/coli_a_00404, https://direct.mit.edu/coli/article-pdf/47/2/333/1938107/coli_a_00404.pdf

Fei N, Lu Z, Gao Y, et al. (2022) Towards artificial general intelligence via a multimodal foundation model. Nature Communications 13(1):3094

Firoozi R, Sun J, Tucker J, et al. (2023) Foundation models in robotics: Applications, challenges, and the future. arXiv preprint

Flach PA, Kakas AC (2000) Abductive and Inductive Reasoning: Background and Issues, Springer Netherlands, Dordrecht, pp 1–27. https://doi.org/10.1007/978-94-017-0606-3_1, URL https://doi.org/10.1007/978-94-017-0606-3_1

Floyd J (2004) Wittgenstein on philosophy of logic and mathematics. Graduate Faculty Philosophy Journal 25(2):227–287

Fried D, Aghajanyan A, Lin J, et al. (2022) Incoder: A generative model for code infilling and synthesis. arXiv preprint arXiv:220405999

Friedman R (2023a) Large language models and logical reasoning. Encyclopedia 3(2):687–697

Friedman R (2023b) Tokenization in the theory of knowledge. Encyclopedia 3(1):380–386

Frohberg J, Binder F (2021) Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. arXiv preprint arXiv:211211941

Fu Y, Peng H, Sabharwal A, et al. (2022) Complexity-based prompting for multi-step reasoning. arXiv preprint arXiv:221000720

Fu Y, Peng H, Khot T, et al. (2023a) Improving language model negotiation with self-play and in-context learning from ai feedback. 2305.10142

Fu Y, Peng H, Ou L, et al. (2023b) Specializing smaller language models towards multi-step reasoning. arXiv preprint arXiv:230112726

Fu Z, Lam W, Yu Q, et al. (2023c) Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. arXiv preprint arXiv:230404052

Fu Z, Yang H, So AMC, et al. (2023d) On the effectiveness of parameter-efficient fine-tuning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 12799–12807

Furbach U, Hölldobler S, Ragni M, et al. (2019) Cognitive reasoning: A personal view. KI-Künstliche Intelligenz 33:209–217

Gadre SY, Ilharco G, Fang A, et al. (2023) Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:230414108

Gao D, Ji L, Zhou L, et al. (2023a) Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. 2306.08640

Gao L, Biderman S, Black S, et al. (2020) The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:210100027

Gao L, Madaan A, Zhou S, et al. (2023b) Pal: Program-aided language models. In: International Conference on Machine Learning, PMLR, pp 10764–10799

Gao P, Han J, Zhang R, et al. (2023c) Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:230415010

Garcez Ad, Besold TR, De Raedt L, et al. (2015) Neural-symbolic learning and reasoning: contributions and challenges. In: 2015 AAAI Spring Symposium Series

Garcez Ad, Bader S, Bowman H, et al. (2022) Neural-symbolic learning and reasoning: A survey and interpretation. Neuro-Symbolic Artificial Intelligence: The State of the Art 342(1):327

Garcez AS, Lamb LC, Gabbay DM (2008) Neural-symbolic cognitive reasoning. Springer Science & Business Media

Gauthier T, Kaliszyk C, Urban J, et al. (2021) Tactictoe: learning to prove with tactics. Journal of Automated Reasoning 65:257–286

Ge Y, Hua W, Mei K, et al. (2023) Openagi: When llm meets domain experts. 2304.04370

Gendron G, Bao Q, Witbrock M, et al. (2023) Large language models are not abstract reasoners. 2305.19555

Geng X, Gudibande A, Liu H, et al. (2023) Koala: A dialogue model for academic research. Blog post, URL https://bair.berkeley.edu/blog/2023/04/03/koala/

Girdhar R, Ramanan D (2020) Cater: A diagnostic dataset for compositional actions and temporal reasoning. In: ICLR

Girdhar R, El-Nouby A, Liu Z, et al. (2023) Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 15180–15190

Gou Z, Shao Z, Gong Y, et al. (2023) Critic: Large language models can self-correct with tool-interactive critiquing. 2305.11738

gravitas/auto gpt S (2023) An experimental open-source attempt to make gpt-4 fully autonomou. 2305.16291

Gu A, Goel K, Ré C (2021) Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:211100396

Gu J, Han Z, Chen S, et al. (2023a) A systematic survey of prompt engineering on vision-language foundation models. 2307.12980

Gu Y, Dong L, Wei F, et al. (2023b) Knowledge distillation of large language models. 2306.08543

Gulati A, Qin J, Chiu CC, et al. (2020) Conformer: Convolution-augmented transformer for speech recognition. In: INTERSPEECH, pp 5036–5040

Guo J, Li J, Li D, et al. (2023a) From images to textual prompts: Zero-shot vqa with frozen large language models. 2212.10846

Guo Z, Zhang R, Zhu X, et al. (2023b) Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. arXiv preprint arXiv:230900615

Gupta A, Gu A, Berant J (2022) Diagonal state spaces are as effective as structured state spaces. Advances in Neural Information Processing Systems 35:22982–22994

Gupta N, Lin K, Roth D, et al. (2019) Neural module networks for reasoning over text. arXiv preprint arXiv:191204971

Gupta T, Kembhavi A (2022) Visual programming: Compositional visual reasoning without training. 2211.11559

Halpern JY (2016) Actual causality. MiT Press

Han JM, Rute J, Wu Y, et al. (2021) Proof artifact co-training for theorem proving with language models. arXiv preprint arXiv:210206203

Han S, Schoelkopf H, Zhao Y, et al. (2022) Folio: Natural language reasoning with first-order logic. 2209.00840

Hao S, Gu Y, Ma H, et al. (2023a) Reasoning with language model is planning with world model. 2305.14992

Hao S, Liu T, Wang Z, et al. (2023b) Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. 2305.11554

Harrison J (2010) Formal methods at intel—an overview. In: Second NASA Formal Methods Symposium, pp 179–195

He H, Zhang J, Xu M, et al. (2023) Scalable mask annotation for video text spotting. arXiv preprint arXiv:230501443

He J, Zhou C, Ma X, et al. (2021) Towards a unified view of parameter-efficient transfer learning. In: International Conference on Learning Representations

He K, Chen X, Xie S, et al. (2022) Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16000–16009

Hemphill CT, Godfrey JJ, Doddington GR (1990) The ATIS spoken language systems pilot corpus. In: Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990, URL https://aclanthology.org/H90-1021

Hendrycks D, Basart S, Kadavath S, et al. (2021a) Measuring coding challenge competence with apps. NeurIPS

Hendrycks D, Burns C, Kadavath S, et al. (2021b) Measuring mathematical problem solving with the math dataset. NeurIPS

Hinton GE (1990) Connectionist learning procedures. In: Machine learning. Elsevier, p 555–610

Ho N, Schmid L, Yun SY (2022) Large language models are reasoning teachers. arXiv preprint arXiv:221210071

Hong Y, Li Q, Zhu SC, et al. (2021a) Vlgrammar: Grounded grammar induction of vision and language

Hong Y, Yi L, Tenenbaum JB, et al. (2021b) Ptr: A benchmark for part-based conceptual, relational, and physical reasoning. 2112.05136

Hong Y, Zhen H, Chen P, et al. (2023) 3d-llm: Injecting the 3d world into large language models. arXiv

Hongjin S, Kasai J, Wu CH, et al. (2022) Selective annotation makes language models better few-shot learners. In: The Eleventh International Conference on Learning Representations

Honovich O, Scialom T, Levy O, et al. (2022) Unnatural instructions: Tuning language models with (almost) no human labor. URL https://arxiv.org/abs/2212.09689

Horawalavithana S, Ayton E, Sharma S, et al. (2022) Foundation models of scientific knowledge for chemistry: Opportunities, challenges and lessons learned. In: Fan A, Ilic S, Wolf T, et al. (eds) Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models. Association for Computational Linguistics, virtual+Dublin, pp 160–172, https://doi.org/10.18653/v1/2022.bigscience-1.12, URL https://aclanthology.org/2022.bigscience-1.12

Hosseini MJ, Hajishirzi H, Etzioni O, et al. (2014) Learning to solve arithmetic word problems with verb categorization. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 523–533, https://doi.org/10.3115/v1/D14-1058, URL https://aclanthology.org/D14-1058

Houlsby N, Giurgiu A, Jastrzebski S, et al. (2019) Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning, PMLR, pp 2790–2799

Hsieh CY, Li CL, Yeh Ck, et al. (2023) 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. Association for Computational Linguistics, Toronto, Canada, pp 8003–8017, https://doi.org/10.18653/v1/2023.findings-acl.507, URL https://aclanthology.org/2023.findings-acl.507

Hsu WN, Bolte B, Tsai YHH, et al. (2021) HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29:3451–3460

Hsu YC, Hua T, Chang S, et al. (2022) Language model compression with weighted low-rank factorization. 2207.00112

Hu EJ, yelong shen, Wallis P, et al. (2022) LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=nZeVKeeFYf9

Hu M, Mu Y, Yu X, et al. (2023a) Tree-planner: Efficient close-loop task planning with large language models. arXiv preprint arXiv:231008582

Hu Y, Yang H, Lin Z, et al. (2023b) Code prompting: a neural symbolic method for complex reasoning in large language models. 2305.18507

Huang D, Shi S, Lin CY, et al. (2016) How well do computers solve math word problems? large-scale dataset construction and evaluation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 887–896

Huang G, Hu J, He Y, et al. (2021a) Machine learning for electronic design automation: A survey. ACM Transactions on Design Automation of Electronic Systems (TODAES) 26(5):1–46

Huang H, Tang T, Zhang D, et al. (2023a) Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. 2305.07004

Huang J, Chang KCC (2022) Towards reasoning in large language models: A survey. arXiv preprint arXiv:221210403

Huang J, Xie S, Sun J, et al. (2021b) Learning a decision module by imitating driver's control behaviors. In: Kober J, Ramos F, Tomlin C (eds) Proceedings of the 2020 Conference on Robot Learning, Proceedings of Machine Learning Research, vol 155. PMLR, pp 1–10, URL https://proceedings.mlr.press/v155/huang21a.html

Huang J, Gu SS, Hou L, et al. (2022a) Large language models can self-improve. arXiv preprint arXiv:221011610

Huang J, Zhu WY, Jia B, et al. (2023b) Perceive, ground, reason, and act: A benchmark for general-purpose visual representation. URL https://openreview.net/forum?id=f6cywgfd11

Huang K, Sun K, Xie E, et al. (2023c) T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. arXiv preprint arXiv:230706350

Huang L, Bras RL, Bhagavatula C, et al. (2019) Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. arXiv preprint arXiv:190900277

Huang S, Dong L, Wang W, et al. (2023d) Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:230214045

Huang W, Abbeel P, Pathak D, et al. (2022b) Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In: International Conference on Machine Learning, PMLR, pp 9118–9147

