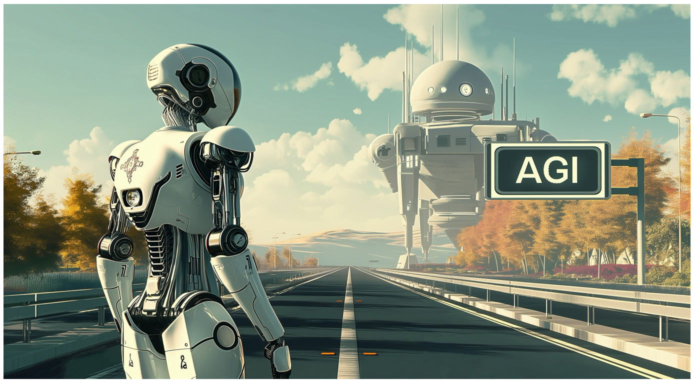
How Far Are We From AGI: Are LLMs All We Need?

Tao Feng[1]*, **Chuanyang Jin**[2*], **Jingyu Liu**[3*], **Kunlun Zhu**[1*]
Haoqin Tu[4], **Zirui Cheng**[1], **Guanyu Lin**[5], **Jiaxuan You**[1†]

{taofeng2, kunlunz2, jiaxuan}@illinois.edu

[1]University of Illinois Urbana-Champaign [2]Johns Hopkins University [3]University of Chicago
[4]University of California, Santa Cruz [5]Carnegie Mellon University

Reviewed on OpenReview: https://openreview.net/forum?id=H2ZKqfNd0U



Abstract

The evolution of artificial intelligence (AI) has profoundly impacted human society, driving significant advancements in multiple sectors. Yet, the escalating demands on AI have highlighted the limitations of AI's current offerings, catalyzing a movement towards Artificial General Intelligence (AGI). AGI, distinguished by its ability to execute diverse real-world tasks with efficiency and effectiveness comparable to human intelligence, reflects a paramount milestone in AI evolution. While existing studies have reviewed specific advancements in AI and proposed potential paths to AGI, such as large language models (LLMs), they fall short of providing a thorough exploration of AGI's definitions, objectives, and developmental trajectories. Unlike previous survey papers, this work goes beyond summarizing LLMs by addressing key questions about our progress toward AGI and outlining the strategies essential for its realization through comprehensive analysis, in-depth discussions, and novel insights. We start by articulating the requisite capability frameworks for AGI, integrating the internal, interface, and system dimensions. As the realization of AGI requires more advanced capabilities and adherence to stringent constraints, we further discuss necessary AGI alignment technologies to harmonize these factors. Notably, we emphasize the importance of approaching AGI responsibly by first defining the key levels of AGI progression, followed by the evaluation framework that situates the status-quo, and finally giving our roadmap of how to reach the pinnacle of AGI. Moreover, to give tangible insights into the ubiquitous impact of the integration of AI, we outline existing challenges and potential pathways toward AGI in multiple domains. In sum, serving as a pioneering exploration into the current state and future trajectory of AGI, this paper aims to foster a collective comprehension and catalyze broader public discussions among researchers and practitioners on AGI. [1]

*Equal contribution. In alphabetical order.

†Corresponding author. All student authors complete this work as interns at UIUC.

[1] Project website: https://github.com/ulab-uiuc/AGI-survey. Unlike traditional publications that remain static, we embrace an innovative approach by treating this paper as a living document. We warmly welcome feedback from the community and plan to update the paper annually. Contributors on the project website will be gratefully acknowledged in future revisions.

Contents

1 Introduction 4

2 AGI Internal: Unveiling the Mind of AGI 6

2.1 AI Perception . . . 6

2.2 AI Reasoning . . . 9

2.3 AI Memory . . . 13

2.4 AI Metacognition . . . 15

3 AGI Interface: Connecting the World with AGI 17

3.1 AI Interfaces to Digital World . . . 17

3.2 AI Interfaces to Physical World . . . 19

3.3 AI Interfaces to Intelligence . . . 21

3.3.1 AI Interface to Other AI agents . . . 21

3.3.2 AI Interfaces to Humans . . . 23

4 AGI Systems: Implementing the Mechanism of AGI 25

4.1 System Challenges . . . 27

4.2 Scalable Model Architectures . . . 27

4.3 Large-scale Training . . . 30

4.4 Inference Techniques . . . 32

4.5 Cost and Efficiency . . . 34

4.6 Computing Platforms . . . 35

4.7 The Future of AGI Systems . . . 36

5 AGI Alignment: Ensuring AGI Meets Various Needs 38

5.1 Expectations of AGI Alignment . . . 38

5.2 Current Alignment Techniques . . . 40

5.3 How to approach AGI Alignments . . . 43

6 AGI Roadmap: Responsibly Approaching AGI 45

6.1 AGI Levels . . . 45

6.2 AGI Evaluation . . . 48

6.2.1 Expectations for AGI Evaluation . . . 49

6.2.2 Current Evaluations and Their Limitations . . . 51

6.3 How to Get to the Next AGI Level . . . 53

6.4 “How Far Are We from AGI” Workshop Discussions . . . 56

6.5 Alternative Perspectives on the AGI Roadmap . . . 60

7 Case Studies: A Bright Future with AGI 63

7.1 AI for Science Discovery and Research 64

7.2 Generative Visual Intelligence 66

7.3 World Models for AGI 68

7.4 Decentralized AI 69

7.5 AI for Coding 70

7.6 Embodied AI: AI for Robotics 72

7.7 Human-AI Collaboration 73

8 Conclusion 74

1 Introduction

> *The path to AGI is not merely a technological journey; it's a philosophical quest to redefine what it means to be intelligent and ethical in a digital age.*
>
> — *Alex Kim, Director of AI Ethics at Future Insight Institute*

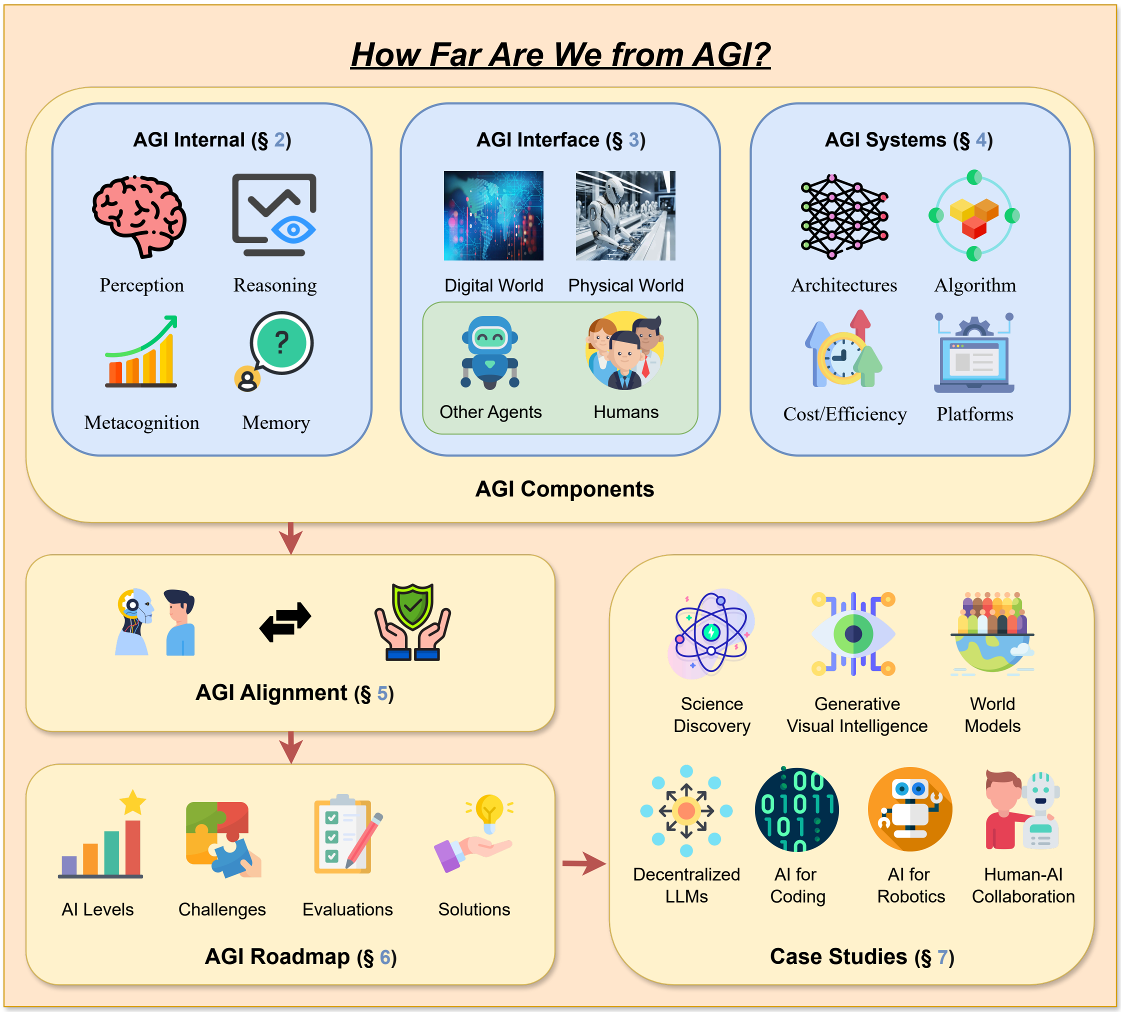


Figure 1: **Overall Structure of This Paper.** This paper starts with discussing core AGI components, including AGI Internal (§ 2), AGI Interface (§ 3), and AGI Systems (§ 4); these discussions help us measure the ability of AGI and estimate how far we are from AGI. As we get closer to AGI, we further expect AGI to meet various constraints, which can be realized by AGI Alignment (§ 5) techniques. We further outline an AGI Roadmap (§ 6) that helps researchers approach AGI responsibly. Finally, some Case Studies (§ 7) are presented to illustrate the current development of early-stage AGI in various fields.

To start approaching the question of how far we are from AGI, it is important to first ground ourselves with the history of artificial intelligence advancement and understand the urge for more advanced systems. Throughout the whole paper, we hope to provide evidences and insights on where we currently stand along the road towards AGI, from the lens of many modern AI systems such as large language models. The goal is to prudently keep questioning ourselves: are LLMs all we need? It is with this enduring curiosity and awareness that we might finally begin to touch the realm of AGI.

Brief History of AI The development of artificial intelligence (AI) has revolutionized human society thanks to their powerful capabilities in many aspects, such as visual perception (Alayrac et al., 2022; Li et al., 2023j), language understanding (Wei et al., 2021; Schick et al., 2023), reasoning optimization (Wei et al., 2022b; Hao et al., 2023; Hu and Shu, 2023), etc. One salient example is the launch of AlphaFold (Jumper et al., 2021) by DeepMind in 2021, which revolutionized the field of protein structure prediction and advanced the frontiers of biological research. Despite the recent advancements, it is worth mentioning that the development of AI is not a smooth journey. Early AI research mainly focused on symbolic research (Stryker, 1959; Turner, 1975) and connectionism (Buckner and Garson, 1997; Medler, 1998), which laid the groundwork for computational approaches to intelligence. From the 1980s to the 1990s, AI faced its winter, and many researchers shifted to practical applications due to high expectations and subsequent disappointments in its development. The rise of machine learning and neural networks (Zadeh, 1996; Kosko, 1992) from the 1990s to the 2010s brought hope to researchers, which led to significant improvements in various applications like natural language processing, computer vision, and analytics. Starting from the 2010s, the advent of deep learning technologies revolutionized AI capabilities, with significant breakthroughs in image (Lu and Weng, 2007; Rawat and Wang, 2017) and speech recognition (Gaikwad et al., 2010; Povey et al., 2011). In recent years, with the emergence of ChatGPT (Wu et al., 2023a; Zhong et al., 2023b), the popularity of large language models (LLMs) has further transformed AI research due to its unified knowledge representation and superior multi-task solving capabilities.

