A Survey of Reasoning with Foundation Models: Concepts, Methodologies, and Outlook

JIANKAI SUN, The Chinese University of Hong Kong, Hong Kong, Hong Kong
CHUANYANG ZHENG, The Chinese University of Hong Kong, Hong Kong, Hong Kong
ENZE XIE, The University of Hong Kong, Hong Kong, Hong Kong
ZHENGYING LIU, Noah Ark's Lab, Hong Kong, Hong Kong
RUIHANG CHU, The Chinese University of Hong Kong, Hong Kong, Hong Kong
JIANING QIU, The Chinese University of Hong Kong, Hong Kong, Hong Kong
JIAQI XU, The Chinese University of Hong Kong, Hong Kong, Hong Kong
MINGYU DING, The University of Hong Kong, Hong Kong, Hong Kong
HONGYANG LI, Shanghai AI Lab, Shanghai, China
MENGZHE GENG, The Chinese University of Hong Kong, Hong Kong, Hong Kong
YUE WU, Noah Ark's Lab, Hong Kong, Hong Kong

Authors' Contact Information: Jiankai Sun and Chuanyang Zheng, The Chinese University of Hong Kong, Hong Kong, Hong Kong; e-mails: 1155136477@link.cuhk.edu.hk, cyzheng21@link.cuhk.edu.hk; Enze Xie, The University of Hong Kong, Hong Kong, Hong Kong; e-mail: xieenze@connect.hku.hk; Zhengying Liu, Noah Ark's Lab, Hong Kong, Hong Kong; e-mail: liuzhengying2@huawei.com; Ruihang Chu, Jianing Qiu, and Jiaqi Xu, The Chinese University of Hong Kong, Hong Kong, Hong Kong; e-mails: ruihangchu@link.cuhk.edu.hk, jianingqiu@cuhk.edu.hk, jiaqixuac@gmail.com; Mingyu Ding, The University of Hong Kong, Hong Kong, Hong Kong; e-mail: myding@berkeley.edu; Hongyang Li, Shanghai AI Lab, Shanghai, China; e-mail: hy@opendrivelab.com; Mengzhe Geng, The Chinese University of Hong Kong, Hong Kong, Hong Kong; e-mail: mzgeng@link.cuhk.edu.hk; Yue Wu, Noah Ark's Lab, Hong Kong, Hong Kong; e-mail: yue.wu@connect.ust.hk; Wenhai Wang, The Chinese University of Hong Kong, Hong Kong, Hong Kong; e-mail: whwang@ie.cuhk.edu.hk; Junsong Chen, Dalian University of Technology, Dalian, Liaoning, China; e-mail: jschen@mail.dlut.edu.cn; Zhangyue Yin, Fudan University, Shanghai, Shanghai, China; e-mail: yinzy21@m.fudan.edu.cn; Xiaozhe Ren, Noah Ark's Lab, Hong Kong, China; e-mail: renxiaozhe@huawei.com; Jie Fu and Junxian He, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong; e-mails: jiefu@ust.hk, junxianh@ust.hk; Yuan Wu, The Chinese University of Hong Kong, Hong Kong, Hong Kong; e-mail: wyuan@cuhk.edu.hk; Qi Liu and Xihui Liu, The University of Hong Kong, Hong Kong, Hong Kong; e-mails: liuqi@cs.hku.hk, xihuiliu@eee.hku.hk; Yu Li, The Chinese University of Hong Kong, Hong Kong, Hong Kong; e-mail: yuli@cuhk.edu.hk; Hao Dong, Peking University, Beijing, Beijing, China; e-mail: dhsig552@163.com; Yu Cheng, The Chinese University of Hong Kong, Hong Kong, Hong Kong; e-mail: chengyu05@gmail.com; Ming Zhang, Peking University, Beijing, Beijing, China; e-mail: mzhang_cs@pku.edu.cn; Pheng Ann Heng, The Chinese University of Hong Kong, Hong Kong, Hong Kong; e-mail: pheng@cse.cuhk.edu.hk; Jifeng Dai, Tsinghua University, Beijing, Beijing, China; e-mail: daijifeng001@gmail.com; Ping Luo, The University of Hong Kong, Hong Kong, Hong Kong; e-mail: pluo.lhi@gmail.com; Jingdong Wang, Hefei University of Technology, Hefei, Anhui, China; e-mail: welleast@gmail.com; Ji-Rong Wen, Renmin University of China, Beijing, China; e-mail: jrwen@ruc.edu.cn; Xipeng Qiu, School of Computer Science, Fudan University, Shanghai, China; e-mail: xpqiu@fudan.edu.cn; Yike Guo, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong; e-mail: yikeguo@ust.hk; Hui Xiong, The Hong Kong University of Science and Technology - Guangzhou Campus, Guangzhou, Guangdong, China; e-mail: xionghui@hkust-gz.edu.cn; Qun Liu, Noah's Ark Lab, Hong Kong, United States; e-mail: qun.liu@huawei.com; Zhenguo Li, Noah Ark's Lab, Hong Kong, Hong Kong; e-mail: Li.Zhenguo@huawei.com.

ACM 0360-0300/2025/06-ART278
https://doi.org/10.1145/3729218

WENHAI WANG, The Chinese University of Hong Kong, Hong Kong, Hong Kong
JUNSONG CHEN, Dalian University of Technology, Dalian, China
ZHANGYUE YIN, Fudan University, Shanghai, China
XIAOZHE REN, Noah Ark's Lab, Hong Kong, China
JIE FU, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong
JUNXIAN HE, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong
YUAN WU, The Chinese University of Hong Kong, Hong Kong, Hong Kong
QI LIU, The University of Hong Kong, Hong Kong, Hong Kong
XIHUI LIU, The University of Hong Kong, Hong Kong, Hong Kong
YU LI, The Chinese University of Hong Kong, Hong Kong, Hong Kong
HAO DONG, Peking University, Beijing, China
YU CHENG, The Chinese University of Hong Kong, Hong Kong, Hong Kong
MING ZHANG, Peking University, Beijing, China
PHENG ANN HENG, The Chinese University of Hong Kong, Hong Kong, Hong Kong
JIFENG DAI, Tsinghua University, Beijing, China
PING LUO, The University of Hong Kong, Hong Kong, Hong Kong
JINGDONG WANG, Hefei University of Technology, Hefei, China
JI-RONG WEN, Renmin University of China, Beijing, China
XIPENG QIU, School of Computer Science, Fudan University, Shanghai, China
YIKE GUO, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong
HUI XIONG, The Hong Kong University of Science and Technology - Guangzhou Campus, Guangzhou, China
QUN LIU, Noah's Ark Lab, Hong Kong, United States
ZHENGUO LI, Noah Ark's Lab, Hong Kong, Hong Kong

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 article, we introduce seminal foundation models proposed or adaptable for 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, e.g., Large Language Models (LLMs), and contribute to the development of AGI.

CCS Concepts: • **Computer methodologies → Nature language processing;**

Additional Key Words and Phrases: Reasoning, foundation models, multimodal, AI agent, artificial general intelligence

ACM Reference Format:
Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Yuan Wu, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. 2025. A Survey of Reasoning with Foundation Models: Concepts, Methodologies, and Outlook. *ACM Comput. Surv.* 57, 11, Article 278 (June 2025), 43 pages. https://doi.org/10.1145/3729218

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 [180]. *Thinking, Fast and Slow* [55] 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" [300].

Since their inception, foundation models [22] have demonstrated remarkable efficacy across various domains, including **natural language processing** (**NLP**) [217], computer vision [283], and multimodal tasks [140]. 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 [91], 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 1 presents an overview of reasoning with regard to tasks and techniques.

Foundation models typically consist of billions of parameters and undergo (pre-)training using self-supervised learning on a broad dataset [22]. 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 [22, 304]. In contrast to the soft attention 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 [183], these models can perform complex tasks such as deduction and inference. This allows them to make decisions based on explicit knowledge [183] and logical reasoning, rather than relying solely on statistical patterns [322]. 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 [277]. 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.

As Figure 1 shows, we provide a concise overview of various reasoning tasks, including **Commonsense Reasoning, Mathematical Reasoning, Logical Reasoning, Multimodal Reasoning,**

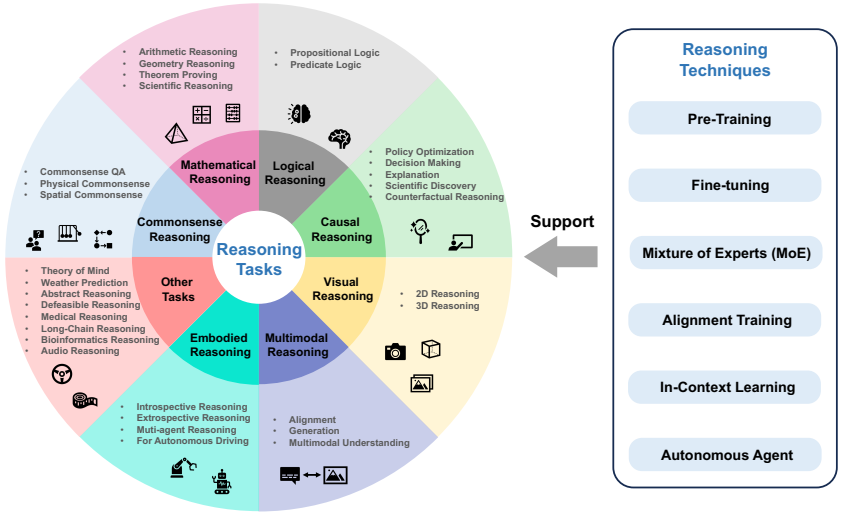

Fig. 1. Left: Overview of the reasoning tasks introduced in this survey, as detailed in Section 3. Right: Overview of the reasoning techniques for foundation models, as detailed in Section 2.

Embodied 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 review a range of techniques, tasks, and benchmarks for reasoning with foundation models. We also explore various application domains that can benefit from reasoning with foundation models, such as question-answering, automated reasoning, and knowledge representation. Additionally, we 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 article aims to provide a comprehensive understanding of reasoning with foundation models, its current state, and future possibilities.

2 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 2.1): Exploring data and architecture of reasoning foundation models.
— Fine-tuning (Section 2.2): Focusing on reasoning foundation models' fine-tuning data and techniques.
— Alignment Training (Section 2.3): Examining the alignment techniques employed by reasoning foundation models.
— **Mixture-of-Expert** (**MoE**) (Section 2.4): Introducing the MoE techniques in the context of reasoning.
— **In-Context Learning** (**ICL**) (Section 2.5): Introducing ICL in reasoning foundation models.
— Autonomous Agent (Section 2.6): Focusing on the reasoning foundation model as an agent for multiple tasks.

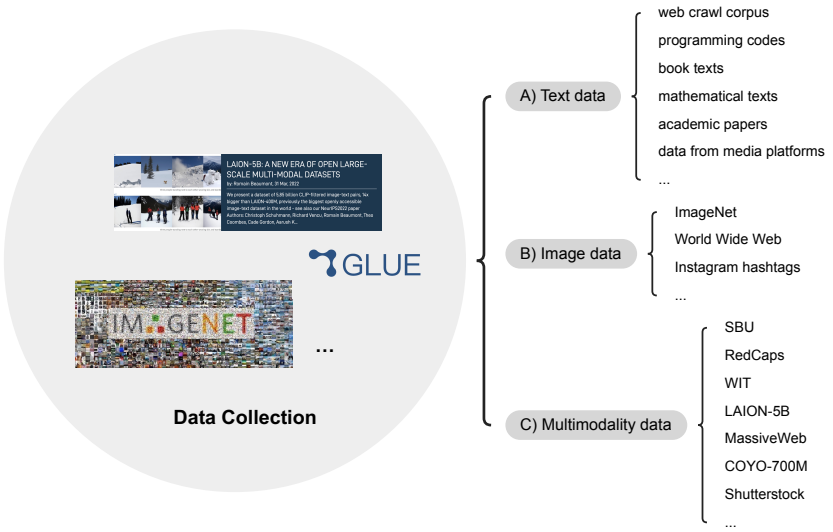

Fig. 2. A diverse suite of data sources and datasets for pre-training foundation models, mainly including text data, image data, and multimodality data.

2.1 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.