Huang W, Xia F, Xiao T, et al. (2022c) Inner monologue: Embodied reasoning through planning with language models. In: arXiv preprint arXiv:2207.05608

Huang X, Ruan W, Huang W, et al. (2023e) A survey of safety and trustworthiness of large language models through the lens of verification and validation. 2305.11391

Huang Y, Kleindessner M, Munishkin A, et al. (2021c) Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. Frontiers in big Data 4:642182

Hudson DA, Manning CD (2019) Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6700–6709

Huo Y, Zhang M, Liu G, et al. (2021) Wenlan: Bridging vision and language by large-scale multi-modal pre-training. arXiv preprint arXiv:210306561

Imani S, Du L, Shrivastava H (2023) Mathprompter: Mathematical reasoning using large language models. 2303.05398

Inaba T, Kiyomaru H, Cheng F, et al. (2023) Multitool-cot: Gpt-3 can use multiple external tools with chain of thought prompting. 2305.16896

Iyer S, Lin XV, Pasunuru R, et al. (2022) Opt-iml: Scaling language model instruction meta learning through the lens of generalization. arXiv preprint arXiv:221212017

Jaderberg M, Simonyan K, Zisserman A, et al. (2016) Spatial transformer networks. 1506.02025

Jain N, Saifullah K, Wen Y, et al. (2023) Bring your own data! self-supervised evaluation for large language models. arXiv preprint arXiv:230613651

Jansen P (2020) Visually-grounded planning without vision: Language models infer detailed plans from high-level instructions. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 4412–4417, https://doi.org/10.18653/v1/2020.findings-emnlp.395, URL https://aclanthology.org/2020.findings-emnlp.395

Jia C, Yang Y, Xia Y, et al. (2021) Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning, PMLR, pp 4904–4916

Jiang AQ, Li W, Han JM, et al. (2021a) Lisa: Language models of isabelle proofs. In: 6th Conference on Artificial Intelligence and Theorem Proving, pp 378–392

Jiang AQ, Welleck S, Zhou JP, et al. (2023) Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In: The Eleventh International Conference on Learning Representations

Jiang D, Li W, Cao M, et al. (2020) Speech SIMCLR: Combining contrastive and reconstruction objective for self-supervised speech representation learning. In: INTERSPEECH, pp 1544–1548

Jiang D, Li W, Zhang R, et al. (2021b) A further study of unsupervised pretraining for transformer based speech recognition. In: ICASSP, pp 6538–6542

Jiang Z, Araki J, Ding H, et al. (2021c) How can we know when language models know? on the calibration of language models for question answering. Transactions of the Association for Computational Linguistics 9:962–977

Jin Q, Yang Y, Chen Q, et al. (2023a) Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. ArXiv

Jin Z, Liu J, Lyu Z, et al. (2023b) Can large language models infer causation from correlation? arXiv preprint arXiv:230605836

Johnson J, Hariharan B, Van Der Maaten L, et al. (2017) Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2901–2910

Joshi M, Choi E, Weld D, et al. (2017) TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vancouver, Canada, pp 1601–1611, https://doi.org/10.18653/v1/P17-1147, URL https://aclanthology.org/P17-1147

Jumper J, Evans R, Pritzel A, et al. (2021) Highly accurate protein structure prediction with alphafold. Nature 596(7873):583–589

Kahn J, Rivière M, Zheng W, et al. (2020) Libri-light: A benchmark for asr with limited or no supervision. In: ICASSP, pp 7669–7673

Kahneman D, Miller DT (1986) Norm theory: Comparing reality to its alternatives. Psychological review 93(2):136

Karimi Mahabadi R, Henderson J, Ruder S (2021) Compacter: Efficient low-rank hypercomplex adapter layers. Advances in Neural Information Processing Systems 34:1022–1035

Katsis Y, Chemmengath S, Kumar V, et al. (2022) AIT-QA: Question answering dataset over complex tables in the airline industry. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies: Industry Track. Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, pp 305–314, https://doi.org/10.18653/v1/2022.naacl-industry.34, URL https://aclanthology.org/2022.naacl-industry.34

Kazemi M, Yuan Q, Bhatia D, et al. (2023) Boardgameqa: A dataset for natural language reasoning with contradictory information. arXiv preprint arXiv:230607934

Kenton JDMWC, Toutanova LK (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, p 2

Khan W, Kamran M, Naqvi SR, et al. (2020) Formal verification of hardware components in critical systems. Wireless Communications and Mobile Computing 2020:1–15

Khashabi D, Min S, Khot T, et al. (2020) Unifiedqa: Crossing format boundaries with a single qa system. arXiv preprint arXiv:200500700

Kıcıman E, Ness R, Sharma A, et al. (2023) Causal reasoning and large language models: Opening a new frontier for causality. arXiv preprint arXiv:230500050

Kim J, Misu T, Chen YT, et al. (2019) Grounding human-to-vehicle advice for self-driving vehicles. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition

Kim S, Joo SJ, Kim D, et al. (2023) The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. arXiv preprint arXiv:230514045

Kirchenbauer J, Geiping J, Wen Y, et al. (2023) A watermark for large language models. 2301.10226

Kirillov A, Mintun E, Ravi N, et al. (2023) Segment anything. arXiv:230402643

Kocetkov D, Li R, Allal LB, et al. (2022) The stack: 3 tb of permissively licensed source code. arXiv preprint arXiv:221115533

Kojima T, Gu SS, Reid M, et al. (2022) Large language models are zero-shot reasoners. Advances in neural information processing systems 35:22199–22213

Koncel-Kedziorski R, Hajishirzi H, Sabharwal A, et al. (2015) Parsing algebraic word problems into equations. Transactions of the Association for Computational Linguistics 3:585–597. https://doi.org/10.1162/tacl_a_00160, URL https://aclanthology.org/Q15-1042

Koncel-Kedziorski R, Roy S, Amini A, et al. (2016) Mawps: A math word problem repository. In: North American Chapter of the Association for Computational Linguistics

Kondo K, Sugawara S, Aizawa A (2023) Probing physical reasoning with counter-commonsense context. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Toronto, Canada, pp 603–612, https://doi.org/10.18653/v1/2023.acl-short.53, URL https://aclanthology.org/2023.acl-short.53

Koons R (2005) Defeasible reasoning. arXiv

Kosinski M (2023) Theory of mind may have spontaneously emerged in large language models. 2302.02083

Kreps S, McCain RM, Brundage M (2022) All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. Journal of experimental political science 9(1):104–117

Kurita K, Michel P, Neubig G (2020) Weight poisoning attacks on pre-trained models. arXiv preprint arXiv:200406660

Kushman N, Artzi Y, Zettlemoyer L, et al. (2014) Learning to automatically solve algebra word problems. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Baltimore, Maryland, pp 271–281, https://doi.org/10.3115/v1/P14-1026, URL https://aclanthology.org/P14-1026

Kwiatkowski T, Palomaki J, Redfield O, et al. (2019) Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics 7:452–466. https://doi.org/10.1162/tacl_a_00276, URL https://aclanthology.org/Q19-1026

Köksal A, Schick T, Korhonen A, et al. (2023) Longform: Optimizing instruction tuning for long text generation with corpus extraction. 2304.08460

Köpf A, Kilcher Y, von Rütte D, et al. (2023) Openassistant conversations – democratizing large language model alignment. 2304.07327

Laban P, Kryściński W, Agarwal D, et al. (2023) Llms as factual reasoners: Insights from existing benchmarks and beyond. 2305.14540

Lahiri S (2014) Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In: Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Gothenburg, Sweden, pp 96–105, URL http://www.aclweb.org/anthology/E14-3011

Lample G, Lacroix T, Lachaux MA, et al. (2022) Hypertree proof search for neural theorem proving. Advances in Neural Information Processing Systems 35:26337–26349

Laurençon H, Saulnier L, Wang T, et al. (2022) The bigscience roots corpus: A 1.6 tb composite multilingual dataset. Advances in Neural Information Processing Systems 35:31809–31826

Laurent J, Platzer A (2022) Learning to find proofs and theorems by learning to refine search strategies: The case of loop invariant synthesis. Advances in Neural Information Processing Systems 35:4843–4856

Le H, Wang Y, Gotmare AD, et al. (2022) Coderl: Mastering code generation through pretrained models and deep reinforcement learning. Advances in Neural Information Processing Systems 35:21314–21328

Leake DB (2012) Introspective Learning and Reasoning, Springer US, Boston, MA, pp 1638–1640. https://doi.org/10.1007/978-1-4419-1428-6_1802, URL https://doi.org/10.1007/978-1-4419-1428-6_1802

Lee YJ, Lim CG, Choi HJ (2022) Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In: Proceedings of the 29th International Conference on Computational Linguistics, pp 669–683

Lester B, Al-Rfou R, Constant N (2021) The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 3045–3059, https://doi.org/10.18653/v1/2021.emnlp-main.243, URL https://aclanthology.org/2021.emnlp-main.243

Levy I, Bogin B, Berant J (2022) Diverse demonstrations improve in-context compositional generalization. arXiv preprint arXiv:221206800

Lewis M, Liu Y, Goyal N, et al. (2020) BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 7871–7880, https://doi.org/10.18653/v1/2020.acl-main.703, URL https://aclanthology.org/2020.acl-main.703

Lewkowycz A, Andreassen AJ, Dohan D, et al. (2022) Solving quantitative reasoning problems with language models. In: Oh AH, Agarwal A, Belgrave D, et al. (eds) Advances in Neural Information Processing Systems, URL https://openreview.net/forum?id=IFXTZERXdM7

Li C (2023) Large multimodal models: Notes on cvpr 2023 tutorial. 2306.14895

Li C, Xia F, Martín-Martín R, et al. (2021a) igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In: 5th Annual Conference on Robot Learning, URL https://openreview.net/forum?id=2uGN5jNJROR

Li C, Zhang R, Wong J, et al. (2022a) BEHAVIOR-1k: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation. In: 6th Annual Conference on Robot Learning, URL https://openreview.net/forum?id=_8DoIe8G3t

Li C, Wong C, Zhang S, et al. (2023a) Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:230600890

Li G, Hammoud HAAK, Itani H, et al. (2023b) Camel: Communicative agents for" mind" exploration of large language model society. In: Thirty-seventh Conference on Neural Information Processing Systems

Li H, Guo D, Fan W, et al. (2023c) Multi-step jailbreaking privacy attacks on chatgpt. arXiv preprint arXiv:230405197

Li H, Sima C, Dai J, et al. (2023d) Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. 2209.05324

Li J, Seltzer ML, Wang X, et al. (2017) Large-scale domain adaptation via teacher-student learning. In: INTERSPEECH, pp 2386–2390

Li J, Li D, Xiong C, et al. (2022b) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. 2201.12086

Li J, Li D, Savarese S, et al. (2023e) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 2301.12597