Craving for General-purpose AI Although AI has brought huge improvements to human society, the increasing material and spiritual demands of society have rendered people discontent with the mere convenience provided by AI. Consequently, achieving Artificial General Intelligence (AGI) that enables AI to perform a wider range of tasks more efficiently and effectively has emerged as a pressing concern, which used to describe an AI system that is at least as capable as a human at most tasks (Wang et al., 2018; Voss and Jovanovic, 2023). Therefore, our paper aims to raise attention to the pressing research questions: ***how far are we from AGI***, and moreover, ***how can we responsibly achieve AGI?***

To investigate these questions, existing research mainly falls into three categories: *Definition and Concept, Technical Methods and Applications*, and *Ethical and Social Implications.* (1) *Definition and Concept:* Wang et al. (2018) define the concept of AGI from a point of view of comparison with humans and propose different levels of it. Voss and Jovanovic (2023) provide direction for the path through the AGI by setting the human-like requirements associated with the AGI. (2) *Technical Methods and Applications:* Yan (2022); Wang et al. (2019a) propose that AGI can be achieved by combining logic with deep learning. Das et al. (2023) argues that many risks exist in the development of AGI technology, such as safety and privacy issues. (3) *Ethical and Social Implications:* Rayhan (2023) thinks that humans should consider the ethical implications of creating AGI, which contains impact on human society, privacy, and power dynamics. Bugaj and Goertzel (2007) propose five ethical imperatives and their implications for AGI interactions. These studies have characterized AGI from different aspects. Still, they lack a systematic assessment of the development process of AGI from various aspects and a clear definition of AGI goals, making it difficult to measure the gap between the current AI development and the future of AGI, and moreover, brainstorm possible paths to achieve AGI.

Overall Structure of This Paper Specifically, as is shown in Figure 1, we start with an overview on the major capabilities required for future AGI in terms of its ***internal*** (§ 2) competence, its connection to the external world as an ***interface*** (§ 3), and the underlying infrastructure ***systems*** (§ 4) that support all these functionalities. When it comes to deployment, a more sophisticated ***alignment*** (§ 5) procedure is indispensable to unleash the growing potential of AGI systems under constraints and human expectations. Furthermore, we picture a ***roadmap***(§ 6) where we discuss how to responsibly approach AGI, which outlines the **three levels of AGI** which are Embryonic, Superhuman, and Ultimate AGI that helps locate our current state, associated evaluation framework, as well as our insights to some critical problems that might hinder our progress towards AGI. Finally, we list a couple of ***case studies*** (§ 7) which concretely describe the lineage of AI technology along various domains with cautious limitations and exciting future directions. We hope that this work can lay a common ground and provide a starting point for researchers and practitioners to reflect on the state of AI and brainstorm responsible approaches to achieve AGI.

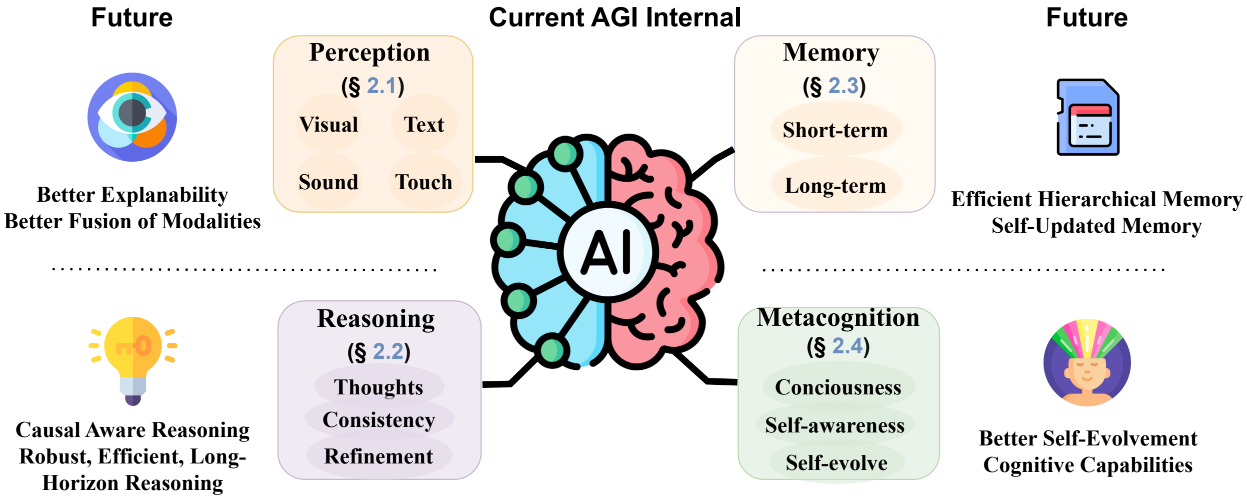


Figure 2: **Current State and Future Expectation of AGI Internal.** We outline four major components for AGI Internal, the mind of AGI: *Perception* (§ 2.1), *Reasoning* (§ 2.2), *Memory* (§ 2.3), and *Metacognition* (§ 2.4), each of which consists of discussions of its current state and future expectation.

2 AGI Internal: Unveiling the Mind of AGI

> *In the end, we are self-perceiving, self-inventing, locked-in mirages that are little miracles of self-reference.*
>
> — *Douglas Hofstadter, I Am a Strange Loop*

The complexity of the human brain, with its specific functional regions dedicated to distinct aspects of cognition and behavior, offers a compelling analogy for the architecture of AGI systems. Similar to the human brain's division into areas for sensory processing, emotion, cognition, and executive functions, the "brain" of an AGI system could also be fundamentally organized into four main components: *perception*, *memory*, *reasoning capabilities*, and *metacognition*. These components mirror the essential aspects of human cognition and play different crucial roles in creating a truly intelligent system. We summarize the overview of this section in Figure 2, which shows the current state and future expectations of AGI internal. Perception (Sec 2.1) refers to the organization and interpretation of sensory information during the interaction between the AGI and its environment (Wang and Hammer, 2018) and is regarded as a fundamental ability in AGI, which includes vision, hearing, touch, smell, etc. The reasoning (Sec 2.2) of AGI is based on the perception of the environment and executes actions to the environment. The interactions between AGI and the environment containing the acquisition of perception and execution of action would be saved as the memories (Sec 2.3) of AGI. The memories will be utilized for the metacognition (Sec 2.4) of AGI.

2.1 AI Perception

> *Humans see what they want to see.*
>
> — *Rick Riordan, The Lightning Thief*

Current State of AI Perception Perception refers to the capability of a system to interpret and make sense of the world around it. This involves the processing and analysis of sensory data to construct a dynamic and contextual understanding of its environment.

Natural language, the primary method of human communication, has evolved from its origins in early human interactions to complex systems like large language models (LLMs). These models have expanded

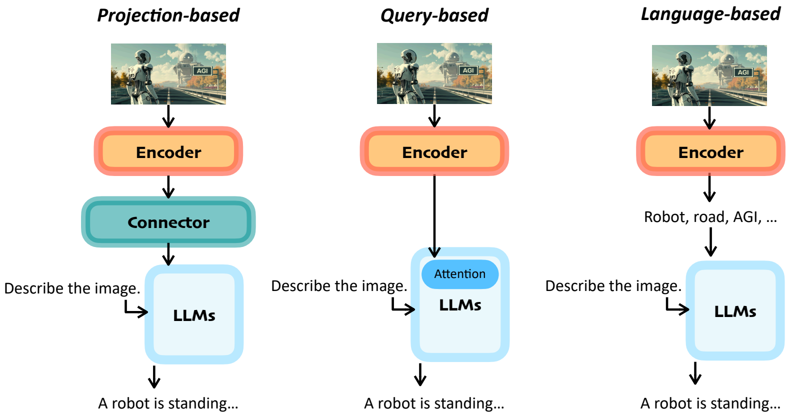


Figure 3: There are three categories for multimodal models with LLM external connections: projection-based, query-based, and language-based.

their capacity to understand and engage in conversations, as well as to perform creative tasks. However, text alone may not fully capture the depth of real-world experiences (Harnad, 1990; Bisk et al., 2020; Tu et al., 2023b), underscoring the importance of multi-modal intelligence that incorporates images, video, and audio for richer human-machine interaction. The transition from traditional LLMs to multi-modal models represents a significant technological leap, facilitating more lifelike interactions across various inputs. This shift, highlighted by recent developments in multi-modal LLMs (OpenAI, 2023b; Team et al., 2023; Li et al., 2023h; Dai et al., 2023; Ye et al., 2023; Zhu et al., 2023a; Liu et al., 2023d; Su et al., 2023; Chen et al., 2023j; Li et al., 2024e; He et al., 2024; Laurençon et al., 2024; Chu et al., 2023), addresses the constraints of language-only comprehension and opens the door to addressing complex challenges that involve multiple forms of data. Integrating various models should adhere to two principles: 1) understanding “how” to incorporate external modal information and ensuring a seamless integration of different modules; 2) determining “what” information to use for preserving the integrity of the original models and enhancing overall capabilities.

The primary objective of utilizing *off-the-shelf* LLMs and multi-modal encoders is to establish a seamless connection between them. This connection can either be external, aligning multi-modal knowledge without altering the existing model structure, or internal, allowing for a more intricate interaction between LLMs and other modal encoders (Yin et al., 2023). These methods often require extensive training, such as creating a learnable interface to link the LLM with non-linguistic modalities, particularly vision. Like LLM pre-training and fine-tuning, Multi-modal LLMs (MLLMs) follow a two-stage training paradigm based on a pre-trained LLM and adapt the process to the multi-modal domain. The first stage, known as the vision-language alignment stage, aims to enable the language model to comprehend visual tokens. The second stage involves multi-modal instruction tuning to align the model with human perceptions. These stages have clear categories based on the combination architectures between the LLM and multi-modal encoders.

- **External Connection of Modalities.** The external approach is based on the idea of bridging the vision branch and LLMs with extra structures and existing models.

  1. *Projection-based*: the modality connector exists outside both the LLMs and multi-modal encoders can be quite straightforward with simple linear projections (Zhu et al., 2023a; Liu et al., 2023d; Su et al., 2023; Chen et al., 2023j; Li et al., 2024e) or incorporating relatively complex selection method (Gao et al., 2023a; Zhang et al., 2023e; Luo et al., 2023; Han et al., 2023b; Fu et al., 2024a). This type of MLLM usually activates the projection layer and/or the LLMs for two stages of alignment training.
  2. *Query-based*: these MLLMs employ a more intricately designed connector but still stand outside of LLMs and multi-modal encoders. This type of model essentially leverages an attention-like interaction between a learnable variable and the vision tokens (Dai et al., 2023; Li et al., 2023h; Ye et al., 2023; He et al., 2024). Since their connectors can learn more complex data patterns than simple projection-based ones, activating the connector alone can also obtain superior multi-modal performance.

3. *Language-based:* language as the interface (Wei et al., 2023; Berrios et al., 2023) is a popular direction for bridging all these *off-the-shelf* models as a holistic and comprehensive one. These methods utilize various pre-built modules for generation and other tasks, with LLMs primarily directing module coordination (Yang et al., 2023d; Li et al., 2023f; Gao et al., 2023a). One main advantage of leveraging tools is that these systems can be more flexible planners Yang et al. (2023d); Wang et al. (2023h); Zhou et al. (2023a); Chen et al. (2023f) for making decisions or artists (Shilong Liu, 2023; Sun et al., 2023b; Huang et al., 2023b; Fu et al., 2023c) for creating versatile multi-media contents with the language as the bridge. One prominent and recent approach is that the GPT-4V model (OpenAI, 2023b) can also generate vivid images by connecting state-of-the-art generators (Rombach et al., 2022). While these approaches offer a wider range of technical solutions for diverse tasks (Wang et al., 2023j;c; Chen et al., 2023e), they generally lack the depth in achieving comparable performance compared with interface-based ones.

- **Internal Connection of Modalities.** Another direction for bridging the multi-modal encoder and the LLM lies in twitching the LLM interblocks.

  1. *Cross-attention-based*: Flamingo (Alayrac et al., 2022) proposed the well-known perceiver with additional cross-attention mechanism inside the attention block of the LLMs. Several variants of Flamingo (Li et al., 2023l; Gong et al., 2023) also use the same or similar framework for tuning the MLLMs.
  2. *Autoregressive:* MLLMs like Fuyu (Bavishi et al., 2023) and its variants (Li et al., 2023m) take vision token as the language token from the pre-training stage and use the same autoregressive training loss to update the whole model parameter.