2.1.1 Data Source. Foundation models are data-driven, and both quality and quantity of data lie at the core of foundation model development. Figure 2 presents three broad types of data sources for foundation model pre-training.

Text Data. The expansion of publicly accessible large-scale text datasets has significantly enriched resources for various applications. Noteworthy examples include the Pile [79], a vast 825 GB English text corpus with 22 diverse subsets for training **language models** (**LMs**), and the C4 dataset [227], an enhanced Common Crawl web corpus widely used in many fields. The ROOTS dataset [135] stands out with 1.6TB of content in 59 languages, while the Gutenberg project [134] offers 3,036 cleaned English books. The CLUECorpus [314] is notable in Chinese text, and the Proof-Pile dataset [11] excels in mathematical text with 8 billion tokens. The peS2o dataset [256], comprising 40 million academic papers, is crucial for LM pre-training. Public conversation datasets, such as the Reddit corpus [235], provide valuable conversational content. Training LMs on extensive code corpora, as highlighted by recent research [10], improves program quality. The RedPajama project [51] reproduces LLaMA's training dataset with 1.2 trillion tokens from diverse sources, offering a broad resource for model development.

Image Data. Supervised pre-training with large, curated datasets like ImageNet [58] and ImageNet21K [234] has become key in developing visual representations. This method links images to labels for visual concepts. As demand for pre-training grows, using noisy labels from web-sourced image-text pairs has gained importance. Industrial labs have built vast datasets like JFT [261] and I2E [305] using semi-automatic methods and enriched them with sources like Instagram hashtags [254] to improve model accuracy. This approach has greatly enhanced visual recognition systems' ability to identify and categorize diverse visual concepts.

Multimodality Data. Large-scale datasets have become pivotal in advancing data-driven research. Notable examples include SBU [199], with 1 million Flickr-derived image-caption pairs, and RedCaps [59], featuring 12 million image-text pairs from Reddit. The WIT dataset [260] stands out with its 37.6 million image-text pairs across 108 Wikipedia languages, incorporating 11.5 million unique images. Other significant datasets are Shutterstock [196], LAION-400M [245], and COYO-700M [27], with OpenAI's CLIP [222] utilizing 400 million web-sourced image-text pairs. The LAION-5B dataset [244] represents a leap to billion-scale, containing 5.85 billion CLIP-filtered pairs, with 2.32 billion in English. DataComp [77] experiments with 12.8 billion image-text pairs from Common Crawl, and Flamingo [4] introduces the M3W dataset, with text and images from 43 million web pages. ImageBind [86] aims to create a joint embedding for six modalities, including images, text, and more, pointing toward a future where multimodal data comprehension is deepened. These datasets are vital for multimodal learning, pushing the boundaries of how diverse data types are understood and utilized.

2.1.2 *Network Architecture.*

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 [223], GPT-2 [224], and the notably advanced GPT-3 [24]. GPT-3, in particular, serves as a quintessential model within this paradigm, exemplifying the architectural efficacy, especially in ICL, a distinguishing feature of LLMs. The decoder-only architecture's influence transcends the GPT lineage, significantly impacting the broader field of LLMs. Numerous cutting-edge LMs have adopted this architectural framework as their foundational structure. For instance, OPT [344] employs the decoder-only architecture to achieve commendable natural language understanding capabilities. Gopher [225] 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 [242], which utilize its unidirectional information flow for tasks necessitating contextual comprehension. LLaMA [272] and its successor, LLaMA-2 [271], have integrated this architectural style to propel advancements in language modeling, achieving remarkable performances across various NLP benchmarks. GLM [341] 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.

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 [353]. The S4 model [90] 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 [185] emerges as a compelling alternative to the S4 and DSS [93] models, with the advantage of markedly faster training times. In contrast, H3 [74] 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 [210] offers a notable solution. This model is crafted by combining implicitly parametrized long convolutions with data-controlled gating, significantly diminishing computational requirements. RWKV [206] 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. LongNet [61] 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 [309] presents an efficient framework that allows 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.

2.2 Fine-Tuning

A fundamental strategy employed by 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.

2.2.1 *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 LM.

Synthesis Data. This section explores using LLMs to synthesize reasoning data, which is then used for model fine-tuning. Key to this is applying **Chain-of-Thought** (**CoT**) prompting to generate reasoning paths, with subsequent fine-tuning leveraging this generated data [75, 105, 112, 144, 178]. The Finetune-CoT method [101] samples multiple reasoning paths from LLMs for fine-tuning, while Distilling step-by-step [105] aims at training smaller models with less data. The Self-Improve approach [112] uses rationale-augmented answers for fine-tuning, and generating explanations for the training set is another strategy [144]. These methods show promise in improving reasoning tasks performance [178].

In mathematics, WizardMath [173] uses **Reinforcement Learning from Evol-Instruct Feedback** (**RLEIF**) for generating diverse math instruction data. MetaMath [336] employs question bootstrapping for training dataset augmentation. MAmmoTH introduces the MathInstruct dataset, integrating CoT and **Process-of-Thought** (**PoT**) rationales [339]. "Orca" [193] and "Orca2" [189] introduce explanation tuning and Prompt Erasing, respectively, for fine-tuning models with detailed explanations or generic prompts.

2.3 Alignment Training

The methodology of alignment training introduces an innovative approach that employs learning techniques to optimize LMs using human feedback directly. This concept has initiated a new

paradigm in which LMs are fine-tuned to correspond with intricate human values more closely. While LLMs can be prompted to execute a variety of 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 LMs in numerous domains.

2.3.1 Data Source. In this section, we explore the diverse data sources employed in alignment training. The efficacy of alignment techniques hinges on the quality, diversity, and scalability of the data used to train and refine models. Broadly, alignment training data can be categorized into two primary sources: (1) human-generated data and (2) synthetic or model-generated data.

Human Data. Databricks's "databricks-dolly-15k" dataset [52] includes 15,000 instructions, while the OpenAssistant corpus [133] features over 10,000 dialogues with contributions from 13,000+ annotators. UnifiedQA [127] is evaluated on 20 datasets, and CrossFit [327] serves as an NLP benchmark with 160 tasks. P3 [239] gathers 2,000+ prompts from 270+ datasets; MetaICL [187] experiments on 142 NLP datasets; ExMix [7] provides 107 tasks; and Natural Instructions [188] and Super-NaturalInstructions [292] offer 61 and 1.5k+ tasks, respectively. Flan 2022 [168] focuses on instruction tuning. xP3 [192] spans 46 languages and 16 tasks, supporting BLOOMZ and mT0 models. LongForm [132] selects 15,000 texts from C4 and Wikipedia. ShareGPT promotes sharing of ChatGPT/GPT4 conversations.

Synthesis Data. Gathering data from human sources can be a resource-intensive and time-consuming process. Given the remarkable success of 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.

Self-Instruct [290] and subsequent works like Alpaca [267] and its iterations [44, 207] utilize LLMs to create training data for RLHF. Instruction Backtranslation [148] uses self-augmentation and self-curation for generating instructions and responses. Unnatural Instructions [104] provides 64,000 innovative instructions, expanded to 240,000 instances through rephrasing. OPT-IML Bench [117] benchmarks **Instruction Meta-Learning** (**IML**) with 2,000 tasks, evaluating models on 52,000+ instructions. Koala [84] curates a diverse dataset from ChatGPT Distillation Data. GPT4All [6] contains around one million prompt-response pairs from a week in March 2023. Alpaca-GPT4 [207] includes 52,000 instruction-following examples in English and Chinese, enhanced with GPT-4 feedback. LaMini-LM [303] features 2.58 million pairs from GPT-3.5-Turbo, ensuring prompt diversity. CoEdIT [228] provides 82,000 text editing instruction pairs. UltraChat [64] offers a million-scale multi-turn instructional conversation dataset. CoT-Collection [128] augments CoT rationales with 1.88 million instances. Dynosaur [331] dynamically expands instruction tuning datasets from the Huggingface Datasets Platform.

A common method for enhancing LLMs to more accurately interpret and respond to human intentions through specific guidance is known as 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 3.

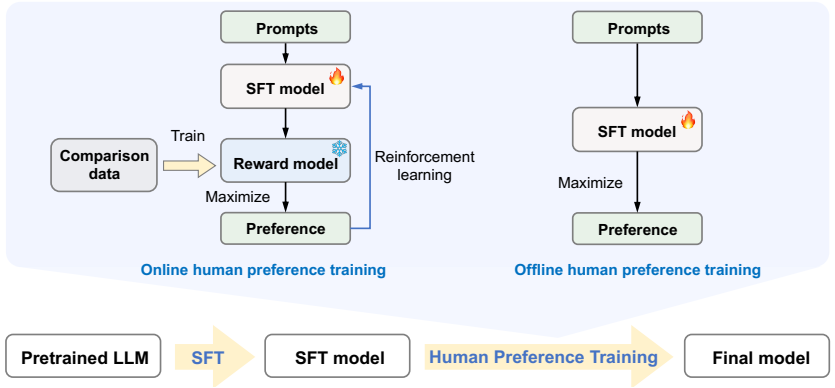

Fig. 3. The development process for LLM's alignment training. First, LLM is conventionally optimized via **Supervised Fine-Tuning** (**SFT**) using high-quality instruction data. Then, it may be further adjusted through Human Preference Training. The related techniques include online human preference training (left) that needs reinforcement learning and offline ones (right) that directly optimizes the policy to satisfy the preferences best.

2.3.2 Alignment Algorithm. In this section, we delve into the algorithmic frameworks underpinning alignment training. Among the diverse methodologies, two dominant paradigms have emerged: (1) online human preference training, which dynamically incorporates real-time human feedback during training, and (2) offline human preference training, which relies on pre-collected human preference datasets.

Online Human Preference Training. RLHF [200] represents a strategy developed to interpret human preferences by incorporating additional reward models within the framework of **Proximal Policy Optimization** (**PPO**) [246]. RLHF is divided into three primary phases: (1) The initial stage includes the creation of a comprehensive set of guidelines and the application of SFT on pre-existing 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**) [65], 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.

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**) [226], which aims to implicitly optimize the same objective as existing RLHF algorithms. **Preference Ranking Optimization** (**PRO**) [258] takes this further by fine-tuning LLMs to better align with human preferences and introduces SFT training objectives for regularization. **Sequence Likelihood Calibration** (**SLiC**) [350] 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**) [338] aligns model probabilities of multiple responses with human preferences using ranking loss, providing a simpler yet effective alternative that retains the performance of the PPO algorithm.

2.4 Mixture of Experts (MoE)

The MoE model represents a sophisticated supervised learning framework consisting of an array of networks, each fine-tuned to process a specific segment of the complete training dataset [118]. In this architecture, individual examples are processed by their respective expert networks. The Sparsely-Gated MoE model [249] integrates thousands of feed-forward sub-networks and employs a selective mechanism to engage a sparse array of these experts for each data instance. This methodology culminates in a model with an astounding 137 billion parameters, assigning a singular expert to every example. The model achieves sparsity through a gating function that directs each input to the top-K experts, where K is at least 2. Expanding upon this concept, GShard [138] adapts the MoE paradigm for transformers by substituting each feed-forward layer with a pairwise MoE layer, equipped with a Top-2 gating network. In a different approach, Switch Transformers [71] refine the MoE's sparsity by selecting either the optimal experts or a single best expert (where K equals 1) for each input. Additionally, GaLM [67] leverages a sparsely activated MoE architecture to amplify model capacity while substantially reducing training costs compared to denser models. The largest variant of GaLM boasts a remarkable 1.2 trillion parameters, significantly surpassing GPT-3 in scale. MoE has also been effectively implemented to enhance the capabilities of vision models [40, 41]. Moreover, MoE finds application in network compression strategies. WideNet [316] represents a parameter-efficient method that utilizes parameter sharing for compression along the network's depth. To optimize modeling capacity, WideNet scales the model's width by replacing standard feed-forward networks with a MoE structure and incorporating distinct layer norms to effectively process diverse semantic representations. MoEBERT [362] adopts a similar strategy, transforming the feed-forward neural networks in a pre-trained model into multiple experts. This modification maintains the robust representational abilities of the pre-trained model while integrating layer-wise distillation during training. In inference, a single expert is activated to optimize performance.