Li J, Yu L, Ettinger A (2023f) Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. arXiv preprint arXiv:230516572

Li J, Zhao Y, Li Y, et al. (2023g) Acecoder: Utilizing existing code to enhance code generation. 2303.17780

Li L, Spratling M (2023) Data augmentation alone can improve adversarial training. arXiv preprint arXiv:230109879

Li L, Szygenda SA, Thornton MA (2005) Combining simulation and formal verification for integrated circuit design validation. In: Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI), pp 92–97

Li L, Qiu J, Spratling M (2023h) Aroid: Improving adversarial robustness through online instance-wise data augmentation. arXiv preprint arXiv:230607197

Li R, Allal LB, Zi Y, et al. (2023i) Starcoder: may the source be with you! arXiv preprint arXiv:230506161

Li S, Chen J, Shen Y, et al. (2022c) Explanations from large language models make small reasoners better. arXiv preprint arXiv:221006726

Li S, Puig X, Paxton C, et al. (2022d) Pre-trained language models for interactive decision-making. Advances in Neural Information Processing Systems 35:31199–31212

Li T, Chen L, Wang H, et al. (2023j) Graph-based topology reasoning for driving scenes. 2304.05277

Li X, Qiu X (2023a) Finding supporting examples for in-context learning. arXiv preprint arXiv:230213539

Li X, Qiu X (2023b) Mot: Memory-of-thought enables chatgpt to self-improve. 2305.05181

Li X, Yin X, Li C, et al. (2020) Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV, pp 121–137

Li X, Sun Y, Cheng G (2021b) Tsqa: Tabular scenario based question answering. ArXiv abs/2101.11429

Li X, Liu M, Zhang H, et al. (2023k) Vision-language foundation models as effective robot imitators. arXiv preprint arXiv:231101378

Li X, Lv K, Yan H, et al. (2023l) Unified demonstration retriever for in-context learning. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp 4644–4668, https://doi.org/10.18653/v1/2023.acl-long.256, URL https://aclanthology.org/2023.acl-long.256

Li X, Yu P, Zhou C, et al. (2023m) Self-alignment with instruction backtranslation. arXiv preprint arXiv:230806259

Li XL, Liang P (2021) Prefix-tuning: Optimizing continuous prompts for generation. 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). Association for Computational Linguistics, Online, pp 4582–4597, https://doi.org/10.18653/v1/2021.acl-long.353, URL https://aclanthology.org/2021.acl-long.353

Li XL, Kuncoro A, Hoffmann J, et al. (2022e) A systematic investigation of commonsense knowledge in large language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp 11838–11855

Li Y, Choi D, Chung J, et al. (2022f) Competition-level code generation with alphacode. Science 378(6624):1092–1097

Li Y, Lin Z, Zhang S, et al. (2022g) On the advance of making language models better reasoners. arXiv preprint arXiv:220602336

Li Y, Du Y, Zhou K, et al. (2023n) Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:230510355

Li Y, Fan H, Hu R, et al. (2023o) Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 23390–23400

Li Y, Gao C, Song X, et al. (2023p) Druggpt: A gpt-based strategy for designing potential ligands targeting specific proteins. bioRxiv pp 2023–06

Li Y, Wang H, Duan Y, et al. (2023q) Clip surgery for better explainability with enhancement in open-vocabulary tasks. arXiv preprint arXiv:230405653

Li Y, Zhou K, Zhao WX, et al. (2023r) Diffusion models for non-autoregressive text generation: A survey. arXiv preprint arXiv:230306574

Lian D, Zhou D, Feng J, et al. (2022) Scaling & shifting your features: A new baseline for efficient model tuning. Advances in Neural Information Processing Systems 35:109–123

Liang J, Huang W, Xia F, et al. (2022a) Code as policies: Language model programs for embodied control. In: arXiv preprint arXiv:2209.07753

Liang Z, Zhang J, Wang L, et al. (2022b) Mwp-bert: Numeracy-augmented pre-training for math word problem solving. In: Findings of NAACL 2022, pp 997–1009

Liao QV, Vaughan JW (2023) Ai transparency in the age of llms: A human-centered research roadmap. arXiv preprint arXiv:230601941

Lightman H, Kosaraju V, Burda Y, et al. (2023) Let's verify step by step. 2305.20050

Lin BY, Sun H, Dhingra B, et al. (2020a) Differentiable open-ended commonsense reasoning. arXiv preprint arXiv:201014439

Lin BY, Zhou W, Shen M, et al. (2020b) CommonGen: A constrained text generation challenge for generative commonsense reasoning. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 1823–1840, https://doi.org/10.18653/v1/2020.findings-emnlp.165, URL https://aclanthology.org/2020.findings-emnlp.165

Lin BY, Fu Y, Yang K, et al. (2023) Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. ArXiv preprint abs/2305.17390. URL https://arxiv.org/abs/2305.17390

Lin Z, Wu YF, Peri S, et al. (2020c) Improving generative imagination in object-centric world models. 2010.02054

Ling W, Yogatama D, Dyer C, et al. (2017a) Program induction by rationale generation: Learning to solve and explain algebraic word problems. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vancouver, Canada, pp 158–167, https://doi.org/10.18653/v1/P17-1015, URL https://aclanthology.org/P17-1015

Ling W, Yogatama D, Dyer C, et al. (2017b) Program induction by rationale generation: Learning to solve and explain algebraic word problems. arXiv preprint arXiv:170504146

Liu AT, Yang Sw, Chi PH, et al. (2020a) Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In: ICASSP, pp 6419–6423

Liu C, Shen J, Xin H, et al. (2023a) Fimo: A challenge formal dataset for automated theorem proving. arXiv preprint arXiv:230904295

Liu F, Eisenschlos JM, Piccinno F, et al. (2023b) Deplot: One-shot visual language reasoning by plot-to-table translation. In: Findings of the 61st Annual Meeting of the Association for Computational Linguistics, URL https://arxiv.org/abs/2212.10505

Liu F, Piccinno F, Krichene S, et al. (2023c) Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, URL https://arxiv.org/abs/2212.09662

Liu H, Li C, Li Y, et al. (2023d) Improved baselines with visual instruction tuning. arXiv preprint arXiv:231003744

Liu H, Li C, Wu Q, et al. (2023e) Visual instruction tuning. arXiv preprint arXiv:230408485

Liu J, Shen D, Zhang Y, et al. (2022a) What makes good in-context examples for gpt-3? In: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pp 100–114

Liu J, Huang Z, Ma Z, et al. (2023f) Guiding mathematical reasoning via mastering commonsense formula knowledge. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, KDD '23, p 1477–1488, https://doi.org/10.1145/3580305.3599375, URL https://doi.org/10.1145/3580305.3599375

Liu NF, Lin K, Hewitt J, et al. (2023g) Lost in the middle: How language models use long contexts. 2307.03172

Liu S, Zeng Z, Ren T, et al. (2023h) Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:230305499

Liu T, Guo Q, Hu X, et al. (2022b) RLET: A reinforcement learning based approach for explainable QA with entailment trees. In: Goldberg Y, Kozareva Z, Zhang Y (eds) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 7177–7189, https://doi.org/10.18653/v1/2022.emnlp-main.483, URL https://aclanthology.org/2022.emnlp-main.483

Liu T, Guo Q, Yang Y, et al. (2023i) Plan, verify and switch: Integrated reasoning with diverse x-of-thoughts. 2310.14628

Liu X, Ji K, Fu Y, et al. (2021a) P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:211007602

Liu X, Yin D, Feng Y, et al. (2022c) Things not written in text: Exploring spatial commonsense from visual signals. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 2365–2376, https://doi.org/10.18653/v1/2022.acl-long.168, URL https://aclanthology.org/2022.acl-long.168

Liu X, Yin D, Zhang C, et al. (2023j) The magic of if: Investigating causal reasoning abilities in large language models of code. arXiv preprint arXiv:230519213

Liu X, Zheng Y, Du Z, et al. (2023k) Gpt understands, too. AI Open

Liu Y, Ott M, Goyal N, et al. (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692

Liu Y, Ott M, Goyal N, et al. (2020b) Ro{bert}a: A robustly optimized {bert} pretraining approach. URL https://openreview.net/forum?id=SyxS0T4tvS

Liu Y, Fabbri AR, Liu P, et al. (2023l) On learning to summarize with large language models as references. arXiv preprint arXiv:230514239

Liu Y, Li Z, Li H, et al. (2023m) On the hidden mystery of ocr in large multimodal models. 2305.07895

Liu Z, Lin Y, Cao Y, et al. (2021b) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022

Long J (2023) Large language model guided tree-of-thought. 2305.08291

Long S, Schuster T, Piché A (2022) Can large language models build causal graphs? In: NeurIPS 2022 Workshop on Causality for Real-world Impact, URL https://openreview.net/forum?id=LQQoJGw8JD1

Long S, Piché A, Zantedeschi V, et al. (2023) Causal discovery with language models as imperfect experts. arXiv preprint arXiv:230702390

Longpre S, Hou L, Vu T, et al. (2023) The flan collection: Designing data and methods for effective instruction tuning. arXiv preprint arXiv:230113688

Lou R, Zhang K, Yin W (2023) Is prompt all you need? no. a comprehensive and broader view of instruction learning. arXiv preprint arXiv:230310475

Lourie N, Le Bras R, Bhagavatula C, et al. (2021) Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 13480–13488

Lu P, Gong R, Jiang S, et al. (2021a) Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. arXiv preprint arXiv:210504165

Lu P, Qiu L, Chen J, et al. (2021b) Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In: The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks

Lu P, Mishra S, Xia T, et al. (2022a) Learn to explain: Multimodal reasoning via thought chains for science question answering. In: The 36th Conference on Neural Information Processing Systems (NeurIPS)

Lu P, Qiu L, Chang KW, et al. (2022b) Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. arXiv preprint arXiv:220914610

Lu P, Peng B, Cheng H, et al. (2023) Chameleon: Plug-and-play compositional reasoning with large language models. 2304.09842

Luo G, Huang M, Zhou Y, et al. (2023a) Towards efficient visual adaption via structural re-parameterization. arXiv preprint arXiv:230208106

Luo G, Zhou Y, Ren T, et al. (2023b) Cheap and quick: Efficient vision-language instruction tuning for large language models. arXiv preprint arXiv:230515023

Luo H, Sun Q, Xu C, et al. (2023c) Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. arXiv preprint arXiv:230809583

Luo M, Kumbhar S, shen M, et al. (2023d) Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. 2310.00836

Luo M, Xu X, Dai Z, et al. (2023e) Dr. icl: Demonstration-retrieved in-context learning. arXiv preprint arXiv:230514128

Luo Z, Xu C, Zhao P, et al. (2023f) Wizardcoder: Empowering code large language models with evol-instruct. arXiv preprint arXiv:230608568