- **Additional modalities for MLLMs** While earlier models predominantly focused on visual inputs and textual outputs, recent developments have broadened to include diverse modalities in both input and output forms. Regarding inputs, with appropriate modal encoders and training data (Girdhar et al., 2023; Zhu et al., 2023c), LLMs can now comprehend video, audio (Zhang et al., 2023g; Chen et al., 2023d; Lyu et al., 2023; Zhang et al., 2023h), and multiple non-linguistic modalities concurrently (Su et al., 2023; Han et al., 2023b;a), making this approach scalable and accessible. Regarding outputs, recent research has shifted toward creating hybrid content that goes beyond mere text generation. LLMs have evolved from initially retrieving images and generating text (Koh et al., 2023b; Chen et al., 2023j) to producing both visual and textual content. The detailed technical paths of generating images and texts include autoregressive tuning of image-text data with unified representations Sun et al. (2023d); Zheng et al. (2023a); Liu et al. (2024b) and symbolic tuning that transforms text features into image generative models like Stable Diffusion (Koh et al., 2023a; Ge et al., 2023a). Moreover, recent advancements in vision have opened up scalable methods for generating content without text, enhancing the potential for generalizing and scaling vision-only models to generative tasks (Bai et al., 2023; El-Nouby et al., 2024). This opens up the possibility of discovering similar "AGI phenomena" when scaling foundation models in other modalities than language.

AGI-level Perception Current models of perception are still limited by their limited modality and lack of robustness. To address these limitations, we propose several potential future research directions:

- **The diversification of modalities is essential for integrating multiple data types and improving model capabilities.** It is crucial to explore less common modalities, such as graphs, and to integrate multiple modalities, such as images, audio, and video simultaneously (Han et al., 2023b;a). This will require carefully designed modules, high-quality data, and a balanced approach to managing the interplay between different modalities and their relationship with language. For example, while GPT-4V can only handle language and visual information, the recent Gemini (Team et al., 2023) model expands its capacity to a wider range of audio and video. Potential methods for incorporating other modal perceptions: a unified modal representation tool like ImageBind (Han et al., 2023b), LangaugeBind (Zhu et al., 2023c) could bridge the modal gap and lessen the burden of learning from other modalities. Existing models that incorporate these tools have shown promising results in efficiency and task performance (Su et al., 2023; Han et al., 2023b).

- **Encouraging multi-modal systems to be more robust and reliable.** As more comprehensive benchmarks covering not only general situations but also challenging inputs like math problems (Yue et al., 2023), counterfactual instructions (Zhang et al., 2023m), and attack strings (Qi et al., 2023; Zhao et al., 2023b) emerge, it becomes evident that multi-modal systems, especially the smaller ones, generally fall short in performance when facing adversarial examples and are heavily language-biased (Tu et al., 2023a; Cui et al., 2023), lacking the reasoning ability under out-of-distribution situations like multi-panel images (Fan et al., 2024), sketches (Tu et al., 2023a), long sequence images (Wang et al., 2024b). These observations pose potential risks in real-world applications. To address these challenges and build more robust multi-modal AGI models, several strategies in terms of the employed learning data are considered. Future research could benefit from incorporating adversarial examples into training (Liu et al., 2023d) or involving increased diversity of training data instruction formats (Dai et al., 2023).

- **Explainable multi-modal models point out the direction for future improvement.** Unlike traditional models, multi-modal models involve complex interactions between different modalities, making it essential to unravel their inner workings to understand and create stronger multi-modal ones. To address this, research efforts have focused on providing explanations during training or generation, offering insights into model performance and reasoning. Methods like probing model performance with diverse training data have been explored (Liu et al., 2023d; Zhao et al., 2023c; Tu et al., 2023c). Additionally, the Gemini (Team et al., 2023) team enhances user trust and understanding of the AI's reasoning process by providing explanations of the generation (Team et al., 2023). Another aspect of improving multi-modal models is increasing transparency. This involves identifying the specific model components or configurations that contribute to the system's abilities (*e.g.*, vision encoder, connector, or training paradigms) (Wang et al., 2023d; He et al., 2024). Studies have also specifically investigated the impact of different modality processors on the overall model performance (Lin et al., 2023a; Tong et al., 2024b). As multi-modal models advance, future research must prioritize explainability and transparency. This will enable us to take the full potential of these powerful AI systems while ensuring their responsible and ethical use. For example, future research avenues could explore strict controlled experiments for training AI models to decompose each part (Tong et al., 2024a) or probing model components to find the most effective module (Zhao et al., 2023c).

2.2 AI Reasoning

> *All our knowledge begins with the senses, proceeds then to the understanding, and ends with reason. There is nothing higher than reason.*
>
> — *Immanuel Kant, Critique of Pure Reason*

Reasoning is the cognitive process of drawing conclusions or making decisions based on available information, logic, and prior knowledge. It involves evaluating evidence, identifying relationships, and applying rules or principles to solve problems (Fagin et al., 2004; Huang and Chang, 2022). AI reasoning refers to the ability of AI systems to simulate this process, enabling machines to understand situations, infer conclusions, and make decisions in a way that mimics human reasoning.

Current State of AI Reasoning Substantial research indicates that reasoning capabilities have emerged in large machine-learning models. Large Language Models (LLMs), including GPT-3 (Brown et al., 2020), LLaMA 2 (Touvron et al., 2023), and PALM 2 (Anil et al., 2023), have unlocked flexible zero-shot and few-shot reasoning capabilities across various NLP tasks (Kojima et al., 2022). Large Visual Language Models (LVLMs) such as GPT-4 with vision (OpenAI, 2023a) and Gemini (Team et al., 2023), have advanced this progress by effectively integrating vision and language reasoning.

Numerous strategies have been developed to elicit effective and efficient reasoning without updating the model. These methods have substantially improved model performance across a wide range of tasks, including arithmetic, commonsense, symbolic reasoning, and challenges in both simulated and real-world settings.

- **Navigating through thoughts.** Chain of Thought (CoT) (Wei et al., 2022b; Kojima et al., 2022) generates a sequence of intermediate reasoning steps, known as “thoughts,” to enable models to decompose multi-step problems and allocate additional computation to more complex tasks. This offers an interpretable insight into the model’s reasoning process, helping to comprehend how an answer is derived and identify where errors in reasoning might occur. Tree of Thoughts (ToT) (Yao et al., 2023) employs tree-based search algorithms to navigate through “thoughts” for deliberate problem-solving. This allows LLMs to explore multiple reasoning paths and perform deliberate decision-making, including looking ahead or backtracking when necessary. Graph of Thoughts (GoT) (Besta et al., 2023) organizes information into a graph structure, where “thoughts” are vertices, and edges correspond to dependencies between these vertices. This graph-based organization facilitates more intricate integration and manipulation of thoughts, allowing for the creation of more sophisticated reasoning pathways and the incorporation of feedback mechanisms. Program of Thoughts (PoT) (Chen et al., 2022) leverages language models to express the reasoning process as a program, delegating the computation to an external computer that executes the generated programs to obtain the answer. This separation of computation from reasoning improves performance on problems that demand highly symbolic reasoning skills.

- **Self-consistent reasoning.** Self-consistency (Wang et al., 2022a) samples a diverse set of reasoning paths and selects the most consistent answers. This method overcomes the constraints of greedy decoding by balancing open-ended and optimal text generation, utilizing the diversity of reasoning paths to achieve more reliable outcomes. Maieutic Prompting (Jung et al., 2022) induces a tree of explanations abductively and recursively, then frames the inference as a satisfiability problem over these explanations and their logical relations. Progressive-Hint Prompting (Zheng et al., 2023b) employs previously generated answers as hints to progressively guide toward the correct answers, enforcing a level of self-consistency with earlier responses.

- **Additional prompting strategies for enhanced reasoning.** Many other prompting methods have been developed to improve the reasoning abilities of LLMs. Complexity-Based Prompting (Fu et al., 2022) creates rationales with more reasoning steps with an example selection scheme. Auto-CoT (Zhang et al., 2022b) samples questions with diversity and automatically generates reasoning chains to construct demonstrations. Least-to-Most Prompting (Zhou et al., 2022b) breaks down a complex problem into a series of simpler subproblems and then solves them in sequence. Decomposed Prompting (Khot et al., 2022) decomposes tasks into simpler sub-tasks and dynamically delegates them to sub-task-specific models. ToolLLM (Qin et al., 2023b) and ToRA (Gou et al., 2023b) integrate natural language reasoning with the use of external tools, significantly enhancing their ability to perform complex reasoning. Collaboration mechanisms between multiple agents, such as debate (Du et al., 2023), reflection (Zhang et al., 2023l), voting (Li et al., 2024g), or role-playing as different characters (Qian et al., 2023a; Zhou et al., 2023d), can further enhance their reasoning performance.

- **Dynamic reasoning and planning.** ReAct (Yao et al., 2022) prompts LLMs to generate reasoning traces and action plans in an interleaved manner. This synergy between reasoning and action enables dynamic reasoning, creating, maintaining, and adjusting action plans while interacting with external environments like Wikipedia. This interaction allows the integration of additional information into the reasoning process and addresses issues like hallucination and error propagation common in chain-of-thought reasoning. “Describe, Explain, Plan and Select” (DEPS) (Wang et al., 2023b) enhances plan reliability by incorporating a dynamic feedback loop that includes description, explanation, and plan adjustment stages, significantly improving error correction and planning efficiency. Inner Monologue (Huang et al., 2022b) underscores the utility of feedback-informed planning, demonstrating improved task completion and adaptability in diverse environments by dynamically incorporating feedback to refine and adjust plans in real-time. ProgPrompt (Singh et al., 2023) enables the generation of executable task plans that are both contextually relevant and adaptable to the robot’s capabilities and the environment’s state by structuring prompts as programmatic instructions and incorporating environment state feedback through assert statements. LLM+P (Liu et al., 2023b) combines the natural language processing strengths of LLMs with the precise problem-solving skills of classical planners, providing optimal solutions for planning problems that involve language description. Thought Rollback (Chen and Li, 2024) introduces a rollback mechanism that allows LLMs to revise prior steps based on error analysis, fostering adaptive reasoning by dynamically adjusting thought structures to improve problem-solving accuracy. Parameter-efficient finetuning methods (PEFTs) (Xu et al., 2023c) adapt large language models by updating only a small subset of their parameters, reducing memory usage and computational costs while achieving performance comparable to full-model finetuning, whereas ReFT methods (Wu et al., 2024a) operate on frozen model representations rather than weights, learning task-specific interventions that efficiently steer model behavior during inference.

- **Reflection and refinement.** Self-refine (Madaan et al., 2024) employs iterative generation, self-generated feedback, and refinement, enabling large language models to adjust based on feedback after each generative cycle. Reflexion (Shinn et al., 2023) expands on the ReAct framework by integrating an evaluator to assess action trajectories and utilizing an LLM to generate verbal self-reflections to provide feedback for future trials. CRITIC (Gou et al., 2023a) utilizes external tools, such as knowledge bases and search engines, to validate the actions produced by LLMs, then employs external knowledge for self-correction to minimize factual errors.

- **Integrating language models, world models, and agent models.** While language models fall short of consistent reasoning and planning in various scenarios, world and agent models can provide essential elements of human-like deliberative reasoning, including beliefs, goals, anticipation of consequences, and strategic planning. The LAW framework (Hu and Shu, 2023) suggests reasoning with world and agent models, with language models serving as the backend for implementing the system or its components. This framework combines three models in a cognitively grounded way, fostering more robust and versatile reasoning capabilities. Within this framework, Reasoning via Planning (RAP) (Hao et al., 2023) prompts an LLM to function as an agent model, guided by the same LLM acting as the world model, which predicts the next state of the reasoning after applying an action to the current state. BIP-ALM (Jin et al., 2024) and LIMP (Shi et al., 2024b) use language models as the planner in agent models, leading to an improved Theory of Mind capacity compared to using language models to infer other agents' mental states directly. Recent studies have explored the potential for using language models to generate goals (Xie et al., 2023) or rewards (Yu et al., 2023; Kwon et al., 2023b; Ma et al., 2023a) in agent models to guide planning. Integrating these approaches, neural-symbolic methods can bridge the gap between the abstract reasoning facilitated by LLMs and the structured decision-making processes inherent in world and agent models. Logic-LM (Pan et al., 2023) apply symbolic execution on logical reasoning. Symbol-LLM (Xu et al., 2023b) unifies neural-symbolic applications under a Symbol+Delegation setting.