2.5 In-Context Learning

ICL employs natural language prompts, including task descriptions and example demonstrations, to guide LMs toward generating desired outputs without explicit retraining, as outlined by Brown et al. [24]. This method involves selecting relevant task examples, organizing them into prompts, and inputting them alongside test instances for model processing. **Vision-Language Models** (**VLMs**) also leverage this technique for task execution [39]. ICL is distinct from instruction tuning, which modifies models directly, but both utilize natural language for task presentation. Interestingly, instruction tuning can enhance LLMs' zero-shot performance in ICL scenarios [48].

2.5.1 Demonstration Example Selection. The effectiveness of 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 LLMs. Two primary methods for demonstration selection are prevalent: heuristic approaches and LLM-based approaches, as explored in the works of Liu et al. [161] and Lee et al. [137].

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. [161] and Lee et al. [137]. 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. [139] and Hongjin et al. [103]. Moreover, research conducted by Ye et al. [328] takes into account both relevance and diversity in the demonstration selection process. Intriguingly, Complex CoT [76] advocates the inclusion of intricate examples that involve extensive reasoning steps, while Auto-CoT [346] 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 LMs 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 [147]. In a related vein, Rubin et al. [237] 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 [175] applies the EPR approach to a broader spectrum of evaluation tasks, encompassing QA, NLI, MathR, and BC. Within the context of ICL, **Compositional Exemplars for In-context Learning** (**CEIL**) [326] 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 [285] 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**) [146] 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 [312] utilize Dual Queries and Low-rank approximation Re-ranking to automatically select exemplars for ICL.

2.5.2 *Chain-of-Thought.*

Zero-Shot CoT. Zero-shot CoT [129] 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 [284] 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. CoT [299] has charted a significant course for enhancing the reasoning capabilities of LLMs by employing detailed reasoning paths as prompts. This directional trend has given rise to various CoT variants, such as least-to-most [356], complex CoT [76], and program-of-thought [38]. However, it is worth noting that all these methods require annotations, which impose practical limitations on their application. To address this constraint, Auto-CoT [346] proposes a novel approach that utilizes Zero-Shot-CoT [129] to generate CoT reasoning paths. Taking a step further, Tree-of-Thought [323] models the human thought process not only as a chain but also as a tree, whereas Graph-of-Thought [325] extends this concept to represent human thought processes as both chains and graphs. Additionally, Skeleton-of-Thought [197] 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 [152] employs a voting verifier to consolidate final answers derived from multiple reasoning paths. In a similar vein, the Self-Consistency method [289] 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 [76]. Furthermore, Model Selection [348] takes a different approach by sampling two answers via CoT and **Plan-of-Thought** (**PoT**) and then employing an LM to select the correct one. Instead of generating complete reasoning paths, Self-Evaluation Guided Decoding [310] 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 [2] progressively samples reasoning paths until predefined criteria are met. Two concurrent approaches related to Tree-of-Thought [167, 323] gradually sample reasoning steps rather than complete reasoning paths. Additionally, **Reasoning via Planning** (**RAP**) [95] repurposes the LLM as both a world model and a reasoning agent. It incorporates a principled planning algorithm, based on **Monte Carlo Tree Search** (**MCTS**), to facilitate strategic exploration within the extensive reasoning space. Exchange-of-Thought [333] and X-of-Thoughts [164] introduce a variety of external reasoning insights and reasoning methods to enhance reasoning performance.

2.6 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 LLMs to control and orchestrate various tools [295], including web browsers and code interpreters, to complete their designated tasks.

VISPROG [94] offers a neuro-symbolic approach for generating interpretable Python-like programs for visual tasks without specific training. ToolFormer [243] self-supervisedly selects and uses APIs for token prediction. CAMEL [141] uses inception prompting in a communicative agent framework for task achievement. GPT4Tools [320], HuggingGPT [251], and Chameleon [171] enhance LLMs with multimodal and tool-augmented capabilities for complex tasks. TRICE [216], ChatCoT [43], and MultiTool-CoT [116] propose frameworks for tool integration and CoT reasoning. AssistGPT [78], OpenAGI [82], and ToolkenGPT [96] introduce novel approaches for tool usage, planning, and action execution. AutoGPT [89], ReAct [324], Reflexion [252], and CREATOR [215] focus on problem-solving, reasoning, and tool creation. Voyager [280] and AutoAgents [31] explore agent-based learning and team-building for tasks. SwiftSage [155] combines behavior cloning with LLMs for efficient problem-solving, showcasing diverse methods to boost LLM capabilities.

3 Reasoning Tasks

In this section, we provide a concise overview of various reasoning tasks, as Figure 1 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.

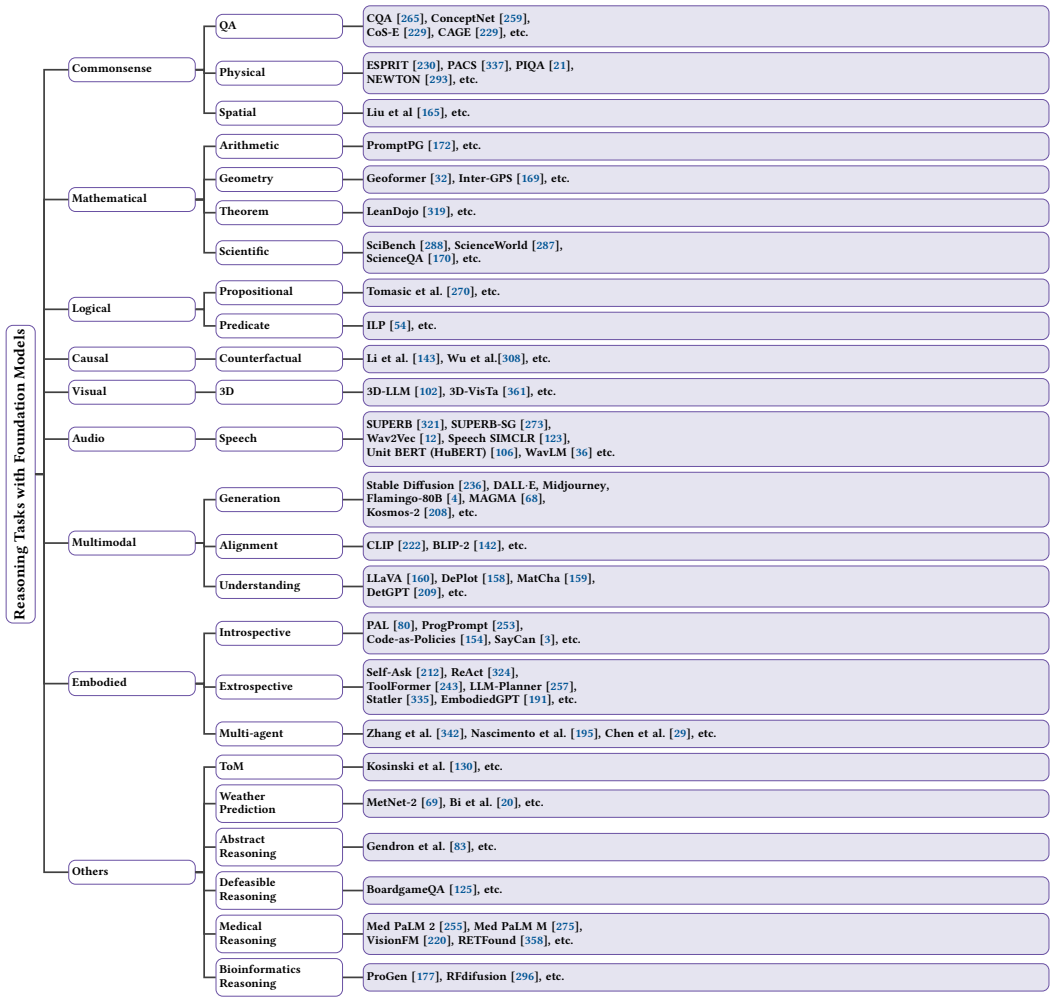

Fig. 4. Taxonomy of Reasoning Tasks with Foundation Models. Only the representative approaches for each type of task are listed.

— Multimodal Reasoning (Section 3.4): Involving reasoning across multiple data modalities, such as text, images, and sensory information.
— Embodied Reasoning (Section 3.5): Exploring reasoning in the context of embodied agents interacting with their environment.

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 4. There are also many other forms of reasoning (e.g., semantic reasoning [269], temporal reasoning [278], automatic planning [201]).

Key Observations: Recent performance evaluations across major LM benchmarks reveal significant advancements in model capabilities. Our analysis in Table 1 demonstrates that leading models like OpenAI o1 and Claude 3.5 Sonnet have achieved exceptional results across diverse tasks, with OpenAI o1 setting new standards in multitask language understanding (92.30% on MMLU [98]) and

Table 1. Comparison of Model Performance Across Various Reasoning Tasks, Including Mathematical Reasoning (MATH [99], FrontierMath [88]), Multitask Understanding (MMLU [98]), Commonsense Reasoning (HellaSwag [340]), and Code Generation (HumanEval [35])

Model	MMLU [98]	HellaSwag [340]	HumanEval [35]	BBHard [264]	GSM8K [49]	Math [99]	FrontierMath [88]
Claude 3.5 Sonnet	88.70%	89.00%	92.00%	93.10%	96.40%	71.10%	-
Claude 3 Opus	86.80%	95.40%	84.90%	86.80%	95.00%	60.10%	-
Gemini 1.5 Pro [268]	81.90%	92.50%	71.90%	84.00%	91.70%	58.50%	-
Gemini Ultra	83.70%	87.80%	74.40%	83.60%	94.40%	53.20%	-
GPT-4 [1]	86.40%	95.30%	67.00%	83.10%	92.00%	52.90%	-
Llama 3 Instruct - 70B	82.00%	87.00%	81.70%	81.30%	93.00%	50.40%	-
Claude 3 Haiku	75.20%	85.90%	75.90%	73.70%	88.90%	38.90%	-
GPT-3.5	70.00%	85.50%	48.10%	66.60%	57.10%	34.10%	-
Mixtral 8x7B [121]	70.60%	84.40%	40.20%	60.76%	74.40%	28.40%	-
GPT-4o	88.70%	-	90.20%	-	-	76.60%	-
GPT-4o mini	82.00%	-	87.00%	-	-	70.20%	-
Llama 3 Instruct - 8B	68.40%	-	62.00%	61.00%	79.60%	30.00%	-
Grok 1.5	73.00%	-	63.00%	-	62.90%	23.90%	-
Mistral Large	81.20%	89.20%	45.10%	-	81.00%	45.00%	-
Gemini 1.5 Flash	78.90%	-	-	89.20%	-	67.70%	-
GPT-4T 2024-04-09	86.50%	-	-	87.60%	-	72.20%	-
OpenAI o1	92.30%	-	92.40%	-	-	94.80%	<2%
OpenAI o1-mini	85.20%	-	92.40%	-	-	90.00%	-
QWen2.5	85.20%		59.1%	86.3%	91.5%	62.1%	-
QWen Plus	-	-	86.6%	86.3%	95.8%	83.1%	-
DeepSeek 2.5	80.4%	-	89.0%	84.3%	95.1%	74.7%	-

mathematical reasoning (94.80% on Math [99]). While both models excel in code generation tasks (92.40% and 92.00% on HumanEval [35] respectively), the most notable progress is evident in complex reasoning tasks, where current models substantially outperform their 2023 predecessors. Claude 3.5 Sonnet's average performance of 88.38% surpasses GPT-4's 79.45%, highlighting rapid year-over-year improvements. Particularly noteworthy is GPT-4o's strong performance on FrontierMath [88] (76.60%), suggesting that advanced mathematical reasoning capabilities are becoming a key differentiator in model evaluation. Even scaled-down versions like OpenAI o1-mini (85.20%) and GPT-4T (86.50%) demonstrate competitive performance, indicating significant improvements in model efficiency. While these results show remarkable progress in solving many benchmark tasks, performance on complex reasoning datasets like FrontierMath suggests this area will likely drive future developments in LM capabilities and evaluation metrics.

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.[1]

Recent research indicates that LMs are capable of acquiring certain aspects of common sense knowledge [352]. In the domain of structured commonsense reasoning, Madaan et al. [176] tackle the task by generating a graph based on natural language input. They formalize this problem as a code generation challenge, utilizing LLMs that are prompted with code to construct the graph representation. Berglund et al. [18] also point out that LMs 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. [149] take a systematic approach to evaluate the performance of large pre-trained LMs on various commonsense benchmarks.