LYU Z, Jin Z, Mihalcea R, et al. (2022) Can large language models distinguish cause from effect? In: UAI 2022 Workshop on Causal Representation Learning, URL https://openreview.net/forum?id=ucHh-ytUkOH

Ma X, Yong S, Zheng Z, et al. (2023) Sqa3d: Situated question answering in 3d scenes. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=IDJx97BC38

Madaan A, Tandon N, Rajagopal D, et al. (2021) Think about it! improving defeasible reasoning by first modeling the question scenario. arXiv preprint arXiv:211012349

Madaan A, Zhou S, Alon U, et al. (2022a) Language models of code are few-shot commonsense learners. arXiv preprint arXiv:221007128

Madaan A, Zhou S, Alon U, et al. (2022b) Language models of code are few-shot commonsense learners. In: Goldberg Y, Kozareva Z, Zhang Y (eds) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 1384–1403, https://doi.org/10.18653/v1/2022.emnlp-main.90, URL https://aclanthology.org/2022.emnlp-main.90

Madaan A, Tandon N, Gupta P, et al. (2023) Self-refine: Iterative refinement with self-feedback. 2303.17651

Madani A, Krause B, Greene ER, et al. (2023) Large language models generate functional protein sequences across diverse families. Nature Biotechnology pp 1–8

Magister LC, Mallinson J, Adamek J, et al. (2022) Teaching small language models to reason. arXiv preprint arXiv:221208410

Manhaeve R, Dumančić S, Kimmig A, et al. (2021) Neural probabilistic logic programming in deepproblog. Artificial Intelligence 298:103504. https://doi.org/https://doi.org/10.1016/j.artint.2021.103504, URL https://www.sciencedirect.com/science/article/pii/S0004370221000552

Manica M, Born J, Cadow J, et al. (2023) Accelerating material design with the generative toolkit for scientific discovery. npj Computational Materials 9(1):69

Manning CD (2022) Human language understanding & reasoning. Daedalus 151(2):127–138

Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, et al. (2019) Habitat: A Platform for Embodied AI Research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

Mao J, Gan C, Kohli P, et al. (2019) The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=rJgMlhRctm

Mao J, Yang X, Zhang X, et al. (2022) Clevrer-humans: Describing physical and causal events the human way. Advances in Neural Information Processing Systems 35:7755–7768

Mao J, Qian Y, Zhao H, et al. (2023a) Gpt-driver: Learning to drive with gpt. 2310.01415

Mao J, Ye J, Qian Y, et al. (2023b) A language agent for autonomous driving. arXiv

Marino K, Rastegari M, Farhadi A, et al. (2019) Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Conference on Computer Vision and Pattern Recognition (CVPR)

Megill N, Wheeler DA (2019) A computer language for mathematical proofs. arXiv

Mehta H, Gupta A, Cutkosky A, et al. (2022) Long range language modeling via gated state spaces. arXiv preprint arXiv:220613947

Miao Sy, Liang CC, Su KY (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, pp 975–984

Mikuła M, Antoniak S, Tworkowski S, et al. (2023) Magnushammer: A transformer-based approach to premise selection. arXiv preprint arXiv:230304488

Miller A, Fisch A, Dodge J, et al. (2016) Key-value memory networks for directly reading documents. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pp 1400–1409, https://doi.org/10.18653/v1/D16-1147, URL https://aclanthology.org/D16-1147

Min S, Lewis M, Zettlemoyer L, et al. (2022) MetaICL: Learning to learn in context. In: NAACL-HLT

Mishra S, Khashabi D, Baral C, et al. (2022) Cross-task generalization via natural language crowdsourcing instructions. In: ACL

Misra K, Rayz JT, Ettinger A (2022) A property induction framework for neural language models. arXiv preprint arXiv:220506910

Mitra A, Del Corro L, Mahajan S, et al. (2023) Orca 2: Teaching small language models how to reason. arXiv preprint arXiv:231111045

Mohamed A, Lee Hy, Borgholt L, et al. (2022) Self-supervised speech representation learning: A review. IEEE Journal of Selected Topics in Signal Processing

Mooij JM, Peters J, Janzing D, et al. (2016) Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research 17(1):1103–1204

Morris BJ, Croker S, Masnick AM, et al. (2012) The emergence of scientific reasoning. In: Kloos H, Morris BJ, Amaral JL (eds) Current Topics in Children's Learning and Cognition. IntechOpen, Rijeka, chap 4, https://doi.org/10.5772/53885, URL https://doi.org/10.5772/53885

de Moura L, Kong S, Avigad J, et al. (2015) The lean theorem prover (system description). In: Automated Deduction-CADE-25: 25th International Conference on Automated Deduction, Berlin, Germany, August 1-7, 2015, Proceedings 25, Springer, pp 378–388

Moura Ld, Ullrich S (2021) The lean 4 theorem prover and programming language. In: Automated Deduction–CADE 28: 28th International Conference on Automated Deduction, Virtual Event, July 12–15, 2021, Proceedings 28, Springer, pp 625–635

Mu Y, Zhang Q, Hu M, et al. (2023) Embodiedgpt: Vision-language pre-training via embodied chain of thought. arXiv preprint arXiv:230515021

Muennighoff N, Wang T, Sutawika L, et al. (2022) Crosslingual generalization through multitask finetuning. arXiv preprint arXiv:221101786

Mukherjee S, Mitra A, Jawahar G, et al. (2023) Orca: Progressive learning from complex explanation traces of gpt-4. arXiv preprint arXiv:230602707

Mündler N, He J, Jenko S, et al. (2023) Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. 2305.15852

Nascimento N, Alencar P, Cowan D (2023) Self-adaptive large language model (llm)-based multiagent systems. 2307.06187

Nelson B, Barreno M, Chi FJ, et al. (2008) Exploiting machine learning to subvert your spam filter. LEET 8(1-9):16–17

Nguyen E, Poli M, Faizi M, et al. (2023) Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. arXiv preprint arXiv:230615794

Nguyen T, Ilharco G, Wortsman M, et al. (2022) Quality not quantity: On the interaction between dataset design and robustness of clip. Advances in Neural Information Processing Systems 35:21455–21469

Nijkamp E, Pang B, Hayashi H, et al. (2022a) Codegen: An open large language model for code with multi-turn program synthesis. arXiv preprint arXiv:220313474

Nijkamp E, Ruffolo JA, Weinstein EN, et al. (2022b) Progen2: exploring the boundaries of protein language models. Cell Systems

Nijkamp E, Hayashi H, Xiong C, et al. (2023) Codegen2: Lessons for training llms on programming and natural languages. arXiv preprint arXiv:230502309

Ning X, Lin Z, Zhou Z, et al. (2023) Skeleton-of-thought: Large language models can do parallel decoding. 2307.15337

Nunes T (2012) Logical Reasoning and Learning, Springer US, Boston, MA, pp 2066–2069. https://doi.org/10.1007/978-1-4419-1428-6_790, URL https://doi.org/10.1007/978-1-4419-1428-6_790

Oberlander J, Cox R, Stenning K (1996) Proof styles in multimodal reasoning. In: Logic, Language and Computation. CSLI Publications, p 403–414

Odouard VV, Mitchell M (2022) Evaluating understanding on conceptual abstraction benchmarks. arXiv preprint arXiv:220614187

Olausson TX, Inala JP, Wang C, et al. (2023) Demystifying gpt self-repair for code generation. 2306.09896

Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:180703748

OpenAI (2023a) Gpt-4 technical report. 2303.08774

OpenAI (2023b) Gpt-4v(ision) system card. arXiv URL https://cdn.openai.com/papers/GPTV_System_Card.pdf

Ordonez V, Kulkarni G, Berg T (2011) Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems 24

Ouyang L, Wu J, Jiang X, et al. (2022) Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35:27730–27744

Padalkar A, Pooley A, Jain A, et al. (2023) Open x-embodiment: Robotic learning datasets and rt-x models. arXiv preprint arXiv:231008864

Pan B, Sun J, Leung HYT, et al. (2020) Cross-view semantic segmentation for sensing surroundings. IEEE Robotics and Automation Letters 5(3):4867–4873. https://doi.org/10.1109/LRA.2020.3004325

Pan L, Albalak A, Wang X, et al. (2023a) Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. arXiv preprint arXiv:230512295

Pan Y, Pan L, Chen W, et al. (2023b) On the risk of misinformation pollution with large language models. 2305.13661

Papineni K, Roukos S, Ward T, et al. (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp 311–318

Paranjape B, Lundberg S, Singh S, et al. (2023) Art: Automatic multi-step reasoning and tool-use for large language models. 2303.09014

Pasupat P, Liang P (2015) Compositional semantic parsing on semi-structured tables. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Beijing, China, pp 1470–1480, https://doi.org/10.3115/v1/P15-1142, URL https://aclanthology.org/P15-1142

Patel A, Bhattamishra S, Goyal N (2021a) 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. Association for Computational Linguistics, Online, pp 2080–2094, https://doi.org/10.18653/v1/2021.naacl-main.168, URL https://aclanthology.org/2021.naacl-main.168

Patel A, Bhattamishra S, Goyal N (2021b) Are nlp models really able to solve simple math word problems? 2103.07191

Patel A, Bhattamishra S, Goyal N (2021c) Are nlp models really able to solve simple math word problems? In: North American Chapter of the Association for Computational Linguistics

Paul D, Ismayilzada M, Peyrard M, et al. (2023) Refiner: Reasoning feedback on intermediate representations. 2304.01904

Paulson LC (1994) Isabelle: A generic theorem prover. Springer

Peng B, Alcaide E, Anthony Q, et al. (2023a) Rwkv: Reinventing rnns for the transformer era. arXiv preprint arXiv:230513048

Peng B, Galley M, He P, et al. (2023b) Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv preprint arXiv:230212813

Peng B, Li C, He P, et al. (2023c) Instruction tuning with gpt-4. arXiv preprint arXiv:230403277

Peng Z, Wang W, Dong L, et al. (2023d) Kosmos-2: Grounding multimodal large language models to the world. 2306.14824

Perez F, Ribeiro I (2022) Ignore previous prompt: Attack techniques for language models. arXiv preprint arXiv:221109527

Peters J, Janzing D, Schölkopf B (2017) Elements of causal inference: foundations and learning algorithms. The MIT Press

Pfeiffer J, Vulić I, Gurevych I, et al. (2020) Mad-x: An adapter-based framework for multi-task cross-lingual transfer. arXiv preprint arXiv:200500052

Pi R, Gao J, Diao S, et al. (2023) Detgpt: Detect what you need via reasoning. 2305.14167

Poli M, Massaroli S, Nguyen E, et al. (2023) Hyena hierarchy: Towards larger convolutional language models. arXiv preprint arXiv:230210866

Pollock JL (1987) Defeasible reasoning. Cognitive science 11(4):481–518

Pollock JL (1991) A theory of defeasible reasoning. International Journal of Intelligent Systems 6(1):33–54