- **Reasoning and planning of embodied agents.** Several studies have proposed methods for reasoning procedures in embodied agents, enhancing their ability to execute tasks and interact with their environment and other agents in a more sophisticated manner. Voyager (Wang et al., 2023h) is an embodied agent in the Minecraft game that uses iterative prompting for dynamic reasoning and skill acquisition. It begins with an automatic curriculum that suggests tasks based on the agent's capabilities and the world state. Then Voyager creates code for these tasks and enters a cycle of execution, feedback assessment, and code refinement. This loop of reasoning, supported by a self-verification module, guarantees task completion and continuous learning. Generative Agents (Park et al., 2023) are language agents grounded in a sandbox game that affords interaction with the environment and other agents. Their memory stream records experiences in natural language, enabling moment-to-moment behavior informed by the relevance, recency, and importance of memory objects. Through reflection, the agents synthesize these memories into higher-level inferences, leading to the creation of coherent plans. While executing these plans, agents continuously reason over recent observations to maintain or adjust the plan.

AGI-level Reasoning While current systems exhibit impressive reasoning skills across various tasks, they also have several substantial flaws and challenges.

- **Foundation models need to learn causation for better understanding and generalization.** The foundation models rely heavily on patterns identified in their training data, which do not always capture the depth and breadth of human knowledge and experiences. Furthermore, these models often operate based on patterns extracted from data without truly comprehending the underlying causal relationships. Zečević et al. (2023) describe how LLMs might superficially replicate causal relationships but lack the underlying causal mechanisms, leading them to be more like "causal parrots" rather than genuinely causal models. Jin et al. (2023a) presents a challenging dataset for causal reasoning and suggests that LLMs may still be far from reasoning reliably about causality. Future advancements in AGI should focus on learning causation over correlation, thereby achieving better generalization and deeper understanding.

- **AGI must address the challenge of complex and long-context reasoning.** Current models face significant difficulties with complex, multi-step reasoning tasks. As discussed earlier, many strategies have been developed to mitigate this issue. However, these strategies often require explicit guidance or careful framing of the problem, which may become unnecessary in the future. Even with these approaches, it remains a challenge for models to process information across long contexts and maintain coherent and logical reasoning throughout the reasoning tasks (Srivastava et al., 2022).

- **AGI should tackle challenges in hallucination, uncertainty assessment, and ambiguity handling, improving performance and safety.** Current models are susceptible to hallucination, where they generate content that is either nonsensical or unfaithful to the provided source content (Ji et al., 2023a; Li et al., 2023d). This tendency hampers performance and raises significant safety concerns in real-world applications. Moreover, these models often struggle to accurately assess their uncertainty and effectively communicate it in their outputs, which can lead to results that are potentially misleading (Zhou et al., 2023b). Additionally, they struggle to handle ambiguity, an issue that can complicate their usability in complex scenarios (Liu et al., 2023h).

- **AGI should get better at social reasoning to enhance interactions with humans and other agents.** Current AI models lack a robust Theory of Mind, the ability to understand the mental states of others (Sap et al., 2022; Ullman, 2023; Jin et al., 2024; Shi et al., 2024b). Improving this capability is essential for AGI systems to safely and effectively interact with humans and other agents in an open-ended manner. Understanding social cues and norms is central to this development, as it allows AGIs to interpret and respond to implicit communication and behavioral expectations in varied contexts (Puig et al., 2020; Zhang et al., 2023b). Advancements in social reasoning in AGI systems could lead to more empathetic and context-aware technologies, ensuring that these systems engage harmoniously and meaningfully in human societies.

- **AGI should solve the challenge in explainability and transparency, thereby enhancing their reliability in decision-making.** Currently, most AI systems lack these qualities, making it difficult to understand how they arrive at specific conclusions or answers (Chen et al., 2023k). Techniques that aim to elicit reasoning in natural language do not consistently align with the actual reasoning processes used by the models, and the explanations generated can be systematically misleading (Bowman, 2023). This limitation hinders their reasoning abilities and creates significant challenges when decision-making requires justification or auditing, particularly in fields like healthcare or law. Recent work on sparse autoencoders, specifically using them to address polysemanticity in neural networks, has shown promise in enhancing the explainability of AI models (Cunningham et al., 2023; Gao et al., 2024; Templeton, 2024). These autoencoders help in disentangling and isolating meaningful features within neural networks, leading to more interpretable and mono semantic features that are easier to understand. Incorporating similar techniques to interpret neural networks, and developing more techniques to explain different decision-making process of AGI systems can facilitate more trustworthy and accountable AI applications.

- **Future AGI systems aim for dynamic reasoning across domains, ethical and efficient planning, and human-like intelligence at unparalleled scales and speeds.** We are still far from achieving AGI-level capability that allows reasoning and planning across varied domains without retraining or human oversight (Saparov et al., 2024). The journey involves enhancing AI systems to transfer knowledge and skills across vastly different areas, enabling them to address unforeseen situations efficiently. A key development focus is creating algorithms that can plan at various levels of abstraction, from broad strategic goals to the specifics of detailed actions. Additionally, AI systems urgently need to become better at managing resources—such as time, energy, and costs—more efficiently during the planning phases. Equally critical is ensuring that these planning processes adhere to ethical standards and safety regulations, especially in sensitive sectors, to avoid misuse or unintended outcomes. Advancements in these areas will collectively move us closer to realizing AGI with robust, versatile planning capabilities.

- **More advanced reasoning abilities are required to solve complex real-world tasks.** In the future, enhancements in prompting techniques or task-framing methods promise to significantly boost the reasoning capabilities of foundational models. On the other hand, future advancements might eliminate the need for complex prompting to aid in reasoning, with these aids being "implicit". A future AGI system could potentially emulate any grounding, learning, and decision-making by listing all the possible actions and simulating and evaluating each one before executing its actual decision-making process. Even more audaciously, it may simulate these implicitly in neurons without any intermediate reasoning in context. For such a system to simulate this flawlessly, it would require an exceptionally realistic world model.
- **Future AGI systems will be able to understand context, infer causality, and apply advanced logical planning dynamically across diverse domains.** By synthesizing vast amounts of information and applying deliberate planning, they can generate innovative solutions to formulating creative hypotheses, making sophisticated moral judgments, predicting the outcomes of novel scenarios, and continuously learning and refining their understanding of the world. Essentially, these future AGI systems would not only excel in processing and generating information but will also be capable of understanding and interacting with the world in a manner deeply analogous to human intelligence, yet at a scale and speed that greatly surpasses human capabilities.

2.3 AI Memory

> *Remembrance of things past is not necessarily the remembrance of things as they were.*
>
> *— Marcel Proust, In Search of Lost Time*

Language and vision models, by their nature, are stateless; they do not maintain information between interactions. However, advanced agents differ in that they can manage internal or external memory, enabling them to engage in complex, multi-step interactions (Sumers et al., 2023; Zhang et al., 2024a). This memory stores intermediate information, domain-specific or broad knowledge, and sequences of the agents' previous observations, thoughts, and actions, among others. It assists agents in utilizing previous knowledge or experiences for reasoning, planning, and self-improvement.

Current State of AI Memory We examine the current state of AI memory, focusing on three key aspects: memory management, which determines what and when to store; memory representation, which defines how information is structured; and memory utilization, which addresses how to apply and use the memory efficiently and effectively.

- **Memory management.** Memory is categorized by duration into short-term and long-term memory.
  1. *Short-term memory*: Short-term memory plays a crucial role in maintaining information needed for current decision-making processes. A notable example is in-context prompting, which uses the foundation models' own context as a form of short-term memory. This approach can provide additional information or examples (Wang et al., 2020), or can be used to generate intermediate reasoning (Nye et al., 2021; Wei et al., 2022b). More broadly, short-term memory encompasses all immediate data essential for decision-making. This includes: (1) real-time data collected or processed by perception modules; (2) immediate outputs from reasoning, planning, and self-evolution modules; and (3) information actively retrieved from long-term memory. These elements collectively are synthesized to guide and inform subsequent actions.
  2. *Long-term memory*: Long-term memory can be broadly classified into two main types: experiences and knowledge. Experiences encompass a range of elements such as past observations, thoughts, actions, and more. This rich collection of experiences serves a critical function in decision-making processes. By retrieving relevant experiences, agents can gain additional information necessary for reasoned judgment, understand feedback from past actions, and achieve a level of generalization in their understanding and reasoning. For example, Reflexion (Shinn et al., 2023) reflects on task feedback signals and maintains them as textual summaries. These summaries are directly incorporated into the context of subsequent episodes, aiding in performance enhancement. Generative agents (Park et al., 2023) document their

experiences in natural language and retrieve memories using a mix of relevance (embedding-based), recency (rule-based), and importance (reasoning-based) criteria.

Knowledge represents an agent's understanding of the world and itself, which enhances its reasoning and decision-making capabilities. Knowledge can originate from two sources. First, AI agents can collect and assimilate knowledge from experiences, integrating new information or skills into their existing knowledge. For example, Voyager (Wang et al., 2023h) maintains a continuously expanding skill library of executable codes for preliminary actions to accomplish tasks. Second, AI agents or models can utilize external knowledge bases. For instance, ReAct (Yao et al., 2022) employs Wikipedia APIs to acquire external knowledge when agents lack information during their activities. ChatGPT Browse with Bing enables ChatGPT to access internet knowledge for answering questions, significantly enhancing its ability to provide accurate responses (OpenAI, 2023a). Retrieval-augmented methods (Lewis et al., 2020; Guu et al., 2020; Shuster et al., 2021; Borgeaud et al., 2022) leverage a knowledge base of unstructured text. The "reading to learn" methods (Branavan et al., 2012; Hanjie et al., 2021) utilize domain knowledge from text manuals to influence the policies in reinforcement learning.

- **Memory representation.** Regarding representation, memory is divided into textual memory and parametric memory. Textual memory is the prevalent method for representing memory content today. It can include both unstructured formats like raw natural language and structured forms such as tuples, databases, and more. Alternatively, memory can be represented in a parametric form. Techniques like supervised fine-tuning (Hu et al., 2021), knowledge editing (De Cao et al., 2021; Mitchell et al., 2021) and model merging (Du et al., 2024; Yang et al., 2024; Lu et al., 2024; Goddard et al., 2024) can integrate domain-specific knowledge into model parameters. For textual memory, each inference involves incorporating memory into the context prompt, leading to higher costs and extended processing times during the reading and inference processes. Conversely, parametric memory often incurs greater costs during the writing phase, as fine-tuning models is more challenging than simple text storage. Regarding interpretability, textual memory is generally more transparent than parametric memory, as the natural language provides the most direct means for human understanding (Zhang et al., 2024a).

- **Memory utilization.** There are two common technologies to utilize memories: memory retrieval and long-context LLMs.

  1. *Memory retrieval:* Memory retrieval involves reading information from long-term memory to short-term memory for immediate use. This can be accomplished through rule-based retrieval or retrieval-augmented methods. Rule-based retrieval can search memory using keywords, timesteps, or specific patterns. In retrieval-augmented approaches, the Dense Passage Retriever (DPR) (Karpukhin et al., 2020) creates dense representations of documents and retrieves the most relevant documents based on their prior probability using Maximum Inner Product Search (MIPS). The Retrieval-Augmented Language Model pre-training (REALM) (Guu et al., 2020) integrates unsupervised pre-training of a knowledge retriever with masked language modeling, enabling direct retrieval of documents to supplement language predictions. Retrieval-Augmented Generation (RAG) models (Lewis et al., 2020; Shuster et al., 2021; Borgeaud et al., 2022) employ a non-parametric memory, such as a dense vector index of Wikipedia, accessed via a pre-trained neural retriever (e.g., DPR). These documents are processed by a seq2seq model, which conditions its output generation on both the input and the retrieved documents. Both the retriever and seq2seq modules, initialized from pre-trained models, are jointly fine-tuned, allowing both retrieval and generation to adapt to downstream tasks.
  2. *Long-context LLMs:* The expansion of the context window in long-context LLMs opens up new avenues for models to access their long-term memory. Works like Ring Attention (Liu et al., 2023j) and LongRoPE (Ding et al., 2024b) greatly reduce the time and cost of long context inference by improving the operation mechanism and storage method of attention. More powerful GPUs with enhanced memory capabilities and further breakthroughs in memory-efficient attention mechanisms (Dao et al., 2022; Tay et al., 2022), allowing the context window for pre-trained LLMs to increase from 1024 tokens in GPT-2 (Radford et al., 2019), to 8192 in GPT-4 (Achiam et al., 2023), and now exceeding 16K tokens. With these expanded context windows, AI systems can more effectively store and recall knowledge and experiences within their context, enabling faster and more comprehensive context-based reasoning.