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

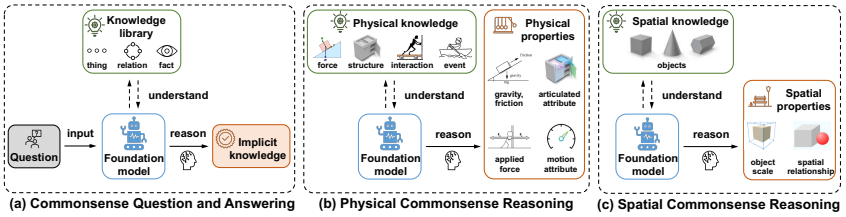

Fig. 5. Three areas of research of foundation models in commonsense reasoning. (a) By understanding everyday knowledge, foundation models can reason about implicit knowledge from questions and deduce answers. (b) Foundation models infer a wide range of physical properties from general physical knowledge. (c) Foundation models reason about spatial properties from a set of objects.

They conduct zero-shot and few-shot commonsense evaluations across four different benchmarks, considering six different model sizes. Notably, their evaluation includes a remarkably LLM 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 LMs with commonsense-specific fine-tuning techniques. Chang et al. [28] propose several architectural variations, leverage external commonsense corpora, and employ commonsense-specific fine-tuning techniques for the Social IQA task [240]. Through their work, they demonstrate that these optimizations can enhance the model's performance in tasks related to social intelligence. Furthermore, Yang et al. [317] 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 5. These areas of research contribute to a deeper understanding of how LMs can effectively capture and reason about commonsense knowledge in various contexts.

3.1.1 Commonsense Question and Answering. As a subfield of commonsense reasoning, commonsense 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 5(a).

The **Commonsense Question Answering** (**CommonsenseQA**) dataset [265] is a challenging multiple-choice dataset specifically designed for commonsense question answering. It is derived from ConceptNet [259] 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 [229] contains human commonsense explanations for the CommonsenseQA 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 [229] is a framework that involves training an LM 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 LMs, 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 [62], shown in Figure 5(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 [46].

Explaining Solutions to Physical ReasonIng Tasks (**ESPRIT**) framework [230] 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**) [337] 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**) [21] is a dataset that focuses on multiple-choice QA 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 [293] is a comprehensive platform that serves as a repository, pipeline, and benchmark specifically created to assess the physical reasoning capabilities of LLMs. CATER [87] mainly focuses on physics-related visual scenes. CLEVRER [330] is a video QA benchmark that targets the physical and causal relations grounded in dynamic videos of rigid-body collisions. CLEVRER-Humans [182] further extends it to the causal judgment of physical events with human labels. Physion [16], Physion++ [276], and ComPhy [42] 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 [60, 307] and neuro-symbolic frameworks with differentiable physics [62] are developed. **Aloe** (**Attention over Learned Object Embeddings**) [60] integrates MONet [26] for unsupervised object segmentation with self-attention mechanisms, facilitating spatio-temporal physical reasoning about objects. SlotFormer [307], 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 [119], **Generative Structured World Models** (**G-SWM**) [156] 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 5(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 [8] and spatial relationship [115]. Liu et al. [165] 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 VLMs 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 [72]. 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 [73].

Experimental findings suggest that the performance of LLMs shows a weak correlation with question difficulty. Ling et al. [157] 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 [351] is a specialized model designed to tackle the MultiHiertt dataset [351]. 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 LMs. While these issues may appear uncomplicated to humans, LMs frequently encounter challenges when it comes to tasks involving arithmetic reasoning [100].

Previous research has explored various approaches to address these challenges. Template-based statistical learning methods like KAZB [131], ZDC [357], and similarity-based method SIM [109] have been utilized. Wang et al. [291] employ a **Recurrent Neural Network** (**RNN**) to convert MWPs 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 [311] introduce an innovative neural approach to construct expression trees in a goal-oriented manner for solving MWPs. Shen et al. [250] introduce a novel ranking task for MWPs and present the Generate & Rank framework, which combines a generative pre-trained LM 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 CoT prompting, along with an LM containing an impressive 540 billion parameters, yields performance comparable to task-specific fine-tuned models across multiple tasks [299]. Unlike traditional symbolic reasoning tasks such as program synthesis and knowledge graph

reasoning, solving MWPs requires additional emphasis on numerical reasoning. PromptPG [172] 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 MWPs. This adaptive approach enhances the model's ability to handle numerical reasoning tasks effectively. Wang et al. [286] introduce MATH-SHEPHERD, a novel process-oriented math verifier that evaluates and assigns a reward score to each step in LLMs' solutions to math problems.

3.2.2 Geometry Reasoning. GeoS [247] provides a system for mapping geometry word problems into a logical representation, facilitating the process of problem-solving. Chen et al. [33] introduce **Neural Geometric Solver** (**NGS**) as an approach to addressing challenges posed by geometric problems in the GeoQA benchmark [33]. NGS adopts a holistic approach, adeptly parsing multi-modal information and generating interpretable programs. Geoformer [32] 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 [32]. This technique enhances the model's ability to handle mathematical expressions effectively. Inter-GPS [169] 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. In the context of hardware verification, it has found successful application in the design of integrated circuits [126]. In the realm of software verification, a notable achievement is the development of CompCert, a verified C compiler [17]. 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 [97]. Consequently, theorem proving has played a pivotal role in verifying floating-point firmware [97]. 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 [205], Lean [56], Coq [15], and Metamath [184]. In recent years, various approaches have integrated machine learning with ITPs [81]. Validated on various datasets (PISA [120], miniF2F [355], LeanDojo [319], and TRIGO [313]), these approaches leverage advancements in LMs [186, 211] 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 LM. Methods like MCTS [136] or dynamic-tree MCTS [282] are employed for this purpose. Previous work has demonstrated the few-shot statement autoformalization capability of LLMs [306, 329]. To investigate the applicability of these findings to proof autoformalization, DSP conducted a thorough analysis using Draft, Sketch, and Proof [122]. Subgoal-Learning [349] utilizes the subgoal-goal informal proof and demonstration selection. LeanDojo [319] is an open-source project for Lean [190], which contains toolkits, data, models, and benchmarks. Lyra [354] proposes the use of Tool Correction to mitigate LLM hallucinations and Conjecture Correction to improve the quality of generated formal proof

Table 2. Comparison between Propositional Logic and Predicate Logic in Terms of Basic Elements, Complexity, Expressive Power, and Applications

	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	$p \vee q$; $p \wedge q$; $\neg p$; $p \rightarrow q$	$\forall x, P(x)$; $\exists x, P(x)$

conjectures. Following the direction of Lyra, the LEGO-Prover [281] employs a growing skill library containing verified lemmas as skills to enhance the capability of LLMs used in theorem proving.

3.3 Logical Reasoning

Logical reasoning covers propositional and predicate logic (Table 2). While traditionally associated with deductive reasoning, which derives conclusions logically entailed by premises, it also includes abductive processes, which hypothesize premises that could lead to a desired or observed conclusion. It serves as a foundational basis across various domains, including computer science, mathematics, and medicine. For example, in diagnosis, a doctor observes symptoms and hypothesizes a disease as the cause [214]. Similarly, in automated planning, one hypothesizes sequences of actions to achieve pre-established goals [85].

Previous studies have explored the combination of neural networks and symbolic reasoning in neuro-symbolic methods [181, 213]. 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 [204] leverages LLMs and symbolic reasoning to enhance logical problem-solving [174]. 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. [25] 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 [298]. Zhao et al. [347] employ LMs 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. [53] 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. The first approach is ICL, where specific prompts are used to elicit step-by-step reasoning from LLMs. Notable methods in this category include CoT prompting [299] and the least-to-most prompting approach [356]. 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, where the reasoning capabilities of LLMs are optimized through fine-tuning or training specialized modules [318].

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).

In the context of propositional logic resolution, Tomasic et al. [270] 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 LMs, they aimed to enhance the logical reasoning capabilities in propositional logic problems.

The use of LMs 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 [54]. ILP leverages first-order logic to represent hypotheses and data, making logical language a crucial component in knowledge representation and reasoning [57].

By incorporating predicate logical representations and reasoning, LLMs offer the potential for more interpretable and explainable models [163]. It enables the discovery of logical patterns and rules from data, facilitating the extraction of human-understandable knowledge.

3.4 Multimodal Reasoning

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 [221, 332].[2]

In the pursuit of developing 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 LMs typically excel in 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**

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

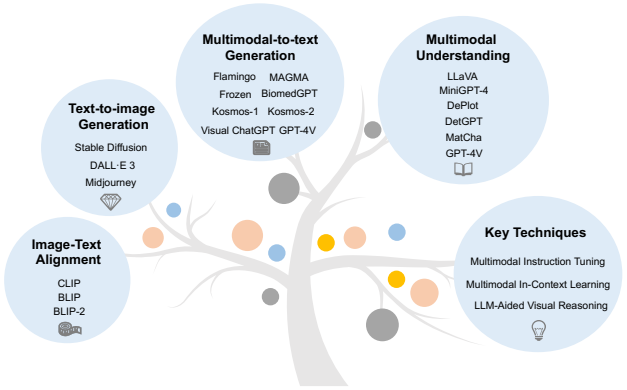

Fig. 6. Multimodal reasoning tasks can be broadly categorized into image-text alignment, T2I generation, multimodal-to-text generation, and multimodal understanding. Current multimodal foundation models mainly involve three key techniques to approach reasoning tasks, including M-IT, M-ICL, and LAVR. The figure style credits from tutorial [140].

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 6 shows multiple multimodal reasoning tasks and the key techniques behind, which are introduced as follows.

3.4.1 Alignment.

Image-Text Alignment. CLIP [222] 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 [142] 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 [4] bridges pretrained vision and language backbones by token fusion with cross-attentions.

3.4.2 Generation.

Text-to-image Generation. Stable Diffusion [236] 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[3] [19] 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[4] can translate textual prompts into visual compositions, providing a visual

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

representation of the given description. ImageGen [238] leverages the capabilities of expansive transformer LMs for text comprehension and combines this with the efficacy of diffusion models for creating high-quality images. PixArt [34] is a Transformer-driven **Text-to-Image** (**T2I**) diffusion model. It rivals leading image generation systems such as Imagen, SDXL, and Midjourney in terms of quality, approaching the benchmarks set by commercial applications.

Multimodal-to-text Generation. Flamingo-80B [4] 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 QA and captioning, as well as closed-ended tasks such as multiple-choice visual QA. Frozen [274] accomplishes few-shot learning ability within a multimodal context by preserving the language capabilities of an 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 [68] follows a similar approach to Frozen by incorporating a new image prefix encoder while keeping the LM frozen. It trains a series of VLMs capable of generating text autoregressively from combined visual and textual inputs. Visual ChatGPT [302] and GPT-4 [198] 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 [113] and Kosmos-2 [208]. 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 [343], which focuses specifically on biomedical research. These domain-specific adaptations aim to enhance the LM's performance and applicability within specialized fields.

3.4.3 Multimodal Understanding. Visual Instruction Tuning [160] 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 Vision Assistant**) [160] represents an extensively trained, large-scale multimodal model. It seamlessly integrates a vision encoder with Vicuna [44], facilitating versatile visual and language comprehension for general-purpose applications. LLaVA excels across a diverse spectrum of tasks necessitating multimodal understanding, encompassing visual QA, image captioning, and instruction-following. Notably, it achieves impressive performance on Science QA [170], a multimodal reasoning dataset in the science domain.

In the domain of reasoning on charts, DePlot [158] 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) [159] 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 [209] 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. LLaMA-VID [153] enhances LLMs for more efficient video and image understanding. It represents each video frame with two tokens, which decreases the burden of processing long videos without sacrificing essential information. To allow users to interactively control the focus of multimodal understanding, Prompt Highlighter [345] highlights specific prompt spans and effectively guides autoregressive generation to produce more targeted outputs.

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 [179].

3.5 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 [219]. While interaction is often emphasized as a tool for gathering information from and altering the environment, it is also important to recognize that interaction serves as an integral link between perception and action, forming a dynamic cycle. This cycle creates a feedback loop that enhances decision-making and adaptive behavior. Autonomous Agents in the context of LLMs 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 LLMs to autonomously perform tasks, utilizing their extensive knowledge, reasoning skills, and vast informational resources [5].