Pollock JL (2009) A recursive semantics for defeasible reasoning. Argumentation in artificial intelligence pp 173–197

Polu S, Sutskever I (2020) Generative language modeling for automated theorem proving. arXiv preprint arXiv:200903393

Polu S, Han JM, Zheng K, et al. (2023) Formal mathematics statement curriculum learning. In: The Eleventh International Conference on Learning Representations

Pratap V, Xu Q, Sriram A, et al. (2020) MLS: A large-scale multilingual dataset for speech research. In: INTERSPEECH, pp 2757–2761

Press O, Zhang M, Min S, et al. (2023) Measuring and narrowing the compositionality gap in language models. 2210.03350

Pryor C, Dickens C, Augustine E, et al. (2023) Neupsl: Neural probabilistic soft logic. 2205.14268

Puig X, Ra K, Boben M, et al. (2018) Virtualhome: Simulating household activities via programs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8494–8502

Qian C, Han C, Fung YR, et al. (2023a) Creator: Disentangling abstract and concrete reasonings of large language models through tool creation. 2305.14318

Qian T, Chen J, Zhuo L, et al. (2023b) Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. arXiv preprint arXiv:230514836

Qiao S, Ou Y, Zhang N, et al. (2022) Reasoning with language model prompting: A survey. arXiv preprint arXiv:221209597

Qiao S, Gui H, Chen H, et al. (2023) Making language models better tool learners with execution feedback. 2305.13068

Qin J, Lin L, Liang X, et al. (2020) Semantically-aligned universal tree-structured solver for math word problems. ArXiv abs/2010.06823

Qin Y, Hu S, Lin Y, et al. (2023) Tool learning with foundation models. 2304.08354

Qiu J, Chen L, Gu X, et al. (2022) Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion. IEEE Robotics and Automation Letters 7(4):8799–8806

Qiu J, Li L, Sun J, et al. (2023a) Large ai models in health informatics: Applications, challenges, and the future. IEEE Journal of Biomedical and Health Informatics

Qiu J, Wu J, Wei H, et al. (2023b) Visionfm: a multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence. arXiv preprint arXiv:231004992

Radford A, Narasimhan K, Salimans T, et al. (2018) Improving language understanding by generative pre-training. arXiv

Radford A, Wu J, Child R, et al. (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9

Radford A, Kim JW, Hallacy C, et al. (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PMLR, pp 8748–8763

Rae JW, Borgeaud S, Cai T, et al. (2021) Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:211211446

Rafailov R, Sharma A, Mitchell E, et al. (2023) Direct preference optimization: Your language model is secretly a reward model. arXiv preprint arXiv:230518290

Raffel C, Shazeer N, Roberts A, et al. (2019) Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv e-prints arXiv:1910.10683

Raheja V, Kumar D, Koo R, et al. (2023) Coedit: Text editing by task-specific instruction tuning. arXiv arXiv:2305.09857 [cs.CL]

Rajani NF, McCann B, Xiong C, et al. (2019) Explain yourself! leveraging language models for commonsense reasoning. In: Proceedings of the 2019 Conference of the Association for Computational Linguistics (ACL2019), URL https://arxiv.org/abs/1906.02361

Rajani NF, Zhang R, Tan YC, et al. (2020) ESPRIT: Explaining solutions to physical reasoning tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 7906–7917, https://doi.org/10.18653/v1/2020.acl-main.706, URL https://aclanthology.org/2020.acl-main.706

Rajpurkar P, Zhang J, Lopyrev K, et al. (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pp 2383–2392, https://doi.org/10.18653/v1/D16-1264, URL https://aclanthology.org/D16-1264

Raven JC, Court J (1938) Raven's progressive matrices. Western Psychological Services Los Angeles, CA

Rawte V, Sheth A, Das A (2023) A survey of hallucination in large foundation models. 2309.05922

Redbooks I (2004) Practical Guide to the IBM Autonomic Computing Toolkit. IBM

Reed S, Zolna K, Parisotto E, et al. (2022) A generalist agent. arXiv preprint arXiv:220506175

Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:190810084

Reiter R (1975) Formal reasoning and language understanding system. In: Theoretical Issues in Natural Language Processing

Ren S, Zhu KQ (2023) Low-rank prune-and-factorize for language model compression. 2306.14152

Ren X, Zhou P, Meng X, et al. (2023) Pangu-*sigma*: Towards trillion parameter language model with sparse heterogeneous computing. arXiv preprint arXiv:230310845

Ridnik T, Ben-Baruch E, Noy A, et al. (2021) Imagenet-21k pretraining for the masses. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)

Roller S, Dinan E, Goyal N, et al. (2020) Recipes for building an open-domain chatbot. arXiv preprint arXiv:200413637

Rombach R, Blattmann A, Lorenz D, et al. (2021) High-resolution image synthesis with latent diffusion models. 2112.10752

Roziere B, Gehring J, Gloeckle F, et al. (2023) Code llama: Open foundation models for code. arXiv preprint arXiv:230812950

Rubin O, Herzig J, Berant J (2022) Learning to retrieve prompts for in-context learning. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 2655–2671

Rudinger R, Shwartz V, Hwang JD, et al. (2020) Thinking like a skeptic: Defeasible inference in natural language. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 4661–4675, https://doi.org/10.18653/v1/2020.findings-emnlp.418, URL https://aclanthology.org/2020.findings-emnlp.418

Sachan M, Xing E (2017) Learning to solve geometry problems from natural language demonstrations in textbooks. In: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017). Association for Computational Linguistics, Vancouver, Canada, pp 251–261, https://doi.org/10.18653/v1/S17-1029, URL https://aclanthology.org/S17-1029

Sachan M, Dubey K, Xing E (2017) From textbooks to knowledge: A case study in harvesting axiomatic knowledge from textbooks to solve geometry problems. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp 773–784, https://doi.org/10.18653/v1/D17-1081, URL https://aclanthology.org/D17-1081

Sakaguchi K, Bras RL, Bhagavatula C, et al. (2021) Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM 64(9):99–106

Salmon W, Salmon M, Kitcher P (1989) Scientific explanation. Minneapolis

Sanh V, Webson A, Raffel C, et al. (2021) Multitask prompted training enables zero-shot task generalization. 2110.08207

Sap M, Rashkin H, Chen D, et al. (2019) Social IQa: Commonsense reasoning about social interactions. 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). Association for Computational Linguistics, Hong Kong, China, pp 4463–4473, https://doi.org/10.18653/v1/D19-1454, URL https://aclanthology.org/D19-1454

Saparov A, He H (2023) Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In: The Eleventh International Conference on Learning

Representations, URL https://openreview.net/forum?id=qFVVBzXxR2V

Savage N (2023) Drug discovery companies are customizing chatgpt: here's how. Nature Biotechnology

Sawada T, Paleka D, Havrilla A, et al. (2023) Arb: Advanced reasoning benchmark for large language models. 2307.13692

Scao TL, Fan A, Akiki C, et al. (2022) Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:221105100

Schick T, Dwivedi-Yu J, Jiang Z, et al. (2022) Peer: A collaborative language model. arXiv preprint arXiv:220811663

Schick T, Dwivedi-Yu J, Dessì R, et al. (2023) Toolformer: Language models can teach themselves to use tools. 2302.04761

Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural networks 61:85–117

Schuhmann C, Vencu R, Beaumont R, et al. (2021) Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:211102114

Schuhmann C, Beaumont R, Vencu R, et al. (2022) Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems 35:25278–25294

Schulman J, Wolski F, Dhariwal P, et al. (2017) Proximal policy optimization algorithms. arXiv preprint arXiv:170706347

Schölkopf B, Locatello F, Bauer S, et al. (2021) Toward causal representation learning. Proceedings of the IEEE 109(5):612–634. https://doi.org/10.1109/JPROC.2021.3058954

Seff A, Cera B, Chen D, et al. (2023) MotionLM: Multi-Agent Motion Forecasting as Language Modeling. In: ICCV

Seo M, Hajishirzi H, Farhadi A, et al. (2015) Solving geometry problems: Combining text and diagram interpretation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, pp 1466–1476, https://doi.org/10.18653/v1/D15-1171, URL https://aclanthology.org/D15-1171

Sha H, Mu Y, Jiang Y, et al. (2023) Languagempc: Large language models as decision makers for autonomous driving. arXiv preprint arXiv:231003026

Shah D, Xu P, Lu Y, et al. (2021) Value function spaces: Skill-centric state abstractions for long-horizon reasoning. arXiv preprint arXiv:211103189

Shao Z, Yu Z, Wang M, et al. (2023) Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14974–14983

Sharma P, Torralba A, Andreas J (2021) Skill induction and planning with latent language. arXiv preprint arXiv:211001517

Shen J, Yin Y, Li L, et al. (2021a) Generate & rank: A multi-task framework for math word problems. In: Moens MF, Huang X, Specia L, et al. (eds) Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp 2269–2279, https://doi.org/10.18653/v1/2021.findings-emnlp.195, URL https://aclanthology.org/2021.findings-emnlp.195

Shen L, Ji S, Zhang X, et al. (2021b) Backdoor pre-trained models can transfer to all. arXiv preprint arXiv:211100197

Shen S, Li C, Hu X, et al. (2022) K-lite: Learning transferable visual models with external knowledge. Advances in Neural Information Processing Systems 35:15558–15573

Shen Y, Song K, Tan X, et al. (2023) Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. 2303.17580

Shi F, Suzgun M, Freitag M, et al. (2023) Language models are multilingual chain-of-thought reasoners. In: The Eleventh International Conference on Learning Representations, URL https://openreview.net/forum?id=fR3wGCk-IXp

Shi S, Wang Y, Lin CY, et al. (2015) Automatically solving number word problems by semantic parsing and reasoning. In: Conference on Empirical Methods in Natural Language Processing

Shi W, Shea R, Chen S, et al. (2022) Just fine-tune twice: Selective differential privacy for large language models. arXiv preprint arXiv:220407667

Shinn N, Cassano F, Labash B, et al. (2023) Reflexion: Language agents with verbal reinforcement learning. 2303.11366

Shreya G, Khapra MM (2022) A survey in adversarial defences and robustness in nlp. arXiv preprint arXiv:220306414

Silver D, Hubert T, Schrittwieser J, et al. (2018) A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science 362(6419):1140–1144

Singh I, Blukis V, Mousavian A, et al. (2023) Progprompt: Generating situated robot task plans using large language models. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 11523–11530

Singh M, Gustafson L, Adcock A, et al. (2022) Revisiting weakly supervised pre-training of visual perception models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 804–814

Singhal K, Tu T, Gottweis J, et al. (2023) Towards expert-level medical question answering with large language models. arXiv preprint arXiv:230509617

Sinha K, Sodhani S, Dong J, et al. (2019) Clutrr: A diagnostic benchmark for inductive reasoning from text. arXiv preprint arXiv:190806177