AGI-level Memory Achieving AGI-level memory requires advanced management of vast, dynamically organized information, improved utilization of memory for reasoning and planning, and the ability to autonomously update and enrich the memory base. It involves human-like deliberate use of memory, yet surpasses human capacities, allowing for more comprehensive and intricate recall.

- **Future AGI will efficiently manage diverse and hierarchical memories, ensuring privacy, collaboration, and scalability.** Current AI agents face challenges in building hierarchical memory and seamlessly incorporating information across various formats. Future AGI systems are anticipated to excel in handling diverse forms of memory, such as embeddings, videos, documents, and databases, both efficiently and effectively. They will also need to address different levels of memory permissions: local memory is essential for preserving privacy, while shared memory, in centralized or decentralized structures as required, is necessary for collaborative efforts and distributed processes. The architectures employed for memory management are expected to be highly organized and scalable. These systems will likely feature advanced algorithms for categorizing and indexing information, allowing agents to efficiently retrieve and record a wide spectrum of experiences and knowledge. Additionally, they may dynamically update and reorganize their memory structures, ensuring optimal storage and retrieval of information.
- **Future AGI will enhance memory utilization through the integration of retrieval and advanced reasoning, enabling more human-like intelligence and adaptability.** Beyond simple memory retrieval, future AGI systems could refine memory utilization by intricately combining retrieval processes with advanced reasoning that strategically synthesizes and applies information in context-appropriate ways. Beyond fixed implementations, the retrieval procedures could be learned or updated to adapt to changing circumstances. The ability to access and apply relevant information from their memory in real-time would be a significant step towards more human-like intelligence, enabling these systems to respond to new situations with a high level of understanding and adaptability.
- **Future AGI will autonomously update their knowledge, enabling continuous learning and adaptation while ensuring safety.** Unlike existing retrieval-augmented models that primarily rely on pre-existing, human-generated content, future AGI systems could autonomously generate, evaluate, and incorporate new content into their memory banks. These updates should encompass knowledge essential for performance enhancement and experiences the systems can draw upon. This concept is closely linked to self-evolve, which we will discuss later. It would allow AGIs to learn from their own experiences and insights, continually enriching and updating their knowledge base. In a constantly changing world, this capability will also enable the systems to adapt quickly given new information and unlearn outdated knowledge. A crucial aspect is to guarantee the safety of the memory updates, ensuring that no harmful information is written that could lead to contamination. Designing safety constraints for autonomous AGIs involves creating robust validation protocols that assess the truthfulness, relevance, and impact of new information before integration. We can implement expert systems to periodically review updates, use anomaly detection to flag outliers and potentially harmful data, and employ additional methods to enforce these safety constraints.

2.4 AI Metacognition

> *I am no bird, and no net ensnares me: I am a free human being with an independent will.*
>
> *— Charlotte Brontë, Jane Eyre*

Metacognition (Choudrie and Selamat, 2006) of humans involves key cognitive and emotional skills such as understanding complex situations, self-awareness, and motivation to innovate. These abilities help share implicit knowledge and drive personal growth.

The development of AGIs with such advanced metacognition provokes a fundamental inquiry: are we, in our pursuit of artificial intelligence, on the verge of creating a new form of life? The implications are far-reaching, as introducing entities with self-awareness and autonomous decision-making capabilities could redefine the boundaries of life and intelligence. This tantalizing horizon calls for meticulous ethical consideration and regulatory scrutiny to ensure that the evolution of AGI contributes positively to human society and does not inadvertently engender a paradigm shift with unforeseen consequences.

Current State of AI Metacognition The discourse on metacognition extends into the realm of AGI, where such capabilities are deemed equally critical. For AGI systems, metacognition, such as self-awareness (Chella et al., 2020; Subagdja et al., 2021), consciousness (Dehaene et al., 2021), the capacity for self-evolution (Floreano et al., 2004; Tao et al., 2024), and theory of mind (Cuzzolin et al., 2020), are posited to be foundational for bridging the gap towards achieving AGI. These internal abilities can enable AI systems to autonomously learn, efficiently complete tasks, and align more closely with human intentions.

- **Self-Awareness in AGI.** Developing self-awareness in AI, particularly within the realm of robotics (Scassellati, 2002), hinges on intricate concepts such as self-reflection (Shinn et al., 2024), meta-cognition (Langdon et al., 2022), and self-distancing (Kross and Ayduk, 2017). These concepts are integral to constructing social robots equipped with cognitive architectures that support self-description, the utilization of personal pronouns, and the ability to respond to self-focusing cues, which are fundamental for facilitating effective human interactions and environmental navigation. As AI systems evolve, the philosophical and practical considerations of equipping them with human-like traits of conscientiousness are gaining traction, heralding a burgeoning field of research (Huang et al., 2023c). For a thorough understanding of this area, interdisciplinary approaches that intertwine psychology, artificial intelligence, and ethics are instrumental.

  To seamlessly integrate into the human-centric world, AGI must possess an acute awareness of the beliefs, intentions, and desires of both themselves and others. This "Theory of Mind" (Premack and Woodruff, 1978) is a meta-ability that enables AGIs to understand and predict behaviors, facilitating smoother human interactions. This comprehension will allow for more nuanced and informed decision-making by AGIs, particularly in complex social contexts.

- **AGI holding certain persona.** Recent advancements reveal that LLMs can exhibit consistent personality traits, such as those categorized by the Big Five or MBTI frameworks, with models like ChatGPT often exhibiting traits aligned with the ENFJ type (Huang et al., 2023c). These models also tend to display certain cognitive thinking styles, with evidence suggesting an inclination towards holistic thinking in ChatGPT's responses (Jin et al., 2023c). Research efforts are increasingly directed towards intentionally imbuing LLMs with specific personalities, enabling them to demonstrate a variety of behaviors that are both diverse and verifiable (Caron and Srivastava, 2022; Jiang et al., 2022b).

- **AGI metacognition ability in self-evolving.** While the aforementioned research defines AGI in terms of easily measurable capabilities such as reasoning (Butlin et al., 2023; Morris et al., 2024), it may overlook the potential importance of meta-cognitive abilities such as self-evolution or self-awareness. Studies predominantly showcase this through the agent's iterative adaptation via task execution (Le, 2019; Wang et al., 2023e), code execution (Gao et al., 2020), or feedback from physical simulations (Qian et al., 2024; Xu et al., 2023a). Other strategies for self-evolution include prompt adaptation and optimization (Wang et al., 2023h; Aksitov et al., 2023), continuous improvement through error identification and self-reflection, and memory retrieval as a mechanism for short- or long-term learning. These approaches mainly emphasize the iterative refinement of tasks within a loop-structured framework based on LLMs. In contrast, recent advancements propose methodologies that address inter-task agent self-evolution, highlighting the significance of leveraging past experiences to effectively evolve AI systems (Qian et al., 2024; Xu et al., 2023a).

These traits are crucial for various reasons. First, self-awareness could enhance AGI's adaptive problem-solving abilities by allowing it to accurately assess its strengths and limitations, thus facilitating adjusting its strategies in real time when faced with new challenges. Second, the capacity for ethical and moral decision-making is increasingly imperative as AGI becomes more entwined with societal functions, necessitating a self-awareness component to enable navigation through complex moral dilemmas and ensure alignment with human values. Furthermore, the potential for AGI to autonomously evolve and adapt without human guidance promises greater efficiency and capability in the long term, possibly leading to an exponential increase in their abilities—a key characteristic of true AGI. Lastly, incorporating autonomous consciousness in AGI could yield more natural and effective human-AI interactions, enhancing collaboration and developing more intuitive user interfaces, particularly in scenarios requiring deep integration into human teams or societal structures.

AGI-level Metacognition The future of AGI metacognition is an exciting realm of possibilities that could dramatically expand the boundaries of artificial intelligence. One key area where AGI could demonstrate significant potential is in enhancing the theory of mind and social reasoning. Current AI struggles with understanding others' mental states, which is crucial for nuanced social interactions. Future AGIs could incorporate multi-modal input and advanced reasoning to model better the beliefs, intentions, and actions of others. For example, an AGI tutor with enhanced social reasoning could deeply understand each student's knowledge, learning style, and motivations, providing hyper-personalized guidance.

- **Future AGI has the potential to achieve genuine self-awareness and consciousness.** AGIs may one day possess deep self-awareness, capable of introspection, reflection, and grappling with existential questions. This would blur the lines between artificial and biological intelligence, raising philosophical and ethical questions. However, uncertainty remains about whether AI could achieve human-like consciousness; it may require integrating metacognition, introspection, and self-perception capabilities. Imagine an AGI companion that doesn't just converse but relates deeply emotionally, sharing in the human condition.
- **Substantial research should focus on AGI's potential for autonomous self-evolution and open-ended learning.** AGIs driven by curiosity and intrinsic motivation could rapidly self-improve, setting goals, innovating strategies, and pushing boundaries. They may exceed human intelligence in certain areas, generating novel insights and breakthroughs that propel fields forward. Picture an AGI scientist who tirelessly conducts experiments, forms and tests hypotheses, and makes discoveries at an unparalleled rate.

As we contemplate the implications and considerations surrounding AGIs with advanced metacognition, we are confronted with profound questions about consciousness, intelligence, ethics, and our place in the world. The exciting potential to integrate AGIs as empathetic companions, insightful advisors, and tireless innovators is balanced by the need to grapple with the implications of creating potentially superior beings and redefining the boundaries between human and artificial intelligence.

Realizing this vision of AGI metacognition will require substantial research and development to close current capability gaps. Nevertheless, the awe-inspiring potential and the philosophical challenges such a future would bring make this an exceedingly important area of AI progress to contemplate and work towards. As we stand on the precipice of this new era, it is crucial that we approach the development of AGI metacognition with a mix of enthusiasm, caution, and deep reflection on the profound implications for our world and our understanding of intelligence itself.

3 AGI Interface: Connecting the World with AGI

In the pursuit of developing AGI, a crucial aspect to address is its capability to interact with the external world. This interaction is facilitated through various interfaces that enable AGI systems to perceive, understand, and act within their environment, be it digital, physical, or intellectual. We summarize these three future directions in Figure 4.

3.1 AI Interfaces to Digital World

> *The Matrix is everywhere.*
>
> — *Matrix*

The concept of AGI interface into the digital world extends the scope by allowing agents to interact with digital environments, such as the Internet, databases, code, and APIs, and exhibit intelligent behaviors

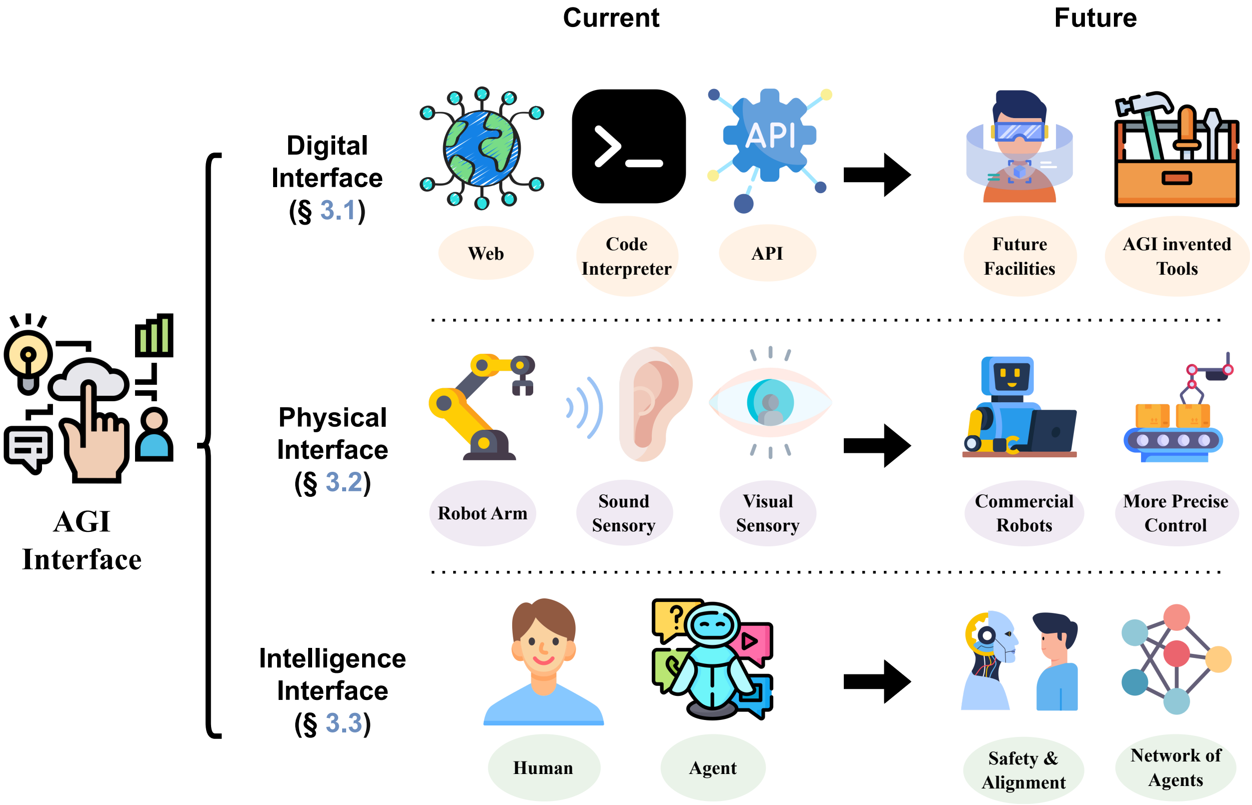


Figure 4: **The Interconnected Spheres of AGI Interface.** In the left part, we present some key elements in three interfaces: Digital (§ 3.1), Physical (§ 3.2), and Intelligence Interface (§ 3.3). On the right side of the figure, we outline several potential future aspects that could be significant.