Several works have investigated the use of language for planning purposes [3, 108, 191]. 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 [3, 114]. Specifically, Huang et al. [114] prompt GPT-3 [24] and Codex [35] to create actions for agents, where each action step is semantically converted into a permissible action through Sentence-RoBERTa [166]. In contrast, SayCan [3] grounds the actions and language by combining the probability of each candidate action, as determined by FLAN [297], with the action's value function. The latter acts as a surrogate for measuring affordance [248]. 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 [155] 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 LM 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 LMs to perform more sophisticated reasoning tasks within the framework. Another noteworthy approach in this regard is RAP [95], which capitalizes on the LM's dual role as both a world model and a reasoning agent. RAP incorporates a well-founded planning algorithm, specifically based on MCTS, 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 [50]), and logical reasoning (e.g., PrOntoQA [241]). The evaluations demonstrate RAP's proficiency in addressing diverse reasoning challenges, effectively showcasing its versatility as a capable reasoning agent. MimicPlay [279] introduces a method for learning robotic policies from human play data, utilizing emergent human and video prompts to direct low-level visuomotor control.

Introspective Reasoning, Extrospective Reasoning, Embodied Reasoning, and Multiagent Reasoning, along with their interconnected aspects, play pivotal roles in the advancement of agent reasoning systems [218]. 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 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 [45], mapping [203], completing [47], grasping [151], planning [183], interaction [124], and control. Safety is a crucial aspect of embodied intelligent systems. In this context, PlanCP [262] suggests the application of conformal prediction to diffusion dynamic models.

3.5.1 Embodied Reasoning. Recent research has highlighted the successful application of LLMs in robotics domains [3, 63]. Moreover, planning is a form of reasoning that may involve multiple aspects, such as temporal and spatial reasoning [277], adding to the significance of integrating LLMs into robotics. Gato [232] 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 toward the realization of AGI. **Robotic Transformer 1 (RT-1)** [23] 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)** [360] further enhances the model's understanding of the world, resulting in more efficient and accurate execution of robotic tasks. By incorporating the CoT 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 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 [202] further extends RT-1 and RT-2 to cross-embodiment settings and shows better transferabilities and zero-shot capabilities. RoboFlamingo [145] leverages pre-trained 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 [315]. 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.

3.5.2 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.

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. [342], they introduce a framework that leverages the capabilities of 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 contrast to the aforementioned studies, Nascimento et al. [195] 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 [231], 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. [29] 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 FL 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 LLMs 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.

4 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 [30, 150, 194]. 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 [334].

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 [24]) and go all the way to 32K (GPT-4 [198]). 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 [9]. This means that the computational resources required to process and train the models become significantly higher as the window size grows. LongNet [61] 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 [37] 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 LMs 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 [162]. 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 combines data from various fields like healthcare and entertainment to enhance understanding and reasoning [4, 66]. Exploring additional data

modalities like video and audio can enrich models' comprehension. Applications extend to EDA [110] and formal methods [301], offering precise, logical reasoning frameworks for system integrity and innovation in software/hardware design.

Efficiency and Cost. Efficiency and cost are critical in foundation model reasoning. Techniques like Model Pruning [263, 294], Compression [359], Quantization [266], Knowledge Distillation [92], and Low-Rank Factorization [107, 233] enhance models' speed and reduce costs, addressing scalability and accessibility challenges.

Human Preference. Mitigating foundation models' risks involves learning from human preference and feedback. Strategies include diverse data collection, continual adaptation, and aligning outputs with real-world evidence. Approaches like Constitutional AI [13] and fine-tuning for diverse approval [14] emphasize integrating human perspectives to build responsible models.

Multilingual Support. Addressing the gap in multilingual reasoning capabilities, Fang et al. [70] use English as a pivot in a commonsense reasoning framework, while Huang et al. [111] introduce cross-lingual thought prompting (XLT). Expanding multilingual support in reasoning models is a key research direction for enhancing global applicability.

Despite the remarkable progress in LLM capabilities, our analysis reveals several critical challenges and limitations that warrant careful consideration. While models like OpenAI o1 and Claude 3.5 Sonnet demonstrate impressive performance across standard benchmarks (Table 1), these results mask underlying limitations in true reasoning capabilities and raise important concerns about their practical deployment. The apparent mastery of benchmarks like MMLU [98] (92.30% by OpenAI o1) and Math [99] (94.80%) may reflect pattern recognition rather than genuine understanding. This is particularly evident in the stark performance drop on FrontierMath [88], where even top models struggle with novel mathematical concepts and complex logical reasoning. Such performance gaps suggest that current evaluation metrics may inadequately capture the depth of understanding required for real-world applications. Beyond technical limitations, several critical risks emerge. The models' increasing capabilities in code generation (>90% on HumanEval [35]) raise security concerns about potential misuse. The high performance of scaled-down models like OpenAI o1-mini (85.20%) and GPT-4T (86.50%), while impressive, may lead to premature deployment of these systems in sensitive domains without fully understanding their limitations. Looking forward, addressing these challenges requires not only technical advances but also more rigorous evaluation frameworks. Future research should focus on developing benchmarks that better assess genuine reasoning capabilities, understanding the bounds of model reliability, and establishing robust safety measures for deployment. The rapid year-over-year improvements, while promising, underscore the need for careful consideration of both the capabilities and limitations of these increasingly powerful systems.

5 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.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. arXiv:2303.08774. Retrieved from https://arxiv.org/abs/2303.08774

[2] Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. 2023. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *EMNLP*. Association for Computational Linguistics, Singapore, 12375–12396.

[3] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as I can, not as I say: Grounding language in robotic affordances. In *CoRL*.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: A visual language model for few-shot learning. In *NeurIPS*. 23716–23736.

[5] Martha W. Alibali, Rebecca Boncoddo, and Autumn B. Hostetter. 2014. Gesture in reasoning: An embodied perspective. In *The Routledge Handbook of Embodied Cognition*. Routledge, 150–159.

[6] Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. GPT4All: Training an Assistant-style Chatbot with Large Scale Data Distillation from GPT-3.5-Turbo. Retrieved from https://github.com/nomic-ai/gpt4all. (2023).

[7] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, et al. 2022. ExT5: Towards extreme multi-task scaling for transfer learning. In *ICLR*.

[8] Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. PROST: Physical reasoning about objects through space and time. In *ACL-IJCNLP 2021*. Association for Computational Linguistics, 4597–4608.

[9] Abi Aryan, Aakash Kumar Nain, Andrew McMahon, Lucas Augusto Meyer, and Harpreet Singh Sahota. 2023. The costly dilemma: Generalization, evaluation and cost-optimal deployment of large language models. arXiv:cs.CL/2308.08061. Retrieved from https://arxiv.org/abs/2308.08061

[10] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. arXiv:2108.07732. Retrieved from https://arxiv.org/abs/2108.07732

[11] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An open language model for mathematics. In *ICLR*.

[12] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*. 12449–12460.

[13] Yuntao Bai et al. 2022. Constitutional AI: Harmlessness from AI feedback. arXiv:cs.CL/2212.08073. Retrieved from https://arxiv.org/abs/2212.08073

[14] Michiel A. Bakker, Martin J. Chadwick, Hannah Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *NeurIPS*.

[15] Bruno Barras, Samuel Boutin, Cristina Cornes, Judicaël Courant, Jean-Christophe Filliatre, Eduardo Gimenez, Hugo Herbelin, Gerard Huet, Cesar Munoz, Chetan Murthy, et al. 1997. *The Coq proof assistant reference manual: Version 6.1*. Ph.D. Dissertation. Inria.

[16] Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. 2021. Physion: Evaluating physical prediction from vision in humans and machines. In *NeurIPS*. 18102–18112.

[17] Stefan Berghofer and Martin Strecker. 2004. Extracting a formally verified, fully executable compiler from a proof assistant. *Electronic Notes in Theoretical Computer Science* 82, 2 (2004), 377–394.

[18] Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *ICLR*.

[19] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science*. Retrieved from https://cdn.openai.com/papers/dall-e-3.pdf (2023).

[20] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* (2023), 1–6.

[21] Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, Yejin Choi . 2020. PIQA: Reasoning about physical commonsense in natural language. In *AAAI*. Vol. 34, 7432–7439.

[22] Rishi Bommasani et al. 2021. On the opportunities and risks of foundation models. arXiv:cs.LG/2108.07258. Retrieved from https://arxiv.org/abs/2108.07258

[23] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. 2023. RT-1: Robotics transformer for real-world control at scale. In *RSS*.

[24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*. 1877–1901.

[25] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:cs.CL/2303.12712. Retrieved from https://arxiv.org/abs/2303.12712

[26] Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. 2019. Monet: Unsupervised scene decomposition and representation. arXiv:1901.11390. Retrieved from https://arxiv.org/abs/1901.11390

[27] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text Pair Dataset. (2022).

[28] Ting-Yun Chang, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tür. 2021. Go beyond plain fine-tuning: Improving pretrained models for social commonsense. In *SLT*. IEEE, 1028–1035.

[29] Chaochao Chen, Xiaohua Feng, Jun Zhou, Jianwei Yin, and Xiaolin Zheng. 2023. Federated large language model: A position paper. arXiv:cs.LG/2307.08925. Retrieved from https://arxiv.org/abs/2307.08925

[30] Canyu Chen and Kai Shu. 2024. Can LLM-generated misinformation be detected?. In *ICLR*.

[31] Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. 2024. AutoAgents: A framework for automatic agent generation. In *IJCAI*.

[32] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *EMNLP*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3313–3323. Retrieved from https://aclanthology.org/2022.emnlp-main.218

[33] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *ACL*.

[34] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*.

[35] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. arXiv:2107.03374. Retrieved from https://arxiv.org/abs/2107.03374

[36] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.

[37] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. arXiv:cs.CL/2306.15595. Retrieved from https://arxiv.org/abs/2306.15595

[38] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research* (2023).

[39] Yixin Chen, Shuai Zhang, Boran Han, and Jiaya Jia. 2023. Lightweight in-context tuning for multimodal unified models. arXiv:cs.CV/2310.05109. Retrieved from https://arxiv.org/abs/2310.05109

[40] Zitian Chen, Mingyu Ding, Yikang Shen, Wei Zhan, Masayoshi Tomizuka, Erik Learned-Miller, and Chuang Gan. 2023. An efficient general-purpose modular vision model via multi-task heterogeneous training. arXiv:2306.17165. Retrieved from https://arxiv.org/abs/2306.17165

[41] Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G. Learned-Miller, and Chuang Gan. 2023. Mod-Squad: Designing mixtures of experts as modular multi-task learners. In *CVPR*. 11828–11837.