Soldaini L, Lo K (2023) peS2o (Pretraining Efficiently on S2ORC) Dataset. Tech. rep., Allen Institute for AI, oDC-By, https://github.com/allenai/pes2o

Song CH, Wu J, Washington C, et al. (2023a) Llm-planner: Few-shot grounded planning for embodied agents with large language models. 2212.04088

Song F, Yu B, Li M, et al. (2023b) Preference ranking optimization for human alignment. arXiv preprint arXiv:230617492

Song X, Shi Y, Chen X, et al. (2018) Explore multi-step reasoning in video question answering. In: Proceedings of the 26th ACM international conference on Multimedia, pp 239–247

Sowa JF (2003) Laws, facts, and contexts: Foundations for multimodal reasoning. In: Knowledge Contributors. Springer, p 145–184

Speer R, Chin J, Havasi C (2017) Conceptnet 5.5: An open multilingual graph of general knowledge. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI Press, AAAI'17, p 4444–4451

Srinivasan K, Raman K, Chen J, et al. (2021) Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 2443–2449

Srivastava A, Rastogi A, Rao A, et al. (2023) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research URL https://openreview.net/forum?id=uyTL5Bvosj

Stolfo A, Jin Z, Shridhar K, et al. (2022) A causal framework to quantify the robustness of mathematical reasoning with language models. arXiv preprint arXiv:221012023

Strehl A, Ghosh J (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research 3(Dec):583–617

Subramanian S, Harrington P, Keutzer K, et al. (2023) Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. 2306.00258

Sun A, Ma P, Yuan Y, et al. (2023a) Explain any concept: Segment anything meets concept-based explanation. 2305.10289

Sun C, Shrivastava A, Singh S, et al. (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision, pp 843–852

Sun J, Sun H, Han T, et al. (2021) Neuro-symbolic program search for autonomous driving decision module design. In: Kober J, Ramos F, Tomlin C (eds) Proceedings of the 2020 Conference on Robot Learning, Proceedings of Machine Learning Research, vol 155. PMLR, pp 21–30, URL https://proceedings.mlr.press/v155/sun21a.html

Sun J, Huang DA, Lu B, et al. (2022a) Plate: Visually-grounded planning with transformers in procedural tasks. IEEE Robotics and Automation Letters 7(2):4924–4930

Sun J, Kousik S, Fridovich-Keil D, et al. (2022b) Self-supervised traffic advisors: Distributed, multi-view traffic prediction for smart cities. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), pp 917–922, https://doi.org/10.1109/ITSC55140.2022.9922340

Sun J, Jiang Y, Qiu J, et al. (2023b) Conformal prediction for uncertainty-aware planning with diffusion dynamics model. In: Thirty-seventh Conference on Neural Information Processing Systems

Sun J, Kousik S, Fridovich-Keil D, et al. (2023c) Connected autonomous vehicle motion planning with video predictions from smart, self-supervised infrastructure. arXiv preprint arXiv:230907504

Sun M, Liu Z, Bair A, et al. (2023d) A simple and effective pruning approach for large language models. 2306.11695

Sun Y, Dong L, Huang S, et al. (2023e) Retentive network: A successor to transformer for large language models. arXiv preprint arXiv:230708621

Suvorov R, Logacheva E, Mashikhin A, et al. (2021) Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:210907161

Svyatkovskiy A, Deng SK, Fu S, et al. (2020) Intellicode compose: Code generation using transformer. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp 1433–1443

Szot A, Clegg A, Undersander E, et al. (2021) Habitat 2.0: Training home assistants to rearrange their habitat. In: Advances in Neural Information Processing Systems (NeurIPS)

Tafjord O, Dalvi B, Clark P (2021) ProofWriter: Generating implications, proofs, and abductive statements over natural language. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, Online, pp 3621–3634, https://doi.org/10.18653/v1/2021.findings-acl.317, URL https://aclanthology.org/2021.findings-acl.317

Tafjord O, Dalvi Mishra B, Clark P (2022) Entailer: Answering questions with faithful and truthful chains of reasoning. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 2078–2093, https://doi.org/10.18653/v1/2022.emnlp-main.134, URL https://aclanthology.org/2022.emnlp-main.134

Talmor A, Berant J (2018) The web as a knowledge-base for answering complex questions. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 641–651, https://doi.org/10.18653/v1/N18-1059, URL https://aclanthology.org/N18-1059

Talmor A, Herzig J, Lourie N, et al. (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). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4149–4158, https://doi.org/10.18653/v1/N19-1421, URL https://aclanthology.org/N19-1421

Talmor A, Yoran O, Catav A, et al. (2021) Multimodal{qa}: complex question answering over text, tables and images. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=ee6W5UgQLa

Tan S, Ivanovic B, Weng X, et al. (2023) Language conditioned traffic generation. CoRL

Tao C, Hou L, Zhang W, et al. (2022) Compression of generative pre-trained language models via quantization. 2203.10705

Tao M, Bao BK, Tang H, et al. (2023) Galip: Generative adversarial clips for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14214–14223

Taori R, Gulrajani I, Zhang T, et al. (2023) Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models https://crfm

stanford edu/2023/03/13/alpaca html 3(6):7

Taylor R, Kardas M, Cucurull G, et al. (2022) Galactica: A large language model for science. arXiv preprint arXiv:221109085

Team G (2022) GT4SD (Generative Toolkit for Scientific Discovery). URL https://github.com/GT4SD/gt4sd-core

Teig N, Scherer R (2016) Bringing formal and informal reasoning together—a new era of assessment? Frontiers in psychology 7:1097

Thoppilan R, De Freitas D, Hall J, et al. (2022) Lamda: Language models for dialog applications. arXiv preprint arXiv:220108239

Tian J, Li Y, Chen W, et al. (2022) Weakly supervised neural symbolic learning for cognitive tasks. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 5888–5896

Tomasic A, Romero OJ, Zimmerman J, et al. (2021) Propositional reasoning via neural transformer language models. arXiv

Tong Z, Song Y, Wang J, et al. (2022) Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. 2203.12602

Touvron H, Lavril T, Izacard G, et al. (2023a) Llama: Open and efficient foundation language models. 2302.13971

Touvron H, Martin L, Stone K, et al. (2023b) Llama 2: Open foundation and fine-tuned chat models. 2307.09288

Tsai HS, Chang HJ, Huang WC, et al. (2022) SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp 8479–8492

Tsimpoukelli M, Menick J, Cabi S, et al. (2021) Multimodal few-shot learning with frozen language models. In: Beygelzimer A, Dauphin Y, Liang P, et al. (eds) Advances in Neural Information Processing Systems, URL https://openreview.net/forum?id=WtmMyno9Tq2

Tu R, Zhang K, Bertilson B, et al. (2019) Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. Advances in Neural Information Processing Systems 32

Tu R, Ma C, Zhang C (2023a) Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. 2301.13819

Tu T, Azizi S, Driess D, et al. (2023b) Towards generalist biomedical ai. arXiv preprint arXiv:230714334

Tung HY, Ding M, Chen Z, et al. (2023) Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. arXiv preprint arXiv:230615668

Tunstall L, Von Werra L, Wolf T (2022) Natural language processing with transformers. " O'Reilly Media, Inc."

Upadhyay S, Chang MW (2015) Draw: A challenging and diverse algebra word problem set

Upadhyay S, Chang MW (2017) Annotating derivations: A new evaluation strategy and dataset for algebra word problems. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Association for Computational Linguistics, Valencia, Spain, pp 494–504, URL https://aclanthology.org/E17-1047

Valipour M, Rezagholizadeh M, Kobyzev I, et al. (2022) Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. arXiv preprint arXiv:221007558

Van Den Oord A, Vinyals O, et al. (2017) Neural discrete representation learning. Advances in neural information processing systems 30

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. Advances in neural information processing systems 30

Vedantam R, Zitnick CL, Parikh D (2015) Cider: Consensus-based image description evaluation. 1411.5726

Waldmann MR, Hagmayer Y (2013) Causal reasoning. arXiv

Wang B, Komatsuzaki A (2021) GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax

Wang D, Zhang J, Du B, et al. (2023a) Scaling-up remote sensing segmentation dataset with segment anything model. arXiv preprint arXiv:230502034

Wang G, Xie Y, Jiang Y, et al. (2023b) Voyager: An open-ended embodied agent with large language models. 2305.16291

Wang H, Yuan Y, Liu Z, et al. (2023c) Dt-solver: Automated theorem proving with dynamic-tree sampling guided by proof-level value function. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 12632–12646

Wang J, Liu Z, Zhao L, et al. (2023d) Review of large vision models and visual prompt engineering. arXiv preprint arXiv:230700855

Wang L, Huang B, Zhao Z, et al. (2023e) Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14549–14560

Wang L, Ma C, Feng X, et al. (2023f) A survey on large language model based autonomous agents. 2308.11432

Wang L, Xu W, Lan Y, et al. (2023g) Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. arXiv preprint arXiv:230504091

Wang L, Yang N, Wei F (2023h) Learning to retrieve in-context examples for large language models. arXiv preprint arXiv:230707164

Wang R, Jansen P, Côté MA, et al. (2022a) Scienceworld: Is your agent smarter than a 5th grader? arXiv preprint arXiv:220307540

Wang R, Zelikman E, Poesia G, et al. (2023i) Hypothesis search: Inductive reasoning with language models. 2309.05660

Wang T, Zhang J, Fei J, et al. (2023j) Caption anything: Interactive image description with diverse multimodal controls. arXiv preprint arXiv:230502677

Wang X, Xu X, Tong W, et al. (2021a) Inferbert: a transformer-based causal inference framework for enhancing pharmacovigilance. Frontiers in Artificial Intelligence 4:659622

Wang X, Caccia L, Ostapenko O, et al. (2023k) Guiding language model reasoning with planning tokens. arXiv preprint arXiv:231005707

Wang X, Gu R, Chen Z, et al. (2023l) Uni-rna: universal pre-trained models revolutionize rna research. bioRxiv pp 2023–07

Wang X, Hu Z, Lu P, et al. (2023m) Scibench: Evaluating college-level scientific problem-solving abilities of large language models. arXiv preprint arXiv:230500970

Wang X, Wei J, Schuurmans D, et al. (2023n) Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations, URL https://openreview.net/forum?id=1PL1NIMMrw

Wang Y, Liu X, Shi S (2017) Deep neural solver for math word problems. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp 845–854, https://doi.org/10.18653/v1/D17-1088, URL https://aclanthology.org/D17-1088

Wang Y, Wang W, Joty S, et al. (2021b) Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. arXiv preprint arXiv:210900859

Wang Y, Kordi Y, Mishra S, et al. (2022b) Self-instruct: Aligning language model with self generated instructions

Wang Y, Mishra S, Alipoormolabashi P, et al. (2022c) Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In: EMNLP