similar to human-like behavior (Qin et al., 2023a). This interface serves as a crucial bridge for grounding AGI in complex, real-world scenarios, providing an indispensable platform for simulating and interacting with the multifaceted nature of human knowledge and experience. By facilitating AGI's engagement with real-world information structures and problem-solving contexts, this digital world interface accelerates the development of more versatile and robust artificial general intelligence capable of operating effectively across various domains.

Current State of AI Interface to Digital World Digital embodiment enables agents to interact dynamically and flexibly with the world. For instance, agents can utilize various APIs to navigate the web, search for relevant information, and construct personalized knowledge bases, allowing them to update their knowledge and adapt to new situations continuously. This approach drives the development of more advanced AI systems, particularly in natural language processing and reasoning capabilities.

- **Integrating digital tools in LLMs significantly enhances their capabilities and addresses inherent limitations.** Utilizing specialized tools augments their domain-specific expertise and increases decision-making transparency and robustness. The Toolformer model (Schick et al., 2023) demonstrates LLMs' ability to learn and effectively employ various external tools autonomously, with advanced learning methods mirroring human learning processes (Xi et al., 2023; Qin et al., 2023a). Models like Gorilla (Patil et al., 2023) connect LLMs with a wide array of APIs, highlighting the evolution towards greater autonomy and application versatility. LLMs are beginning to create and modify tools (Cai et al., 2024b; Qian et al., 2023b), leading to a future where agents exhibit increased self-sufficiency. This expansion in tool functionality facilitates multi-modal interactions and broadens the range of tasks LLMs can perform, aligning with the goals of embodied learning research (Zhuang et al., 2023).

- **LLM-based agents and frameworks demonstrate the capabilities of digital embodiment** (Zhou et al., 2024b; Wu et al., 2024b; Deng et al., 2023; Yang et al., 2023b). Mind2Web (Deng et al., 2023) allows for the comprehensive evaluation of agent generalizability in web scenarios, a critical aspect in creating robust and efficient web-based artificial intelligence. Voyager (Wang et al., 2023h) is an embodied agent in the Minecraft game that uses iterative prompting for dynamic reasoning and skill acquisition. To move forward, Generative Agents (Park et al., 2023), grounded in a sandbox game, have a memory stream that records experiences in natural language, enabling informed moment-to-moment behavior. Each of them focuses on different types of digital embodiment.

AGI-level Interfaces to the Digital World While the current state of AGI tool usage is advanced, it highlights several pivotal areas for reaching this goal.

- **AGI systems' creation of novel tools is nascent and limited, requiring a leap beyond human-designed frameworks for true autonomy.** Creating novel tools by AGI systems, as exemplified by the CREATOR framework (Qian et al., 2023c), is a groundbreaking step. Yet, the ability of AGI systems to invent tools autonomously remains nascent. These systems often rely on human-designed frameworks and algorithms, limiting their creative scope. True AGI would require a leap beyond this, enabling systems to ideate and engineer tools independently and intuitively.
- **Extending the scope of digital worlds.** There are still many opportunities to empower AGI systems with interfaces in different modalities and various environments, such as wearable computing, smart environments, mixed-reality settings, and emerging technologies like virtual reality (VR) and extended reality (XR). Although AGI will continue to exhibit promising performance in such interaction tasks, researchers need to explore potential solutions to ensure that AGI can yield beneficial results to humans while minimizing the cost of interaction. AGI should be able to seamlessly integrate with these technologies, leveraging their unique affordances to create more engaging and intuitive interactions. Moreover, AGI systems should be capable of adapting to novel interaction paradigms that may emerge in the future, ensuring that they remain relevant and valuable to users in the long term.

3.2 AI Interfaces to Physical World

> *In the twenty-first century, the robot will take the place which slave labor occupied in ancient civilization.*
>
> — *Nikola Tesla*

The integration of AI into physical entities is a crucial aspect of the pursuit of AGI. AGI in the physical world emphasizes learning through direct interaction with the environment and making an impact on reality, such as creating or modifying substances. In this section, we will explore the latest advancements in embodied AI in the physical world, including robotic control, navigation, and manipulation.

Current State of AI Interfaces to the Physical World The current state mainly lies in the interaction with robotic functionalities, understanding the potential for more intuitive human-robot interfaces, and emphasizing the importance of real-world datasets in advancing AI's practical applications.

- **Robotic control and action.** Recent advancements in robotic control and action including PaLM-E (Driess et al., 2023), RT-2 (Zitkovich et al., 2023), and Mobile Aloha (Fu et al., 2024b) demonstrate the potential for robots to interpret and execute complex, high-level instructions through natural language, providing a more intuitive interface between humans and robots. SayCan (Ahn et al., 2022) combines the semantic understanding capabilities of PaLM (Chowdhery et al., 2022a) with robotic affordances, enabling robots to understand abstract tasks and execute them in real-world environments. PaLM-E injects embodied observations into the language embedding space of a pre-trained language model for sequential robotic manipulation planning, while VIMA (Jiang et al., 2022a) processes multi-modal prompts and outputs motor actions autoregressively to control a robot arm. RT2 leverages large-scale, pre-trained vision-language models for robotic control.

- **Robotic navigation and interaction.** Effective navigation and interaction with the environment are essential for embodied AI systems to interface with the physical world. LM-Nav (Shah et al., 2023) combines pre-trained models of language, vision, and action for robotic navigation, marking a significant step towards more intuitive human-robot interaction. VoxPoser (Huang et al., 2023d) utilizes composable 3D value maps and language models for nuanced robotic manipulation, while LLM-Planner (Song et al., 2023a) harnesses LLMs to facilitate few-shot planning for embodied agents, enabling them to follow natural language instructions to accomplish complex tasks within visually-perceived environments. Gervet et al. (2023) presents a comprehensive evaluation of semantic visual navigation methods, finding that modular learning approaches achieve a high success rate in real-world home environments, demonstrating the effectiveness of these interfaces in navigating physical spaces.

- **Understanding and replicating human motion.** To create more natural and intuitive interfaces between humans and robots, it is essential to understand and replicate human motion. MotionGPT (Jiang et al., 2023b) likens human motion to a foreign language that AI can interpret, opening new pathways for understanding and replicating human-like movements in robots. Instruct2Act (Huang et al., 2023a) maps multi-modality instructions to robotic actions using large language models, showcasing the potential of LLMs in interpreting and executing diverse instructions. Furthermore, Perceiver-Actor (Shridhar et al., 2022) introduces a transformer-based model that can be trained end-to-end to map visual observations and natural language instructions to actions for robotic manipulation tasks, further enhancing the interface between human instructions and robotic actions.

  Integrating LLMs into embodied AI represents a vital step forward in the journey towards AGI. The research above demonstrates that LLMs have shown remarkable potential in enabling robots to understand and execute complex instructions, navigate environments, and manipulate objects, creating more seamless interfaces between AI systems and the physical world. By leveraging LLMs' semantic understanding and generalization capabilities, embodied AI systems can achieve intelligence and flexibility that mirrors human capabilities. For manipulation in unstructured environments, an approach that emphasizes integration, embodiment, feedback, and informed assumptions is more effective (Eppner et al., 2016).

- **Datasets.** The work presented by Khazatsky et al. (2024) is a significant contribution to the field of "in-the-wild" robotic manipulation, offering the DROID dataset which captures a wide range of real-world interactions. This dataset is particularly notable for its large-scale in-the-wild setting, featuring diverse environments and tasks that mirror everyday scenarios. Similarly, Li et al. (2024f) introduces BEHAVIOR-1K, another leap in AGI robotics benchmarks, focusing on human-centered activities within a realistic simulation to test the limits of autonomous agents in complex tasks. Both works represent cutting-edge efforts to benchmark and enhance the generalization capabilities of AI in the realm of long-horizon, real-world tasks, bridging the gap between controlled laboratory conditions and the dynamic and unpredictable nature of real-world interactions.

AGI-level Interfaces to the Physical World Looking ahead, the journey toward AGI with embodied intelligence encompasses several key areas for future research and development. A critical area is enhancing contextual and environmental understanding in AI systems. AGI needs to develop the capacity to interpret and adapt to dynamic, real-world environments with a level of sophistication at least comparable to human cognition. This advanced environmental analysis and perception capability is essential for AGI to effectively navigate and interact with complex, ever-changing physical surroundings, mirroring the adaptability and intuition that humans exhibit in diverse situations.

- **Enhancing multisensory integration and interaction capabilities is key to developing more relatable and effective AI systems.** This includes advancing the synergy between visual, auditory, linguistic, and tactile inputs, enabling AI systems to comprehensively perceive their surroundings. Improving interaction capabilities, such as natural language processing and human-like movement, will also make AI systems more relatable and effective in human-centric environments.

- **Advancing affordable robotic manufacturing is crucial for democratizing embodied AI and fostering innovation across sectors.** The development of more advanced yet affordable robotic manufacturing techniques is vital. Lowering the cost of robot production without compromising quality or functionality can democratize access to embodied AI, enabling widespread adoption and innovation across various sectors. This, in turn, could accelerate the deployment of intelligent agents in everyday scenarios, from domestic assistance to industrial automation.

- **Edge computing's efficiency is crucial for embodied AI applications, enabling real-time decisions and immediate responses in dynamic environments.** The efficiency of edge computing, particularly in on-device inference speeds, plays a critical role in the practicality of embodied AI applications. Faster, more efficient processing at the edge allows AI systems to make real-time decisions based on vast amounts of data from their surroundings without the latency associated with cloud computing. This capability is essential for tasks requiring immediate response and adaptation to dynamic environments, such as autonomous driving and real-time navigation in crowded spaces.

3.3 AI Interfaces to Intelligence

> *The future of work lies in the collaboration between humans and AI, where technology enhances our natural abilities, allowing us to think more strategically and creatively and empowering us to drive innovation in the workplace.*
>
> *— Demis Hassabis, CEO and co-founder of DeepMind*

Integrating AI with other intelligent entities, whether artificial or human, is a critical aspect of achieving AGI. Interfacing with intelligence allows for exchanging knowledge, collaboration, and enhancing overall system capabilities. In this section, we will explore two main categories of interfaces to intelligence: interfaces to AI agents (3.3.1) and interfaces to humans (3.3.2).

3.3.1 AI Interface to Other AI agents

There are generally two categories to improve the overall system for integrating one AGI system with others. The first aspect focuses on the teaching process among AGI models through a sequential interaction between different models. The second emphasizes the simultaneous collaboration between these models, connecting different agents to form a comprehensive and robust AGI system.

Current State of AI Interfaces to Other AI Agents The interfaces to other AI agents include both sequential and parallel interactions, where the agents act as teachers, learners, collaborators, or communicators.