[42] Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B. Tenenbaum, and Chuang Gan. 2022. ComPhy: Compositional physical reasoning of objects and events from videos. In *ICLR*.
[43] Zhipeng Chen, Kun Zhou, Beichen Zhang, Zheng Gong, Xin Zhao, and Ji-Rong Wen. 2023. ChatCoT: Tool-augmented chain-of-thought reasoning on chat-based large language models. In *Findings of EMNLP 2023*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, 14777–14790.
[44] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, et al. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. (March 2023). Retrieved from https://lmsys.org/blog/2023-03-30-vicuna/
[45] Ruihang Chu, Yukang Chen, Tao Kong, Lu Qi, and Lei Li. 2021. ICM-3D: Instantiated category modeling for 3D instance segmentation. *IEEE Robotics and Automation Letters* (2021).
[46] Ruihang Chu, Zhengzhe Liu, Xiaoqing Ye, Xiao Tan, Xiaojuan Qi, Chi-Wing Fu, and Jiaya Jia. 2023. Command-driven articulated object understanding and manipulation. In *CVPR*. 8813–8823.
[47] Ruihang Chu, Enze Xie, Shentong Mo, Zhenguo Li, Matthias Nießner, Chi-Wing Fu, and Jiaya Jia. 2023. DiffComplete: Diffusion-based generative 3D shape completion. In *NeurIPS*.
[48] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. 25, 70 (2024), 1–53.
[49] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. arXiv:2110.14168. Retrieved from https://arxiv.org/abs/2110.14168
[50] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. arXiv:2110.14168. Retrieved from https://arxiv.org/abs/2110.14168
[51] Together Computer. 2023. RedPajama: An Open Source Recipe to Reproduce LLaMA Training Dataset. (April 2023). Retrieved from https://github.com/togethercomputer/RedPajama-Data
[52] Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, et al. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-tuned LLM. (2023).
[53] Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *ICLR*.
[54] Andrew Cropper, Sebastijan Dumančić, Richard Evans, and Stephen H. Muggleton. 2022. Inductive logic programming at 30. *Machine Learning* (2022), 1–26.
[55] Kahneman Daniel. 2017. *Thinking, Fast and Slow*.
[56] Leonardo de Moura, Soonho Kong, Jeremy Avigad, Floris Van Doorn, and Jakob von Raumer. 2015. The Lean theorem prover (system description). In *Automated Deduction-CADE-25*. Springer, 378–388.
[57] Luc De Raedt and Kristian Kersting. 2010. Statistical relational learning. *Encyclopedia of Machine Learning* (2010).
[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. IEEE, 248–255.
[59] Karan Desai, Gaurav Kaul, Zubin Trivadi Aysola, and Justin Johnson. 2021. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS*.
[60] David Ding, Felix Hill, Adam Santoro, and Matt Botvinick. 2020. Object-based attention for spatio-temporal reasoning: Outperforming neuro-symbolic models with flexible distributed architectures. arXiv:2012.08508 1. Retrieved from https://arxiv.org/abs/2012.08508
[61] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, and Furu Wei. 2023. LongNet: Scaling transformers to 1,000,000,000 tokens. In *ICLR*.
[62] Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. 2021. Dynamic visual reasoning by learning differentiable physics models from video and language. In *NeurIPS*. 887–899.
[63] Mingyu Ding, Yan Xu, Zhenfang Chen, David Daniel Cox, Ping Luo, Joshua B. Tenenbaum, and Chuang Gan. 2023. Embodied concept learner: Self-supervised learning of concepts and mapping through instruction following. In *CoRL*. PMLR, 1743–1754.
[64] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *EMNLP*.
[65] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. 2023. RAFT: Reward rAnked FineTuning for generative foundation model alignment. *Transactions on Machine Learning Research* (2023).
[66] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. PaLM-E: An embodied multimodal language model. arXiv:2303.03378. Retrieved from https://arxiv.org/abs/2303.03378

[67] Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML*. PMLR, 5547–5569.

[68] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2022. MAGMA – Multimodal augmentation of generative models through adapter-based finetuning. In *EMNLP*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2416–2428. Retrieved from https://aclanthology.org/2022.findings-emnlp.179

[69] Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Rob Carver, Marcin Andrychowicz, Jason Hickey, et al. 2022. Deep learning for twelve hour precipitation forecasts. *Nature Communications* 13, 1 (2022), 1–10.

[70] Yuwei Fang, Shuohang Wang, Yichong Xu, Ruochen Xu, Siqi Sun, Chenguang Zhu, and Michael Zeng. 2022. Leveraging knowledge in multilingual commonsense reasoning. In *ACL*. Association for Computational Linguistics, Dublin, Ireland, 3237–3246. DOI: https://doi.org/10.18653/v1/2022.findings-acl.255

[71] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 1 (2022), 5232–5270.

[72] Juliet Floyd. 2004. Wittgenstein on philosophy of logic and mathematics. *Graduate Faculty Philosophy Journal* 25, 2 (2004), 227–287.

[73] Robert Friedman. 2023. Tokenization in the theory of knowledge. *Encyclopedia* 3, 1 (2023), 380–386.

[74] Daniel Y. Fu, Tri Dao, Khaled Kamal Saab, Armin W. Thomas, Atri Rudra, and Christopher Re. 2023. Hungry hungry hippos: Towards language modeling with state space models. In *ICLR*.

[75] Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *ICML*. PMLR, 10421–10430.

[76] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. In *ICLR*.

[77] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS*.

[78] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. 2023. AssistGPT: A general multi-modal assistant that can plan, execute, inspect, and learn. arXiv:cs.CV/2306.08640. Retrieved from https://arxiv.org/abs/2306.08640

[79] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800GB dataset of diverse text for language modeling. arXiv:2101.00027. Retrieved from https://arxiv.org/abs/2101.00027

[80] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *ICML*. PMLR, 10764–10799.

[81] Thibault Gauthier, Cezary Kaliszyk, Josef Urban, Ramana Kumar, and Michael Norrish. 2021. TacticToe: Learning to prove with tactics. *Journal of Automated Reasoning* 65 (2021), 257–286.

[82] Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023. OpenAGI: When LLM meets domain experts. In *NeurIPS*.

[83] Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2024. Large language models are not strong abstract reasoners. In *IJCAI*.

[84] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A Dialogue Model for Academic Research. Blog post. (April 2023). Retrieved from https://bair.berkeley.edu/blog/2023/04/03/koala/

[85] Malik Ghallab, Dana Nau, and Paolo Traverso. 2004. *Automated Planning: Theory and Practice*. Elsevier.

[86] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *CVPR*. 15180–15190.

[87] Rohit Girdhar and Deva Ramanan. 2020. Cater: A diagnostic dataset for compositional actions and temporal reasoning. In *ICLR*.

[88] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in AI. arXiv:2411.04872. Retrieved from https://arxiv.org/abs/2411.04872

[89] Significant gravitas/auto GPT. 2023. An experimental open-source attempt to make GPT-4 fully autonomous. arXiv:cs.AI/2305.16291. Retrieved from https://arxiv.org/abs/2305.16291

[90] Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *ICLR*.

[91] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. arXiv:cs.CV/2307.12980. Retrieved from https://arxiv.org/abs/2307.12980
[92] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *ICLR*.
[93] Ankit Gupta, Albert Gu, and Jonathan Berant. 2022. Diagonal state spaces are as effective as structured state spaces. In *NeurIPS*. 22982–22994.
[94] Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *CVPR*. 14953–14962.
[95] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *EMNLP*. ACL, Singapore.
[96] Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. 2023. ToolkenGPT: Augmenting frozen language models with massive tools via tool embeddings. In *NeurIPS*.
[97] John Harrison. 2010. Formal methods at Intel–An overview. In *NFM*. Vol. 8, 179–195.
[98] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*.
[99] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS (Round 2)*.
[100] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *NIPS*.
[101] Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *ACL*. 14852–14882.
[102] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3D-LLM: Injecting the 3D world into large language models. In *NeurIPS*.
[103] S. U. Hongjin, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, et al. 2022. Selective annotation makes language models better few-shot learners. In *ICLR*.
[104] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *ACL*. ACL, 14409–14428.
[105] Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. In *ACL*. Association for Computational Linguistics, Toronto, Canada, 8003–8017.
[106] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.
[107] Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. 2022. Language model compression with weighted low-rank factorization. In *ICLR*.
[108] Mengkang Hu, Yao Mu, Xinmiao Chelsey Yu, Mingyu Ding, Shiguang Wu, Wenqi Shao, Qiguang Chen, Bin Wang, Yu Qiao, and Ping Luo. 2024. Tree-planner: Efficient close-loop task planning with large language models. In *ICLR*.
[109] Danqing Huang, Shuming Shi, Chin-Yew Lin, Jian Yin, and Wei-Ying Ma. 2016. How well do computers solve math word problems? Large-scale dataset construction and evaluation. In *ACL*. 887–896.
[110] Guyue Huang, Jingbo Hu, Yifan He, Jialong Liu, Mingyuan Ma, Zhaoyang Shen, Juejian Wu, Yuanfan Xu, Hengrui Zhang, Kai Zhong, et al. 2021. Machine learning for electronic design automation: A survey. *ACM Transactions on Design Automation of Electronic Systems (TODAES)* 26, 5 (2021), 1–46.
[111] Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *EMNLP*.
[112] Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *EMNLP*. ACL, 1051–1068.
[113] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al.. 2023. Language is not all you need: Aligning perception with language models. In *NeurIPS*.
[114] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *ICML*. PMLR, 9118–9147.
[115] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*. 6700–6709.

[116] Tatsuro Inaba, Hirokazu Kiyomaru, Fei Cheng, and Sadao Kurohashi. 2023. MultiTool-CoT: GPT-3 can use multiple external tools with chain of thought prompting. In *ACL*. ACL, 1522–1532.

[117] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. OPT-IML: Scaling language model instruction meta learning through the lens of generalization. arXiv:2212.12017. Retrieved from https://arxiv.org/abs/2212.12017

[118] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation* 3, 1 (1991), 79–87.

[119] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2016. Spatial transformer networks. In *NeurIPS*.

[120] Albert Qiaochu Jiang, Wenda Li, Jesse Michael Han, and Yuhuai Wu. 2021. LISA: Language models of ISAbelle proofs. In *AITP*. 378–392.

[121] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. arXiv:2401.04088. Retrieved from https://arxiv.org/abs/2401.04088

[122] Albert Qiaochu Jiang, Sean Welleck, Jin Peng Zhou, Timothee Lacroix, Jiacheng Liu, Wenda Li, Mateja Jamnik, Guillaume Lample, and Yuhuai Wu. 2023. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In *ICLR*.

[123] Dongwei Jiang, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li. 2020. Speech SIMCLR: Combining contrastive and reconstruction objective for self-supervised speech representation learning. In *INTERSPEECH*. 1544–1548.

[124] Ran Jiao, Zhaowei Wang, Ruihang Chu, Mingjie Dong, Yongfeng Rong, and Wusheng Chou. 2020. An intuitive end-to-end human-UAV interaction system for field exploration. *Frontiers in Neurorobotics* (2020).

[125] Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. BoardgameQA: A dataset for natural language reasoning with contradictory information. In *NeurIPS*.

[126] Wilayat Khan, Muhammad Kamran, Syed Rameez Naqvi, Farrukh Aslam Khan, Ahmed S. Alghamdi, and Eesa Alsolami. 2020. Formal verification of hardware components in critical systems. *Wireless Communications and Mobile Computing* 2020 (2020), 1–15.

[127] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *EMNLP 2020*. ACL, Online, 1896–1907.

[128] Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *EMNLP*.

[129] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*. 22199–22213.

[130] Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. arXiv:cs.CL/2302.02083. Retrieved from https://arxiv.org/abs/2302.02083

[131] Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014. Learning to automatically solve algebra word problems. In *ACL*. ACL, Baltimore, Maryland, 271–281.

[132] Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. 2023. LongForm: Optimizing instruction tuning for long text generation with corpus extraction. arXiv:cs.CL/2304.08460. Retrieved from https://arxiv.org/abs/2304.08460

[133] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. OpenAssistant conversations – democratizing large language model alignment. In *NeurIPS*.

[134] Shibamouli Lahiri. 2014. Complexity of word collocation networks: A preliminary structural analysis. In *ACL*. ACL, Gothenburg, Sweden, 96–105.

[135] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The BigScience ROOTS corpus: A 1.6 TB composite multilingual dataset. In *NeurIPS*. 31809–31826.

[136] Jonathan Laurent and André Platzer. 2022. Learning to find proofs and theorems by learning to refine search strategies: The case of loop invariant synthesis. In *NeurIPS*. 4843–4856.

[137] Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does GPT-3 generate empathetic dialogues? A novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *COLING*. 669–683.

[138] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. GShard: Scaling giant models with conditional computation and automatic sharding. In *ICLR*.