Wang Y, Mukherjee S, Liu X, et al. (2022d) Adamix: Mixture-of-adapter for parameter-efficient tuning of large language models. arXiv preprint arXiv:220512410 1(2):4

Wang Y, Le H, Gotmare AD, et al. (2023o) Codet5+: Open code large language models for code understanding and generation. arXiv preprint arXiv:230507922

Wang Y, Zhong W, Li L, et al. (2023p) Aligning large language models with human: A survey. arXiv preprint arXiv:230712966

Wang YR, Duan J, Fox D, et al. (2023q) Newton: Are large language models capable of physical reasoning? 2310.07018

Wang Z, Wohlwend J, Lei T (2020) Structured pruning of large language models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, https://doi.org/10.18653/v1/2020.emnlp-main.496, URL https://doi.org/10.18653%2Fv1%2F2020.emnlp-main.496

Wang Z, Cai S, Liu A, et al. (2023r) Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. 2302.01560

Wang Z, Zhang G, Yang K, et al. (2023s) Interactive natural language processing. arXiv preprint arXiv:230513246

Watson JL, Juergens D, Bennett NR, et al. (2023) De novo design of protein structure and function with rfdiffusion. Nature 620(7976):1089–1100

Wei J, Bosma M, Zhao VY, et al. (2021) Finetuned language models are zero-shot learners. arXiv preprint arXiv:210901652

Wei J, Tay Y, Bommasani R, et al. (2022a) Emergent abilities of large language models. 2206.07682

Wei J, Wang X, Schuurmans D, et al. (2022b) Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing

Systems 35:24824–24837

Weidinger L, Mellor J, Rauh M, et al. (2021) Ethical and social risks of harm from language models. 2112.04359

Welleck S, Lu X, West P, et al. (2022) Generating sequences by learning to self-correct. arXiv preprint arXiv:221100053

Weston J, Sukhbaatar S (2023) System 2 attention (is something you might need too). 2311.11829

Willig M, ZE ˇCEVI C M, Dhami DS, et al. (2023a) Causal parrots: Large language models may talk causality but are not causal. preprint

Willig M, Zečević M, Dhami DS, et al. (2023b) Probing for correlations of causal facts: Large language models and causality. URL https://openreview.net/forum?id=UPwzqPOs4-

Woodcock J, Larsen PG, Bicarregui J, et al. (2009) Formal methods: Practice and experience. ACM computing surveys (CSUR) 41(4):1–36

Wu C, Yin S, Qi W, et al. (2023a) Visual chatgpt: Talking, drawing and editing with visual foundation models. 2303.04671

Wu D, Han W, Wang T, et al. (2023b) Language prompt for autonomous driving. arXiv preprint arXiv:230904379

Wu H, Zhang Z, Zhang E, et al. (2023c) Q-bench: A benchmark for general-purpose foundation models on low-level vision. arXiv preprint arXiv:230914181

Wu M, Norrish M, Walder C, et al. (2021) Tacticzero: Learning to prove theorems from scratch with deep reinforcement learning. Advances in Neural Information Processing Systems 34:9330–9342

Wu M, Waheed A, Zhang C, et al. (2023d) Lamini-lm: A diverse herd of distilled models from large-scale instructions. CoRR abs/2304.14402. URL https://arxiv.org/abs/2304.14402, 2304.14402

Wu W, Timofeev A, Chen C, et al. (2023e) Mofi: Learning image representations from noisy entity annotated images. arXiv preprint arXiv:230607952

Wu Y, Jiang AQ, Li W, et al. (2022a) Autoformalization with large language models. In: Oh AH, Agarwal A, Belgrave D, et al. (eds) Advances in Neural Information Processing Systems, URL https://openreview.net/forum?id=IUikebJ1Bf0

Wu Z, Dvornik N, Greff K, et al. (2022b) Slotformer: Unsupervised visual dynamics simulation with object-centric models. arXiv preprint arXiv:221005861

Wu Z, Qiu L, Ross A, et al. (2023f) Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. 2307.02477

Wu Z, Wang Z, Xu X, et al. (2023g) Embodied task planning with large language models. arXiv preprint arXiv:230701848

Xi Z, Chen W, Guo X, et al. (2023) The rise and potential of large language model based agents: A survey. 2309.07864

Xia F, R. Zamir A, He ZY, et al. (2018) Gibson Env: real-world perception for embodied agents. In: Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on, IEEE

Xiao B, Wu H, Xu W, et al. (2023a) Florence-2: Advancing a unified representation for a variety of vision tasks. arXiv preprint arXiv:231106242

Xiao G, Tian Y, Chen B, et al. (2023b) Efficient streaming language models with attention sinks. arXiv preprint arXiv:230917453

Xie D, Wang R, Ma J, et al. (2023a) Edit everything: A text-guided generative system for images editing. 2304.14006

Xie E, Yao L, Shi H, et al. (2023b) Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. arXiv preprint arXiv:230406648

Xie Y, Kawaguchi K, Zhao Y, et al. (2023c) Decomposition enhances reasoning via self-evaluation guided decoding. 2305.00633

Xie Z, Sun S (2019) A goal-driven tree-structured neural model for math word problems. In: Ijcai, pp 5299–5305

Xin H, Wang H, Zheng C, et al. (2023) Lego-prover: Neural theorem proving with growing libraries. arXiv preprint arXiv:231000656

Xiong J, Li Z, Zheng C, et al. (2023a) Dq-lore: Dual queries with low rank approximation re-ranking for in-context learning. arXiv preprint arXiv:231002954 URL https://api.semanticscholar.org/CorpusID:263620351

Xiong J, Shen J, Yuan Y, et al. (2023b) Trigo: Benchmarking formal mathematical proof reduction for generative language models. arXiv preprint arXiv:231010180

Xu FF, Alon U, Neubig G, et al. (2022) A systematic evaluation of large language models of code. In: Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming, pp 1–10

Xu L, Hu H, Zhang X, et al. (2020) CLUE: A Chinese language understanding evaluation benchmark. In: Proceedings of the 28th International Conference on

Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online), pp 4762–4772, https://doi.org/10.18653/v1/2020.coling-main.419, URL https://aclanthology.org/2020.coling-main.419

Xu P, Shao W, Zhang K, et al. (2023a) Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. arXiv preprint arXiv:230609265

Xu R, Luo F, Zhang Z, et al. (2021a) Raise a child in large language model: Towards effective and generalizable fine-tuning. arXiv preprint arXiv:210905687

Xu R, Wang X, Wang T, et al. (2023b) Pointllm: Empowering large language models to understand point clouds. arXiv preprint arXiv:230816911

Xu S, Yang L, Kelly C, et al. (2023c) Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. arXiv preprint arXiv:230801317

Xu Y, Liu X, Cao X, et al. (2021b) Artificial intelligence: A powerful paradigm for scientific research. The Innovation 2(4):100179. https://doi.org/https://doi.org/10.1016/j.xinn.2021.100179, URL https://www.sciencedirect.com/science/article/pii/S2666675821001041

Xu Z, Zhang Y, Xie E, et al. (2023d) DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model. arXiv preprint arXiv:231001412

Yan Z, Zhang K, Zhou R, et al. (2023) Multimodal chatgpt for medical applications: an experimental study of gpt-4v. arXiv preprint arXiv:231019061

Yang H, Li P, Lam W (2022a) Parameter-efficient tuning by manipulating hidden states of pretrained language models for classification tasks. arXiv preprint arXiv:220404596

Yang H, Wang Y, Li P, et al. (2023a) Bridging the gap between pre-training and fine-tuning for commonsense generation. In: Findings of the Association for Computational Linguistics: EACL 2023, pp 376–383

Yang J, Prabhakar A, Narasimhan K, et al. (2023b) Intercode: Standardizing and benchmarking interactive coding with execution feedback. 2306.14898

Yang K, Deng J (2019) Learning to prove theorems via interacting with proof assistants. In: International Conference on Machine Learning, PMLR, pp 6984–6994

Yang K, Deng J (2021) Learning symbolic rules for reasoning in quasi-natural language. arXiv preprint arXiv:211112038

Yang K, Deng J, Chen D (2022b) Generating natural language proofs with verifier-guided search. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi,

United Arab Emirates, pp 89–105, https://doi.org/10.18653/v1/2022.emnlp-main.7, URL https://aclanthology.org/2022.emnlp-main.7

Yang K, Swope AM, Gu A, et al. (2023c) Leandojo: Theorem proving with retrieval-augmented language models. arXiv preprint arXiv:230615626

Yang S, Nachum O, Du Y, et al. (2023d) Foundation models for decision making: Problems, methods, and opportunities. arXiv preprint arXiv:230304129

Yang Sw, Chi PH, Chuang YS, et al. (2021) SUPERB: Speech Processing Universal PERformance Benchmark. In: INTERSPEECH, pp 1194–1198

Yang Z, Yang Y (2022) Decoupling features in hierarchical propagation for video object segmentation. Advances in Neural Information Processing Systems 35:36324–36336

Yang Z, Qi P, Zhang S, et al. (2018) HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp 2369–2380, https://doi.org/10.18653/v1/D18-1259, URL https://aclanthology.org/D18-1259

Yang Z, Dong L, Du X, et al. (2022c) Language models as inductive reasoners. arXiv preprint arXiv:221210923

Yang Z, Du X, Mao R, et al. (2023e) Logical reasoning over natural language as knowledge representation: A survey. arXiv preprint arXiv:230312023

Yang Z, Li L, Lin K, et al. (2023f) The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:230917421

Yao L, Huang R, Hou L, et al. (2021) Filip: Fine-grained interactive language-image pre-training. In: International Conference on Learning Representations

Yao L, Han J, Wen Y, et al. (2022a) Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. Advances in Neural Information Processing Systems 35:9125–9138

Yao L, Huang R, Hou L, et al. (2022b) FILIP: fine-grained interactive language-image pre-training. In: ICLR

Yao L, Han J, Liang X, et al. (2023a) Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 23497–23506

Yao S, Yu D, Zhao J, et al. (2023b) Tree of thoughts: Deliberate problem solving with large language models. 2305.10601

Yao S, Zhao J, Yu D, et al. (2023c) ReAct: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations (ICLR)

Yao Y, Li Z, Zhao H (2023d) Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. 2305.16582

Yasunaga M, Liang P (2020) Graph-based, self-supervised program repair from diagnostic feedback. In: International Conference on Machine Learning, PMLR, pp 10799–10808

Ye J, Wu Z, Feng J, et al. (2023a) Compositional exemplars for in-context learning. arXiv preprint arXiv:230205698

Ye Q, Lin BY, Ren X (2021) Crossfit: A few-shot learning challenge for cross-task generalization in nlp. arXiv preprint arXiv:210408835

Ye S, Xie Y, Chen D, et al. (2023b) Improving commonsense in vision-language models via knowledge graph riddles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2634–2645

Ye X, Iyer S, Celikyilmaz A, et al. (2022) Complementary explanations for effective in-context learning. arXiv preprint arXiv:221113892