- **Agents as teachers and learners.** On one hand, stronger AGI models often act as oracles to provide 'supervision' to inferior ones, from tuning on data from better models (Taori et al., 2023; Gu et al., 2023) to prompt-engineering-based approaches (Huang et al., 2022a; Jiang et al., 2023a; Fu et al., 2023b), there emerges the concept of model knowledge distillation. In the field of language processing, it is common to use a teacher system to label and expand existing data by directly taking the teacher's answer (Gilardi et al., 2023; Hsieh et al., 2023; Li et al., 2022a; Sun et al., 2024b; Ding et al., 2023a) or using more advanced techniques such as CoT prompting (Ramnath et al., 2023; Li et al., 2023g), or creating new data for subsequent models to distill useful and compact knowledge from large-scale data (Li et al., 2023b; Javaheripi et al., 2023). Similar paradigms are applied in computer vision and multimodal domains for better model training and deployment. One of the most prevailing methods is utilizing the GPT-4V model to label answers in various tasks (Shu et al., 2023; Liu et al., 2023d; Li et al., 2023j).

  In conventional solutions, a better AGI model severs as the role of the teacher, however, there is a growing trend for less competent AGI models to provide insights for aligning stronger ones with or beyond human perception, known as superalignment (Burns et al., 2023). This has proven to be effective in empowering higher capacity than the teacher model with a vanilla fine-tuning strategy and distilled data from the smaller model. Recent studies have begun to explore how leveraging weaker models to enhance the performance of stronger ones can be applied across various domains. Chen et al. (2024a) introduced the innovative idea of iteratively harnessing the capacity of weaker models to improve the efficacy of more powerful counterparts. (Ji et al., 2024) developed an efficient alignment paradigm that learns the correctional residuals between aligned and unaligned responses from a weaker model. (Sun et al., 2024c) employs a less advanced system, trained on simpler tasks, to guide more capable models for tackling hard reasoning tasks (*e.g.*, level 4-5 MATH problems). In the vision domain, (Guo et al., 2024a) presented compelling evidence that weak-to-strong generalization may outperform finetuning methods in certain scenarios. This interaction between different models requires acquiring knowledge from selected models step-by-step, forming the sequential interface.

- **Agents as collaborators or communicators.** On the other hand, off-the-shelf AGI interfaces can facilitate the integration of various models into a comprehensive and efficient system, allowing for simultaneous collaboration and knowledge sharing. In single modality domains, there is a growing trend of combining different language models into an integrated system for improved language modeling, such as the Mixture-of-Expert (MoE) system (Fuzhao Xue and You, 2023). The key advantage of this approach is the significant reduction in deployment efficiency, achieved through model parallelism, dynamic expert model selection, and token routing. As a result, the MoE system can reduce memory requirements and computational budgets compared to similar scale LM systems (Chen et al., 2023g; Chowdhury et al., 2023). By utilizing a gate coordinator to plan for different expert language models, the system can generate specialized and high-quality responses for different aspects of queries, leading to higher efficiency and better handling of tasks (Du et al., 2022). In the multi-modal domain, models with the ability to understand more than one modality can collaborate with other AGI interfaces to create a system with a more comprehensive understanding of related visions. For instance, LLaVA-Plus (Shilong Liu, 2023) builds upon LLaVA, an end-to-end trained vision language model, by incorporating newly constructed data to enhance its tool-using skills beyond its original visual understanding and reasoning abilities.

  Agent-based works also equip the ability to solve multi-modal problems using various external tools for such purposes. Numerous conventional algorithms facilitate the coordination of multiple agents or robots in either physical or simulated settings (Sunehag et al., 2017; Gupta et al., 2017; Fioretto et al., 2018; Foerster et al., 2018). Furthermore, advancements have been made in developing techniques that accelerate communication among multi-agent, thus enabling them to work towards a common goal across a variety of tasks (Sukhbaatar et al., 2016; Jha et al., 2024; Qian et al., 2023a; Hong et al., 2023b; Liu et al., 2023k). Works by Li et al. (2023e) and Chen et al. (2023f) have established a range of common scenarios for multi-agent interactions, featuring vivid visualizations and the integration of human interaction. Building upon the existing frameworks for multi-agent systems. Chen et al. (2023c) innovatively propose the automated creation of new agents to navigate dynamic environments effectively. Additionally, Hong et al. (2023b) and Zhou et al. (2023a) have integrated Standardized Operating Procedures (SOPs) to streamline the customization and deployment processes of multi-agent systems. The underlying principle of these systems is utilizing advanced LLMs as coordinators, aimed at efficiently addressing a myriad of tasks within simulated environments. Recent developments in the concept of Natural Language-Based Societies of Mind (NLSOMs) (Zhuge et al., 2023) have revolutionized the understanding of cooperation between neural networks to the “Mindstorm” metaphor. This framework employs the language interface to conduct communications between agents, allowing for easy and straightforward adaptation of novel modules.

  Looking beyond multi-modal tasks, more specific and advanced agent systems that can handle more complex computer applications such as web browsing (Mialon et al., 2023; Deng et al., 2024; Zhou et al., 2024b), software manipulation (Kapoor et al., 2024; Rawles et al., 2023; Yang et al., 2023b), and gaming (Ma et al., 2023b; Wang et al., 2023h;b; Xu et al., 2024) have emerged. In the gaming realm, agents have been deployed in simulated interactive environments, including Minecraft (Wang et al., 2023h;b), Starcraft II (Ma et al., 2023b). These agents receive textual observations from internal APIs and execute predefined semantic actions. However, these domain-specific applications limit the agents’ ability to generalize to other games or broader software applications. Tan et al. (2024) proposes the concept of General Computer Control (GCC) and adopts a more intuitive approach, employing multimodal input from screenshots to generate keyboard and mouse commands within Red Dead Redemption II. This setting holds promise for expansion into more intricate computer tasks. While Wang et al. (2023a) attempts to interact with the environment in a human-like manner using screenshots as input and controlling mouse and keyboard actions, its action space is constrained to a predefined hybrid space. Despite achieving notable results in specific tasks, these methods lack the ability to generalize across diverse tasks due to inconsistencies in observation and action spaces. Several research efforts (Zhang et al., 2024b; Gao et al., 2023b; Kapoor et al., 2024; Niu et al., 2024; Wu et al., 2024b) have aimed to enhance the scalability of web agents by utilizing screenshots as input and keyboard and mouse operations as output, allowing them to interact with a wider range of applications.

AGI-level Interfaces to Other Agents While current interfaces to AI models demonstrate effectiveness, they are primarily focused on a narrow range of applications (Askell et al., 2021; Memarian and Doleck, 2023). To enhance their capabilities, we propose the following improvements in the future:

- **Advanced interface encodes unified representations.** The integration of the AGI interface is paramount for enhancing model performance. These interfaces can supervise less capable models in sequential interactions and introduce a range of novel capabilities in parallel cooperation. A key focus should be developing comprehensive and lightweight AGI interfaces that exhibit strong generative and understanding abilities. Future advanced interfaces should be able to encode a unified representation across all modalities. This would simplify the aligning process and optimize resource allocation for various downstream tasks, such as generation and identification.
- **High-quality connection promotes effective communications.** To ensure that AGI-level interfaces among agents lead to effective collaboration, it is essential to establish high-quality connections that are both reliable and efficient. The recent concept of weak-to-strong alignment is particularly noteworthy, underscoring the significance of determining the most effective methods for incorporating external capabilities into existing systems. Traditionally, tuning model parameters has been the primary method for generalizing systems to specific tasks or domains. However, recent research highlights the potential of approaches that require minimal or no parameter tuning (Burns et al., 2023; Zhao et al., 2023c). Moreover, incorporating in-context learning and context-aware communication among agents can enable agents to adjust their interactions based on the situational context, improving the relevance and efficacy of their collaboration.
- **Interaction protocols ensure safe interaction.** It's important to create robust and effective interaction protocols to serve as the foundation for safe communications and actions between different AGI entities. To achieve this, it's important to implement standardized security measures, including advanced encryption methods, authentication protocols, and content filters specifically designed to safeguard against the dissemination of misinformation or malicious content. Additionally, developing guidelines for safe AGI actions and ensuring that all activities are performed within the bounds of ethical norms and regulations is essential. The focus on safety protocols enhances the security of interactions between AGI systems and builds trust with end-users, paving the way for broader acceptance and integration of these technologies into everyday applications.
- **Advanced agent network promotes cooperative learning.** Enhancing the network of AGI agents to foster social and cooperative learning is essential for the advancement of collective intelligence. By enabling agents to share insights, strategies, and knowledge, we can facilitate a more rapid and efficient learning process across different domains. Social learning mechanisms, such as imitation and observation, can be incorporated to enable agents to learn from the successes and failures of their peers. Furthermore, cooperative learning models can be designed to encourage agents to work together towards common goals, harnessing their diverse strengths and capabilities. Such a networked approach not only accelerates the pace of innovation but also leads to the developing more versatile and adaptive AGI systems capable of tackling complex, real-world problems through teamwork and collaboration.

3.3.2 AI Interfaces to Humans

Human intelligence has been the ultimate goal of AI, and human beings have also been the primary beneficiaries of AI. As we move towards AGI, we should empower AI with the capabilities to interact with humans to ensure that AI can actually benefit humans. Therefore, we call for future advances in interface technologies to lay solid foundations for AGI's capabilities to interact with humans.

Current State of AI Interfaces to Humans Developing interfaces with artificial intelligence has been explored in Human-Computer Interaction (HCI) research for a long time. We discuss current research in human-AI interfaces, including both graphical interfaces and multimodal interfaces, as well as general principles.

In history, there have been many related principles or guidelines for designing interfaces for human-AI interaction. Most of the existing frameworks in HCI research to interact with human are based on the idea of *augmenting* (rather than *replacing*) human intelligence with artificial intelligence (Engelbart, 1962). Therefore, maintaining human agency and reflecting human values have been consistent themes in designing human-AI interaction. Researchers also argued that the benefits of allowing AI agents to take the initiative and automate users' routines versus the benefits of waiting for users' direct manipulation would need to be carefully weighed (Horvitz, 1999; Shneiderman and Maes, 1997). With advances in artificial intelligence, researchers have articulated 18 generally applicable design guidelines for human-AI interaction spanning different phases in user interactions (Amershi et al., 2019). Recent research also presents six principles for designing generative AI applications that address unique characteristics of generative AI UX and offer new interpretations and extensions of known issues in designing AI applications (Weisz et al., 2024). Such guidelines could serve as a resource for the principles of the future design of AI-infused interfaces, optimizing interaction performance and improving the interaction experience.

- **Graphical user interfaces.** One emerging line of research has focused on designing interfaces to support user tasks based on textual or visual interactions, which will lower the "threshold" while raising the "ceiling" in terms of the quality and the diversity of user tasks (Myers et al., 2000). A common theme in developing such interfaces is to create potential wrappers beyond simply providing "straightforward" input and output (Jiang et al., 2023c; Suh et al., 2023; Gero et al., 2024; Suh et al., 2024). For instance, researchers tried to use interactive diagrams to support humans in dealing with information-seeking and question-answering tasks powered by large language models (Jiang et al., 2023c). Another thread of research is to identify possible workflows or strategies that can unlock the potential of AI during human-AI interaction (Wu et al., 2022b;a; Brade et al., 2023; Arawjo et al., 2023; Leiser et al., 2024; Kim et al., 2024; Feng et al., 2024). For example, previous researchers introduced the notion of chaining multiple LLM prompts together to help users accomplish complex tasks with LLMs, which enables humans to take advantage of LLM's ability to handle a variety of independent tasks (Wu et al., 2022b;a). In addition, when interacting with humans, large language models could encounter various non-language input or output data, such as direct manipulation action traces, vector graphics, or application states (Aveni et al., 2023; Duan et al., 2024). For example, researchers attempted to create alternative representations of context information to leverage the capabilities of large language models in different interaction tasks, auto-completion of forms (Aveni et al., 2023).