[139] Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *ACL*.
[140] Chunyuan Li. 2023. Large multimodal models: Notes on CVPR 2023 tutorial. arXiv:cs.CV/2306.14895. Retrieved from https://arxiv.org/abs/2306.14895
[141] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *NeurIPS*.
[142] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. PMLR, 19730–19742.
[143] Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2023. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. In *ACL*.
[144] Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2024. Explanations from large language models make small reasoners better. In *AAAI*.
[145] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. 2024. Vision-language foundation models as effective robot imitators. In *ICLR*.
[146] Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. In *ACL*., Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, Toronto, Canada, 4644–4668. DOI: https://doi.org/10.18653/v1/2023.acl-long.256
[147] Xiaonan Li and Xipeng Qiu. 2023. Finding supporting examples for in-context learning. In *EMNLP*.
[148] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2024. Self-alignment with instruction backtranslation. In *ICLR*.
[149] Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In *EMNLP*. 11838–11855.
[150] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *EMNLP*. ACL, 292–305.
[151] Yiming Li, Tao Kong, Ruihang Chu, Yifeng Li, Peng Wang, and Lei Li. 2021. Simultaneous semantic and collision learning for 6-DoF grasp pose estimation. In 2021 IEEE. In *IROS*.
[152] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022. On the advance of making language models better reasoners. arXiv:2206.02336. Retrieved from https://arxiv.org/abs/2206.02336
[153] Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024. LLaMA-VID: An image is worth 2 tokens in large language models. In *ECCV*.
[154] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as policies: Language model programs for embodied control. In *ICRA*. IEEE, 9493–9500.
[155] Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. 2023. SwiftSage: A generative agent with fast and slow thinking for complex interactive tasks. In *NeurIPS*.
[156] Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. 2020. Improving generative imagination in object-centric world models. In *ICML*. JMLR.org, Article 570, 10 pages.
[157] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *ACL*. Regina Barzilay and Min-Yen Kan (Eds.), Association for Computational Linguistics, Vancouver, Canada, 158–167.
[158] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023. DePlot: One-shot visual language reasoning by plot-to-table translation. In *ACL*. Retrieved from https://arxiv.org/abs/2212.10505
[159] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2023. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *ACL*. Retrieved from https://arxiv.org/abs/2212.09662
[160] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.
[161] Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B. Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3?. In *DeeLIO*. 100–114.
[162] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.
[163] Tengxiao Liu, Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2022. RLET: A reinforcement learning based approach for explainable QA with entailment trees. In *EMNLP*. Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7177–7189.

[164] Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023. Plan, verify and switch: Integrated reasoning with diverse X-of-thoughts. In *EMNLP*. ACL, 2807–2822.

[165] Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. In *ACL*. Association for Computational Linguistics, Dublin, Ireland, 2365–2376. DOI: https://doi.org/10.18653/v1/2022.acl-long.168

[166] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692

[167] Jieyi Long. 2023. Large language model guided Tree-of-Thought. arXiv:cs.AI/2305.08291. Retrieved from https://arxiv.org/abs/2305.08291

[168] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *ICML*.

[169] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL*.

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

[171] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. In *NeurIPS*.

[172] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *ICLR*.

[173] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. WizardMath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. arXiv:2308.09583. Retrieved from https://arxiv.org/abs/2308.09583

[174] Man Luo, Shrinidhi Kumbhar, Ming shen, Mihir Parmar, Neeraj Varshney, Pratyay Banerjee, Somak Aditya, and Chitta Baral. 2023. Towards LogiGLUE: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. arXiv:cs.CL/2310.00836. Retrieved from https://arxiv.org/abs/2310.00836

[175] Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y. Zhao. 2023. Dr. ICL: Demonstration-retrieved in-context learning. arXiv:2305.14128. Retrieved from https://arxiv.org/abs/2305.14128

[176] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. In *EMNLP*. Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1384–1403. DOI: https://doi.org/10.18653/v1/2022.emnlp-main.90

[177] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z. Sun, Richard Socher, et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* (2023), 1–8.

[178] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *ACL*. Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.), Association for Computational Linguistics, Toronto, Canada, 1773–1781. DOI: https://doi.org/10.18653/v1/2023.acl-short.151

[179] Matteo Manica, Jannis Born, Joris Cadow, Dimitrios Christofidellis, Ashish Dave, Dean Clarke, Yves Gaetan Nana Teukam, Giorgio Giannone, Samuel C. Hoffman, Matthew Buchan, et al. 2023. Accelerating material design with the generative toolkit for scientific discovery. *npj Computational Materials* 9, 1 (2023), 69.

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

[181] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *ICLR*.

[182] Jiayuan Mao, Xuelin Yang, Xikun Zhang, Noah Goodman, and Jiajun Wu. 2022. CLEVRER-humans: Describing physical and causal events the human way. In *NeurIPS*. 7755–7768.

[183] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. 2024. A language agent for autonomous driving. In *CoLM*.

[184] Norman Megill and David A. Wheeler. 2019. A computer language for mathematical proofs. *arXiv* (2019).

[185] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. 2023. Long range language modeling via gated state spaces. In *ICLR*.

[186] Maciej Mikuła, Szymon Tworkowski, Szymon Antoniak, Bartosz Piotrowski, Albert Q. Jiang, Jin Peng Zhou, Christian Szegedy, Łukasz Kuciński, Piotr Miłoś, and Yuhuai Wu. 2024. Magnushammer: A transformer-based approach to premise selection. In *ICLR*.

[187] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *NAACL-HLT*.
[188] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*.
[189] Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2024. Orca 2: Teaching small language models how to reason. In *ACL*.
[190] Leonardo de Moura and Sebastian Ullrich. 2021. The lean 4 theorem prover and programming language. In *Automated Deduction–CADE 28*. Springer, 625–635.
[191] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. EmbodiedGPT: Vision-language pre-training via embodied chain of thought. In *NeurIPS*.
[192] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *ACL*.
[193] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. arXiv:2306.02707. Retrieved from https://arxiv.org/abs/2306.02707
[194] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *ICLR*.
[195] Nathalia Nascimento, Paulo Alencar, and Donald Cowan. 2023. Self-adaptive large language model (LLM)-based multiagent systems. In *ACSOS-C*. 104–109.
[196] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. 2022. Quality not quantity: On the interaction between dataset design and robustness of clip. In *NeurIPS*. 21455–21469.
[197] Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. 2024. Skeleton-of-thought: Prompting LLMs for efficient parallel generation. In *ICLR*.
[198] OpenAI. 2023. GPT-4 technical report. arXiv:cs.CL/2303.08774. Retrieved from https://arxiv.org/abs/2303.08774
[199] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*.
[200] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*. 27730–27744.
[201] Siqi Ouyang and Lei Li. 2023. AutoPlan: Automatic planning of interactive decision-making tasks with large language models. In *EMNLP*. ACL, 3114–3128.
[202] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. 2024. Open X-embodiment: Robotic learning datasets and RT-X models. In *ICRA*. IEEE, 6892–6903.
[203] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. 2020. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters* 5, 3 (2020), 4867–4873.
[204] Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *EMNLP*. ACL, 3806–3824.
[205] Lawrence C. Paulson. 1994. *Isabelle: A Generic Theorem Prover*. Springer.
[206] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, et al. 2023. RWKV: Reinventing RNNs for the transformer era. In *EMNLP*. ACL, Singapore, 14048–14077.
[207] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. arXiv:2304.03277. Retrieved from https://arxiv.org/abs/2304.03277
[208] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *ICLR*.
[209] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, et al. 2023. DetGPT: Detect what you need via reasoning. In *EMNLP*. ACL, 14172–14189.
[210] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. In *ICML*.
[211] Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. 2023. Formal mathematics statement curriculum learning. In *ICLR*.
[212] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *EMNLP*.

[213] Connor Pryor, Charles Dickens, Eriq Augustine, Alon Albalak, William Wang, and Lise Getoor. 2023. NeuPSL: Neural probabilistic soft logic. In *IJCAI*. 4145–4153.
[214] Júlia Pukancová and Martin Homola. 2020. The AAA ABox abduction solver: System description. *KI-Künstliche Intelligenz* 34, 4 (2020), 517–522.
[215] Cheng Qian, Chi Han, Yi Fung, Yujia Qin, Zhiyuan Liu, and Heng Ji. 2023. CREATOR: Tool creation for disentangling abstract and concrete reasoning of large language models. In *EMNLP*. ACL, 6922–6939.
[216] Shuofei Qiao, Honghao Gui, Chengfei Lv, Qianghuai Jia, Huajun Chen, and Ningyu Zhang. 2024. Making language models better tool learners with execution feedback. In *NAACL*. ACL, 3550–3568.
[217] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *ACL*. ACL, Toronto, Canada, 5368–5393.
[218] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, et al. 2024. Tool learning with foundation models. *ACM Computing Surveys* (Nov. 2024).
[219] Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J. Topol. 2024. LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence* (2024), 1–3.
[220] Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Minqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, et al. 2024. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI* 1, 12 (2024), AIoa2300221.
[221] Jianing Qiu, Wu Yuan, and Kyle Lam. 2024. The application of multimodal large language models in medicine. *The Lancet Regional Health–Western Pacific* 45 (2024).
[222] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.
[223] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *arXiv* (2018).
[224] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.
[225] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. arXiv:2112.11446. Retrieved from https://arxiv.org/abs/2112.11446
[226] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*.
[227] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. (2020).
[228] Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *EMNLP*.
[229] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging language models for commonsense reasoning. In *ACL*. ACL, 4932–4942.
[230] Nazneen Fatema Rajani, Rui Zhang, Yi Chern Tan, Stephan Zheng, Jeremy Weiss, Aadit Vyas, Abhijit Gupta, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. ESPRIT: Explaining solutions to physical reasoning tasks. In *ACL*. Association for Computational Linguistics, Online, 7906–7917. DOI: https://doi.org/10.18653/v1/2020.acl-main.706
[231] IBM. Redbooks. 2004. *Practical Guide to the IBM Autonomic Computing Toolkit*. IBM.
[232] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *Transactions on Machine Learning Research* (2022).
[233] Siyu Ren and Kenny Q. Zhu. 2024. Low-rank prune-and-factorize for language model compression. Torino, Italia, 10822–10832.
[234] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. ImageNet-21K pretraining for the masses. In *NeurIPS*.
[235] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. . 2021. Recipes for building an open-domain chatbot. In ACL, 300–325.
[236] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*. 10684–10695.
[237] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *ACL*. 2655–2671.
[238] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*. 36479–36494.

[239] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al.. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR*.

[240] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *EMNLP-IJCNLP*. ACL, Hong Kong, China, 4463–4473.

[241] Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *ICLR*.

[242] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. arXiv:2211.05100. Retrieved from https://arxiv.org/abs/2211.05100

[243] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *NeurIPS*.

[244] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*. 25278–25294.

[245] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv:2111.02114. Retrieved from https://arxiv.org/abs/2111.02114

[246] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. arXiv:1707.06347. Retrieved from https://arxiv.org/abs/1707.06347

[247] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *EMNLP*. Lluís Màrquez, Chris Callison-Burch, and Jian Su (Eds.), Association for Computational Linguistics, Lisbon, Portugal, 1466–1476. DOI: https://doi.org/10.18653/v1/D15-1171

[248] Dhruv Shah, Alexander T. Toshev, Sergey Levine, and Brian Ichter. 2022. Value function spaces: Skill-centric state abstractions for long-horizon reasoning. In *ICLR*.

[249] Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*.

[250] Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu. 2021. Generate & rank: A multi-task framework for math word problems. In *EMNLP*. ACL, 2269–2279.

[251] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI tasks with ChatGPT and its friends in hugging face. In *NeurIPS*.

[252] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*.

[253] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. Progprompt: Generating situated robot task plans using large language models. In *ICRA*. IEEE, 11523–11530.

[254] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. 2022. Revisiting weakly supervised pre-training of visual perception models. In *CVPR*. 804–814.

[255] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. arXiv:2305.09617. Retrieved from https://arxiv.org/abs/2305.09617

[256] Luca Soldaini and Kyle Lo. 2023. *peS2o (Pretraining Efficiently on S2ORC) Dataset*. Technical Report. Allen Institute for AI. ODC-By. Retrieved from https://github.com/allenai/pes2o

[257] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. 2023. LLM-Planner: Few-shot grounded planning for embodied agents with large language models. In *ICCV*.

[258] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment. In *AAAI*.

[259] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *AAAI*. AAAI Press, 4444–4451.

[260] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR*. 2443–2449.

[261] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*. 843–852.

[262] Jiankai Sun, Yiqi Jiang, Jianing Qiu, Parth Talpur Nobel, Mykel Kochenderfer, and Mac Schwager. 2023. Conformal prediction for uncertainty-aware planning with diffusion dynamics model. In *NeurIPS*.

[263] Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *ICLR*.

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

[265] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *ACL*. ACL, Minneapolis, Minnesota, 4149–4158.

[266] Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang, Qun Liu, Ping Luo, and Ngai Wong. 2022. Compression of generative pre-trained language models via quantization. In *ACL*. ACL, 4821–4836.

[267] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models* 3, 6 (2023), 7.

[268] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. 2023. Gemini: A family of highly capable multimodal models. arXiv:2312.11805. Retrieved from https://arxiv.org/abs/2312.11805

[269] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. 2018. Multinet: Real-time joint semantic reasoning for autonomous driving. In *IV*. IEEE, 1013–1020.