Yi K, Gan C, Li Y, et al. (2019) Clevrer: Collision events for video representation and reasoning. arXiv preprint arXiv:191001442

Yih Wt, Richardson M, Meek C, et al. (2016) The value of semantic parse labeling for knowledge base question answering. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Berlin, Germany, pp 201–206, https://doi.org/10.18653/v1/P16-2033, URL https://aclanthology.org/P16-2033

Yin D, Liu X, Yin F, et al. (2023a) Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. arXiv preprint arXiv:230514327

Yin S, Fu C, Zhao S, et al. (2023b) A survey on multimodal large language models. 2306.13549

Yin Z, Sun Q, Chang C, et al. (2023c) Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In: The 2023 Conference on Empirical Methods in Natural Language Processing, URL https://openreview.net/forum?id=30kbnyD9hF

Yin Z, Sun Q, Guo Q, et al. (2023d) Do large language models know what they don't know? In: Rogers A, Boyd-Graber J, Okazaki N (eds) Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada, pp 8653–8665, https://doi.org/10.18653/v1/2023.

findings-acl.551, URL https://aclanthology.org/2023.findings-acl.551

Yin Z, Wang J, Cao J, et al. (2023e) Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. arXiv preprint arXiv:230606687

Yoneda T, Fang J, Li P, et al. (2023) Statler: State-maintaining language models for embodied reasoning. 2306.17840

Young N, Bao Q, Bensemann J, et al. (2022) AbductionRules: Training transformers to explain unexpected inputs. In: Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, Dublin, Ireland, pp 218–227, https://doi.org/10.18653/v1/2022.findings-acl.19, URL https://aclanthology.org/2022.findings-acl.19

Yu F, Zhang H, Tiwari P, et al. (2023a) Natural language reasoning, a survey. 2303.14725

Yu F, Zhang H, Wang B (2023b) Nature language reasoning, a survey. arXiv preprint arXiv:230314725

Yu L, Jiang W, Shi H, et al. (2023c) Metamath: Bootstrap your own mathematical questions for large language models. arXiv preprint arXiv:230912284

Yu S, Mo S, Ahn S, et al. (2021) Abstract reasoning via logic-guided generation. 2107.10493

Yu S, Wu P, Liang PP, et al. (2022) Pacs: A dataset for physical audiovisual commonsense reasoning. 2203.11130

Yu T, Zhang R, Yang K, et al. (2018) Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp 3911–3921, https://doi.org/10.18653/v1/D18-1425, URL https://aclanthology.org/D18-1425

Yu T, Feng R, Feng R, et al. (2023d) Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:230406790

Yuan L, Chen D, Chen YL, et al. (2021) Florence: A new foundation model for computer vision. arXiv preprint arXiv:211111432

Yuan Z, Yuan H, Li C, et al. (2023a) Scaling relationship on learning mathematical reasoning with large language models. arXiv preprint arXiv:230801825

Yuan Z, Yuan H, Tan C, et al. (2023b) Rrhf: Rank responses to align language models with human feedback without tears. arXiv preprint arXiv:230405302

Yue X, Qu X, Zhang G, et al. (2023) Mammoth: Building math generalist models through hybrid instruction tuning. arXiv preprint arXiv:230905653

Zaken EB, Ravfogel S, Goldberg Y (2021) Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:210610199

Zan D, Chen B, Yang D, et al. (2022) Cert: Continual pre-training on sketches for library-oriented code generation. arXiv preprint arXiv:220606888

Zan D, Chen B, Zhang F, et al. (2023) Large language models meet nl2code: A survey. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 7443–7464

Zelikman E, Wu Y, Mu J, et al. (2022) STar: Bootstrapping reasoning with reasoning. In: Oh AH, Agarwal A, Belgrave D, et al. (eds) Advances in Neural Information Processing Systems, URL https://openreview.net/forum?id=_3ELRdg2sgI

Zellers R, Bisk Y, Schwartz R, et al. (2018) Swag: A large-scale adversarial dataset for grounded commonsense inference. arXiv preprint arXiv:180805326

Zellers R, Holtzman A, Bisk Y, et al. (2019) Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:190507830

Zeng A, Liu X, Du Z, et al. (2022) Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:221002414

Zeng A, Attarian M, brian ichter, et al. (2023) Socratic models: Composing zero-shot multimodal reasoning with language. In: The Eleventh International Conference on Learning Representations, URL https://openreview.net/forum?id=G2Q2Mh3avow

Zeng W, Ren X, Su T, et al. (2021) Pangu-\$alpha$: Large-scale autoregressive pre-trained chinese language models with auto-parallel computation. arXiv preprint arXiv:210412369

Zhai X, Wang X, Mustafa B, et al. (2022) Lit: Zero-shot transfer with locked-image text tuning. In: CVPR, pp 18102–18112

Zhang C, Bauer S, Bennett P, et al. (2023a) Understanding causality with large language models: Feasibility and opportunities. arXiv preprint arXiv:230405524

Zhang G, Shi Y, Liu R, et al. (2023b) Chinese open instruction generalist: A preliminary release. arXiv preprint arXiv:230407987

Zhang H, Du W, Shan J, et al. (2023c) Building cooperative embodied agents modularly with large language models. 2307.02485

Zhang J, Zhou Z, Mai G, et al. (2023d) Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. arXiv preprint

arXiv:230410597

Zhang K, Li Z, Li J, et al. (2023e) Self-edit: Fault-aware code editor for code generation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp 769–787, URL https://aclanthology.org/2023.acl-long.45

Zhang K, Yu J, Yan Z, et al. (2023f) Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. 2305.17100

Zhang P, Li X, Hu X, et al. (2021) Vinvl: Making visual representations matter in vision-language models. arXiv preprint arXiv:210100529

Zhang Q, Chen M, Bukharin A, et al. (2023g) Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv:230310512

Zhang R, Han J, Zhou A, et al. (2023h) Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:230316199

Zhang S, Roller S, Goyal N, et al. (2022a) Opt: Open pre-trained transformer language models. arXiv preprint arXiv:220501068

Zhang X, Wang L, Helwig J, et al. (2023i) Artificial intelligence for science in quantum, atomistic, and continuum systems. 2307.08423

Zhang Y, Dai H, Kozareva Z, et al. (2017) Variational reasoning for question answering with knowledge graph. In: AAAI Conference on Artificial Intelligence

Zhang Z, Zhang A, Li M, et al. (2022b) Automatic chain of thought prompting in large language models. arXiv preprint arXiv:221003493

Zhao H, Wang K, Yu M, et al. (2023a) Explicit planning helps language models in logical reasoning. 2303.15714

Zhao WX, Zhou K, Li J, et al. (2023b) A survey of large language models. arXiv preprint arXiv:230318223

Zhao X, Li W, Kong L (2023c) Decomposing the enigma: Subgoal-based demonstration learning for formal theorem proving. arXiv preprint arXiv:230516366

Zhao X, Xie Y, Kawaguchi K, et al. (2023d) Automatic model selection with large language models for reasoning. 2305.14333

Zhao Y, Khalman M, Joshi R, et al. (2022a) Calibrating sequence likelihood improves conditional language generation. In: The Eleventh International Conference on Learning Representations

Zhao Y, Li Y, Li C, et al. (2022b) MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 6588–6600, https://doi.org/10.18653/v1/2022.acl-long.454, URL https://aclanthology.org/2022.acl-long.454

Zhao Y, Pang T, Du C, et al. (2023e) On evaluating adversarial robustness of large vision-language models. arXiv preprint arXiv:230516934

Zhao Z, Lee WS, Hsu D (2023f) Large language models as commonsense knowledge for large-scale task planning. arXiv preprint arXiv:230514078

Zheng C, Liu Z, Xie E, et al. (2023a) Progressive-hint prompting improves reasoning in large language models. arXiv preprint arXiv:230409797

Zheng C, Wang H, Xie E, et al. (2023b) Lyra: Orchestrating dual correction in automated theorem proving. arXiv preprint arXiv:230915806

Zheng K, Han JM, Polu S (2021) Minif2f: a cross-system benchmark for formal olympiad-level mathematics. arXiv preprint arXiv:210900110

Zheng Q, Xia X, Zou X, et al. (2023c) Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. arXiv preprint arXiv:230317568

Zhong V, Xiong C, Socher R (2018) Seq2SQL: Generating structured queries from natural language using reinforcement learning. URL https://openreview.net/forum?id=Syx6bz-Ab

Zhong Z, Friedman D, Chen D (2021) Factual probing is [MASK]: Learning vs. learning to recall. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, pp 5017–5033, https://doi.org/10.18653/v1/2021.naacl-main.398, URL https://aclanthology.org/2021.naacl-main.398

Zhou D, Schärli N, Hou L, et al. (2023a) Least-to-most prompting enables complex reasoning in large language models. In: The Eleventh International Conference on Learning Representations, URL https://openreview.net/forum?id=WZH7099tgfM

Zhou H, Ding M, Peng W, et al. (2023b) Generalizable long-horizon manipulations with large language models. arXiv preprint arXiv:231002264

Zhou J, Hu S, Lv X, et al. (2020) Kacc: A multi-task benchmark for knowledge abstraction, concretization and completion. arXiv preprint arXiv:200413631

Zhou J, Zhang Y, Luo Q, et al. (2023c) Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In: Proceedings of

the 2023 CHI Conference on Human Factors in Computing Systems, pp 1–20

Zhou K, Yang J, Loy CC, et al. (2022) Learning to prompt for vision-language models. International Journal of Computer Vision 130(9):2337–2348

Zhou L, Dai S, Chen L (2015) Learn to solve algebra word problems using quadratic programming. In: Conference on Empirical Methods in Natural Language Processing

Zhou X, Liu M, Zagar BL, et al. (2023d) Vision language models in autonomous driving and intelligent transportation systems. arXiv preprint arXiv:231014414

Zhou Y, Chia MA, Wagner SK, et al. (2023e) A foundation model for generalizable disease detection from retinal images. Nature pp 1–8

Zhu B, Sharma H, Frujeri FV, et al. (2023a) Fine-tuning language models with advantage-induced policy alignment. arXiv preprint arXiv:230602231

Zhu D, Chen J, Shen X, et al. (2023b) Minigpt-4: Enhancing vision-language understanding with advanced large language models. 2304.10592

Zhu F, Lei W, Huang Y, et al. (2021) Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. arXiv preprint arXiv:210507624

Zhu X, Li J, Liu Y, et al. (2023c) A survey on model compression for large language models. 2308.07633

Zhu Y, Yuan H, Wang S, et al. (2023d) Large language models for information retrieval: A survey. 2308.07107

Ziyu Z, Xiaojian M, Yixin C, et al. (2023) 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: ICCV

Zong Y, Aodha OM, Hospedales T (2023) Self-supervised multimodal learning: A survey. 2304.01008