[270] Anthony Tomasic, Oscar J. Romero, John Zimmerman, and Aaron Steinfeld. 2021. Propositional reasoning via neural transformer language models. In *NeSy*.

[271] Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:cs.CL/2307.09288. Retrieved from https://arxiv.org/abs/2307.09288

[272] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al.. 2023. LLaMA: Open and efficient foundation language models. arXiv:cs.CL/2302.13971. Retrieved from https://arxiv.org/abs/2302.13971

[273] Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, Jiatong Shi, et al. 2022. SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities. In *ACL*. 8479–8492.

[274] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *NeurIPS*.

[275] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical AI. *NEJM AI* (2024).

[276] Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Joshua B. Tenenbaum, Daniel L. K. Yamins, Judith E. Fan, and Kevin A. Smith. 2023. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. In *NeurIPS*.

[277] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models-a critical investigation. In *NeurIPS*. 75993–76005.

[278] Lluis Vila. 1994. A survey on temporal reasoning in artificial intelligence. *AI Communications* 7, 1 (1994), 4–28.

[279] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. 2023. Mimicplay: Long-horizon imitation learning by watching human play. In *CoRL*.

[280] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research* (2024).

[281] Haiming Wang, Huajian Xin, Chuanyang Zheng, Zhengying Liu, Qingxing Cao, Yinya Huang, Jing Xiong, Han Shi, Enze Xie, Jian Yin, et al. 2024. LEGO-prover: Neural theorem proving with growing libraries. In *ICLR*.

[282] Haiming Wang, Ye Yuan, Zhengying Liu, Jianhao Shen, Yichun Yin, Jing Xiong, Enze Xie, Han Shi, Yujun Li, Lin Li, et al. 2023. DT-Solver: Automated theorem proving with dynamic-tree sampling guided by proof-level value function. In *ACL*. 12632–12646.

[283] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. 2023. Review of large vision models and visual prompt engineering. arXiv:2307.00855. Retrieved from https://arxiv.org/abs/2307.00855

[284] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *ACL*.

[285] Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *EACL*.

[286] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *ACL*. ACL, 9426–9439.

[287] Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022. ScienceWorld: Is your agent smarter than a 5th grader? In *EMNLP*.

[288] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. SCIBENCH: Evaluating college-level scientific problem-solving abilities of large language models. In *MATH-AI*.

[289] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

[290] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning language model with self generated instructions. In *ACL*.

[291] Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *EMNLP*. ACL, Copenhagen, Denmark, 845–854.

[292] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ tasks. In *EMNLP*.

[293] Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. 2023. NEWTON: Are large language models capable of physical reasoning?. In *EMNLP*.

[294] Ziheng Wang, Jeremy Wohlwend, and Tao Lei. 2020. Structured pruning of large language models. In *EMNLP*. Association for Computational Linguistics.

[295] Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, et al. 2023. Interactive natural language processing. arXiv:2305.13246. Retrieved from https://arxiv.org/abs/2305.13246

[296] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, et al. 2023. De novo design of protein structure and function with RFdiffusion. *Nature* 620, 7976 (2023), 1089–1100.

[297] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *ICLR*.

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

[299] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*. 24824–24837.

[300] Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 Attention (is something you might need too). arXiv:cs.CL/2311.11829. Retrieved from https://arxiv.org/abs/2311.11829

[301] Jim Woodcock, Peter Gorm Larsen, Juan Bicarregui, and John Fitzgerald. 2009. Formal methods: Practice and experience. *ACM Computing Surveys (CSUR)* 41, 4 (2009), 1–36.

[302] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, drawing and editing with visual foundation models. arXiv:cs.CV/2303.04671. Retrieved from https://arxiv.org/abs/2303.04671

[303] Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2024. LaMini-LM: A diverse herd of distilled models from large-scale instructions. In *ACL*.

[304] Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, et al. 2024. A comparative study on reasoning patterns of OpenAI's o1 model. arXiv:2410.13639. Retrieved from https://arxiv.org/abs/2410.13639

[305] Wentao Wu, Aleksei Timofeev, Chen Chen, Bowen Zhang, Kun Duan, Shuangning Liu, Yantao Zheng, Jonathon Shlens, Xianzhi Du, and Yinfei Yang. 2024. MOFI: Learning image representations from noisy entity annotated images. In *ICLR*.

[306] Yuhuai Wu, Albert Qiaochu Jiang, Wenda Li, Markus Norman Rabe, Charles E. Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models. In *NeurIPS*.

[307] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. 2022. Slotformer: Unsupervised visual dynamics simulation with object-centric models. arXiv:2210.05861. Retrieved from https://arxiv.org/abs/2210.05861

[308] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. In *NAACL*. ACL, 1819–1862.

[309] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *ICLR*.

[310] Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. Self-evaluation guided beam search for reasoning. In *NeurIPS*.

[311] Zhipeng Xie and Shichao Sun. 2019. A goal-driven tree-structured neural model for math word problems. In *IJCAI*. 5299–5305.

[312] Jing Xiong, Zixuan Li, Chuanyang Zheng, Zhijiang Guo, Yichun Yin, Enze Xie, Zhicheng YANG, Qingxing Cao, Haiming Wang, Xiongwei Han, et al. 2024. DQ-LoRe: Dual queries with low rank approximation re-ranking for in-context learning. In *ICLR*.

[313] Jing Xiong, Jianhao Shen, Ye Yuan, Haiming Wang, Yichun Yin, Zhengying Liu, Lin Li, Zhijiang Guo, Qingxing Cao, Yinya Huang, et al. 2023. TRIGO: Benchmarking formal mathematical proof reduction for generative language models. In *EMNLP*.

[314] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, 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), 4762–4772.

[315] Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, et al. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation* 2, 4 (2021), 100179.

[316] Fuzhao Xue, Ziji Shi, Futao Wei, Yuxuan Lou, Yong Liu, and Yang You. 2022. Go wider instead of deeper. In *AAAI*. Vol. 36, 8779–8787.

[317] Haoran Yang, Yan Wang, Piji Li, Wei Bi, Wai Lam, and Chen Xu. 2023. Bridging the gap between pre-training and fine-tuning for commonsense generation. 376–383.

[318] Kaiyu Yang, Jia Deng, and Danqi Chen. 2022. Generating natural language proofs with verifier-guided search. In *EMNLP*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 89–105. DOI: https://doi.org/10.18653/v1/2022.emnlp-main.7

[319] Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. LeanDojo: Theorem proving with retrieval-augmented language models. In *NeurIPS*.

[320] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. GPT4Tools: Teaching large language model to use tools via self-instruction. In *NeurIPS*.

[321] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I. Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al.. 2021. SUPERB: Speech processing Universal PERformance Benchmark. In *INTERSPEECH*. 1194–1198.

[322] Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. 2023. Logical reasoning over natural language as knowledge representation: A survey. arXiv:2303.12023. Retrieved from https://arxiv.org/abs/2303.12023

[323] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R. Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*.

[324] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *ICLR*.

[325] Yao Yao, Zuchao Li, and Hai Zhao. 2024. GoT: Effective graph-of-thought reasoning in language models. In *NAACL*. ACL, 2901–2921.

[326] Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *ICML*. JMLR.org, Article 1662, 16 pages.

[327] Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *EMNLP*.

[328] Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023. Complementary explanations for effective in-context learning. In *ACL*. ACL, 4469–4484.

[329] Ziyu Ye, Jiacheng Chen, Jonathan Light, Yifei Wang, Jiankai Sun, Guohao Li, Mac Schwager, Philip Torr, Yuxin Chen, Kaiyu Yang, et al. 2024. Reasoning in reasoning: A hierarchical framework for (better and faster) neural theorem proving. In *MATH-AI*.

[330] Kexin Yi*, Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. CLEVRER: Collision events for video representation and reasoning. In *ICLR*.

[331] Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. In *EMNLP*. ACL, 4031–4047.

[332] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* (2024), nwae403.

[333] Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *EMNLP*.

[334] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know?. In *ACL*. ACL, Toronto, Canada, 8653–8665.

[335] Takuma Yoneda, Jiading Fang, Peng Li, Huanyu Zhang, Tianchong Jiang, Shengjie Lin, Ben Picker, David Yunis, Hongyuan Mei, and Matthew R. Walter. 2024. Statler: State-maintaining language models for embodied reasoning. In *ICRA*.
[336] Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. MetaMath: Bootstrap your own mathematical questions for large language models. In *ICLR*.
[337] Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. PACS: A dataset for physical audiovisual CommonSense reasoning. In *ECCV*.
[338] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: Rank responses to align language models with human feedback. In *NeurIPS*.
[339] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MAmmoTH: Building math generalist models through hybrid instruction tuning. In *ICLR*.
[340] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? arXiv:1905.07830. Retrieved from https://arxiv.org/abs/1905.07830
[341] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al.. 2023. GLM-130B: An open bilingual pre-trained model. In *ICLR*.
[342] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. 2024. Building cooperative embodied agents modularly with large language models. In *ICLR*.
[343] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, et al. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine* (2024), 1–13.
[344] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. arXiv:2205.01068. Retrieved from https://arxiv.org/abs/2205.01068
[345] Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. 2024. Prompt highlighter: Interactive control for multi-modal LLMs. In *CVPR*.
[346] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *ICLR*.
[347] Hongyu Zhao, Kangrui Wang, Mo Yu, and Hongyuan Mei. 2023. Explicit planning helps language models in logical reasoning. In *EMNLP*. ACL, 11155–11173.
[348] James Xu Zhao, Yuxi Xie, Kenji Kawaguchi, Junxian He, and Michael Qizhe Xie. 2023. Automatic model selection with large language models for reasoning. In *EMNLP*.
[349] Xueliang Zhao, Wenda Li, and Lingpeng Kong. 2023. Decomposing the enigma: Subgoal-based demonstration learning for formal theorem proving. arXiv:2305.16366. Retrieved from https://arxiv.org/abs/2305.16366
[350] Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J. Liu. 2022. Calibrating sequence likelihood improves conditional language generation. In *ICLR*.
[351] Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *ACL*. Association for Computational Linguistics, Dublin, Ireland, 6588–6600. DOI: https://doi.org/10.18653/v1/2022.acl-long.454
[352] Zirui Zhao, Wee Sun Lee, and David Hsu. 2023. Large language models as commonsense knowledge for large-scale task planning. In *RSS 2023 LTAMP*.
[353] Chuanyang Zheng, Yihang Gao, Han Shi, Jing Xiong, Jiankai Sun, Jingyao Li, Minbin Huang, Xiaozhe Ren, Michael Ng, Xin Jiang, et al. 2024. DAPE V2: Process attention score as feature map for length extrapolation. arXiv:2410.04798. Retrieved from https://arxiv.org/abs/2410.04798
[354] Chuanyang Zheng, Haiming Wang, Enze Xie, Zhengying Liu, Jiankai Sun, Huajian Xin, Jianhao Shen, Zhenguo Li, and Yu Li. 2024. Lyra: Orchestrating dual correction in automated theorem proving. *Transactions on Machine Learning Research* (2024).
[355] Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2022. miniF2F: A cross-system benchmark for formal Olympiad-level mathematics. In *ICLR*.
[356] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, et al. 2023. Least-to-most prompting enables complex reasoning in large language models. In *ICLR*.
[357] Lipu Zhou, Shuaixiang Dai, and Liwei Chen. 2015. Learn to solve algebra word problems using quadratic programming. In *EMNLP*.
[358] Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, et al. 2023. A foundation model for generalizable disease detection from retinal images. *Nature* (2023), 1–8.

[359] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A survey on model compression for large language models. arXiv:cs.CL/2308.07633. Retrieved from https://arxiv.org/abs/2308.07633

[360] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. 2023. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL* . Vol. 229, PMLR, 2165–2183.

[361] Zhu Ziyu, Ma Xiaojian, Chen Yixin, Deng Zhidong, Huang Siyuan, and Li Qing. 2023. 3D-VisTA: Pre-trained transformer for 3D vision and text alignment. In *ICCV*.

[362] Simiao Zuo, Qingru Zhang, Chen Liang, Pengcheng He, Tuo Zhao, and Weizhu Chen. 2022. MoEBERT: From BERT to mixture-of-experts via importance-guided adaptation. In *NAACL*. Association for Computational Linguistics, 1610–1623.

Received 9 March 2024; revised 4 March 2025; accepted 1 April 2025