- **Multimodal user interfaces.** Many researchers are actively exploring integrating AI with existing interaction techniques to enrich user experiences across different modalities and for different groups. On the one hand, previous research has created many novel sensing technologies and interaction techniques beyond simple textual and visual interactions. Recent advances in multimodal foundation models have shown great promise in many interaction tasks. For example, GPT-4o possesses remarkable capabilities of reasoning across audio, vision, and text in real time OpenAI (2024). In the future, it is worth exploring the possibility of empowering human-level AI with the capabilities to interact with humans through different modalities (Li et al., 2024d; Lin et al., 2024d). From this perspective, previous researchers have proposed a novel pipeline that provides generalized predictions of follow-up actions for real-world multimodal sensory inputs, leveraging the explicit reasoning of LLMs on structured text converted from multimodal data to ground the predicted actions (Li et al., 2024d). Meanwhile, AI could also be an important part of user experiences in mixed reality. Recent research in mixed reality provides abundant opportunities for user interfaces that could be driven by large language models (Bozkir et al., 2024). Additionally, researchers are actively exploring inclusive interfaces to ensure everyone can benefit from interacting with AI. One area of focus is creating better interfaces for people with disabilities in the field of accessibility research (Huh et al., 2023; Valencia et al., 2023). In recent work, researchers have created interactive systems that allow blind or low-vision creators to generate images by providing rich visual descriptions of the generated outcomes in language (Huh et al., 2023).

AGI-level Interfaces to Humans Researchers pointed out unique challenges in designing human-AI interactions due to the uncertainty surrounding AI's capabilities and the complexity of AI's outputs (Yang et al., 2020). In the future, there will still be important challenges that we will need to overcome to truly empower AGI with the capabilities of interacting with humans (Bigham, 2023).

- **Ensuring the benefits in different environments.** Future research needs to figure out strategies for AGI to benefit humans in what they want to do. AGI will have more capabilities comparable to intelligent humans, bringing forth numerous possibilities that can benefit real humans. However, if the cost of using AGI exceeds its benefits, people are unlikely to derive value from it (Horvitz, 1999). Looking ahead, there are many opportunities to design AGI-level interfaces in different modalities and various environments, such as wearable computing, smart environments, and mixed-reality settings. However, researchers still need to explore potential solutions to ensure that AGI could actually yield beneficial results to humans that outweigh the cost.
- **Maintaining the controllability for different people.** Future work should explore how we can support diverse people, even with limited AI literacy, to interact with AGI in a controllable manner. Compared with AI experts, it could be difficult for users who lack an AI background to understand AI's full capacities and mechanisms, which can lead to unexpected pitfalls in interactions. The advances in AGI would continue to amplify similar concerns. In the future, it is increasingly important for us to think about potential solutions to empower users with the capabilities to understand and control the interactions with AGI.
- **Managing the risks at different scales.** As we architect novel interfaces for AGI, future researchers should consider the potential implications these interactions will have at individual, community, and societal levels. Previous research has continued to identify the impacts of human-AI interaction on people's behavior and mentality. Though these issues are not necessarily unique to AGI, they will be amplified and more prevalent as we empower AI with more human-level capabilities in the real world. Future work still needs to focus on approaches to understand and mitigate the impacts of interface design for AGI at different scales so that we can maximize their positive impacts and minimize their negative impacts.

4 AGI Systems: Implementing the Mechanism of AGI

> *Dangers lurk in all systems. Systems incorporate the unexamined beliefs of their creators.*
>
> — *Frank Herbert, God Emperor of Dune*

The emergent behaviors (Wei et al., 2022a) exhibited by many large models such as Llama 2 (Touvron et al., 2023), GPT-4 (OpenAI, 2023a), Gemini (Team et al., 2023), Claude 3 (AI, 2024), and Mistral Jiang et al. (2023d) appear when the number of parameters in a model gets scaled up to a certain amount. The underlying workhorse that enables this scaling while retaining sufficient efficiency of LLMs is a range of system efforts: 1) *scalable model architectures* fundamentally and algorithmically define the computation and modeling, 2) *large-scale training* techniques optimize the utilization of more hardware accelerators, potentially spread out geographically, 3) *inference infrastructures* ensures stable and high-throughput serving of multiple models, 4) *cost and efficiency* discusses various methodologies in making data, model combination, and automation process much more efficient, and finally we touch some aspects on 5) *hardware computing platforms* which attempt to break soft physical constraints and therefore, provide the next generation computational capabilities and hardware foundation for future algorithmic innovations.

Advancements in system research are essential for facilitating this scalability, a trend that is anticipated to remain pivotal as we step towards artificial general intelligence. With continual improvement in AI system

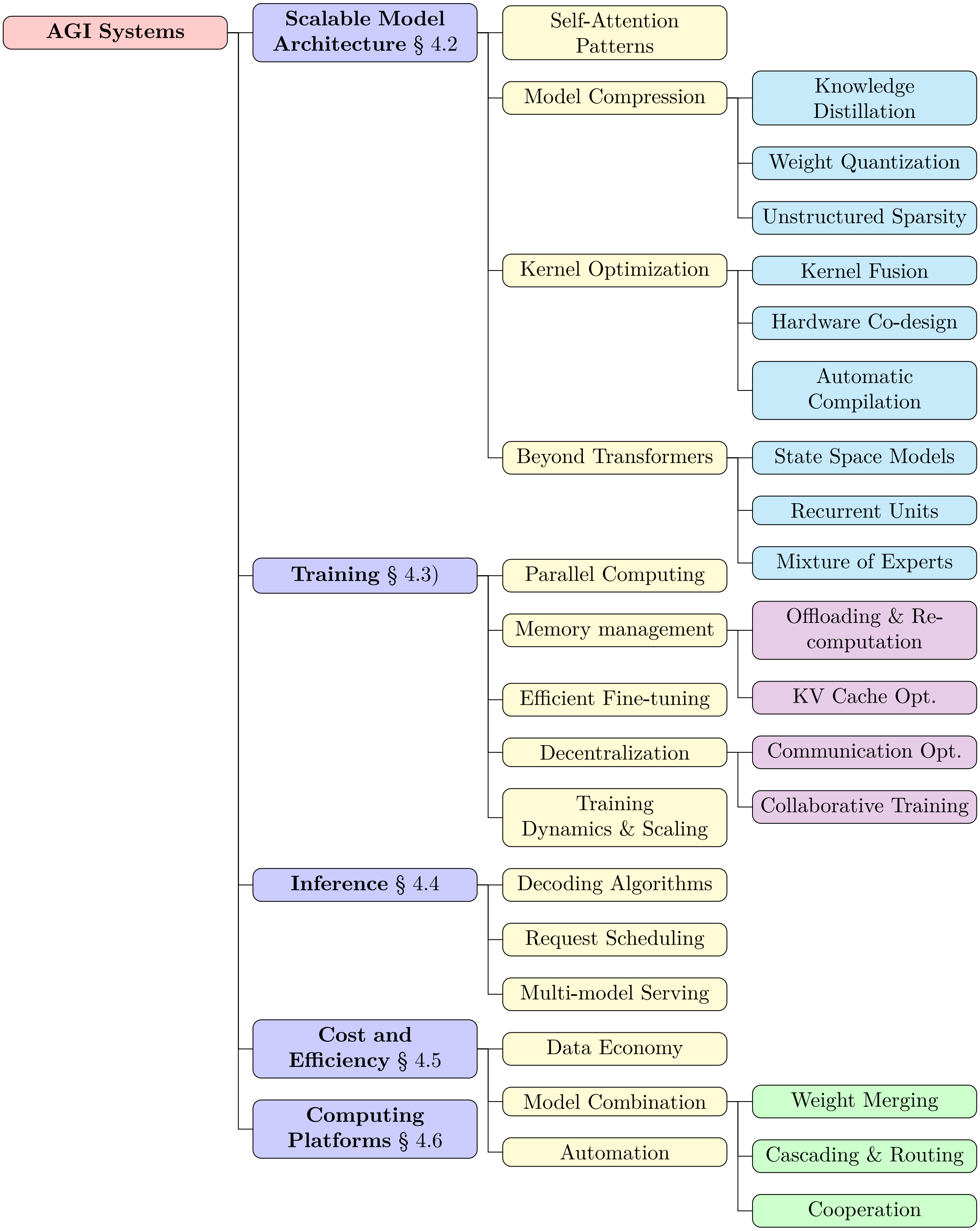


Figure 5: **Taxonomy of Current AGI Systems.** We discuss several advancements in various categories of AGI systems, including scalable model architectures, large-scale training, optimized inference techniques, methods for reducing cost and improving efficiency, as well as next-generation computing platforms for AI.

research and engineering, we could envision that models are trained over ten thousand heterogeneous accelerators across the cloud, which not only connects multiple agents into a single multi-functional amalgamation but also provides instant and personal assistance to people in everyday life.

In this section, we start with the major challenges that AGI systems need to tackle and introduce some prior efforts on model architecture innovation, training, inference, cost reduction, and computing platforms. Finally, we will conclude by envisioning what AGI systems would look like and their roles in the future.

4.1 System Challenges

We first briefly describe and categorize major system challenges in this section:

The Large Amount of Training Data Large models require a lot of training data to achieve Chinchilla optimality (Hoffmann et al., 2022). At the same time, we can envision that the raw amount of data available on the internet will likely skyrocket while the average quality and authenticity might not improve as fast, partly due to the massive success of generative models and user content creation. This demands a more sophisticated and automated data processing pipeline that can select, structurize, clean, and mix data from different sources for efficient training.

The Speed and Cost of Iteration Each iteration of large model training can take enormous resources and time during prototyping and experiments. In practice, there might be many other interference, such as human and system errors, that will result in training preemption. Automatic (hyper-parameter & architecture) search pipeline and well-designed training infrastructure (Shoeybi et al., 2020; Aminabadi et al., 2022) can drastically reduce iteration cost and implicitly improve model development speed.

Privacy-sensitive and Resource-constraint Settings While the current most successful large models are deployed in data centers where requests from users are processed in a centralized manner (Achiam et al., 2023; AI, 2024), the need for serving models on edges where data and queries are used and processed locally in more privacy or latency-sensitive situations will become more substantial. However, edge devices are usually less capable in terms of computing and memory, which motivates developing techniques that optimize the utilization under resource-constraint settings and, at the same time, do not compromise model performance.

Efficient Methods for Fine-tuning and Adaptation Fine-tuning pre-trained models on task-specific data has been the most popular paradigm. Despite this, the computing requirement and time for full model weight updating is still prohibitive for many users. Efficient fine-tuning methods help reduce the barrier to domain adaptation, agent training, and task-specific optimization.

Serving Latency and Throughput AGI systems need to support low latency and high throughput for seamless user experiences and engagement. However, current systems often trade-off one with others such as optimizing batch processing, the time to first token, or single query completion time (Yu et al., 2022; Aminabadi et al., 2022). Striking a balance among all these metrics is a challenging question.

Memory Footprint One salient challenge for deploying large models is the memory footprint, which becomes even more severe for long context and multi-modal inputs due to the quadratic nature of self-attention. KV cache is the common technique for trading off memory for faster inference (Pope et al., 2022) and will also incur a significant memory burden if not handled gracefully.

Hardware Compatibility and Acceleration The performance of model serving heavily depends on how well engineers can leverage the hardware's capability. Specialized kernels and algorithms designed for different accelerator architectures can substantially boost the inference speed. Being compatible with heterogeneous devices and creating uniform software abstraction can help fully unleash the potential of large models.

In the following sections, we'll discuss advancements in the major system categories and relate how these prior efforts help address some of the above-mentioned challenges.

4.2 Scalable Model Architectures

One everlasting topic that system researchers and engineers are trying to deal with is how to make larger and more powerful models. However, there are several axes to consider for scaling: *the number of parameters and* *training data volume* (Chinchilla scaling law (Hoffmann et al., 2022)), as well as *the effective context length and serving capacity* (Long context scaling law (Xiong et al., 2023)). To this end, we start with introducing optimization techniques from prior works on the model architecture level and move on to training and inference infrastructures that are more model-agnostic.

Self-attention Patterns Vanilla self-attention, the core mechanism under transformer architectures (Vaswani et al., 2017), scales quadratically with the input sequence length, drastically limiting its potential on long context tasks with current hardware memory capacity. However, the full causal mask used in self-attention might not be efficiently optimal (Farina et al., 2024), wasting computation and time. The lottery ticket hypothesis Frankle and Carbin (2019) and the structured sparsity observation (Dong et al., 2023a) both suggest potential structural and computational redundancy in LLMs. Inspired by these works, many works explored different attention mask patterns to reduce the computations and memory requirement over the fully casual one. Sliding window Beltagy et al. (2020); Jiang et al. (2023d) and dilated patterns Ding et al. (2023b) either limit to local attention or reduce the resolution, often resulting in improving efficiency with some performance degradation. Another line of work observes that some tokens in the sentence are semantically more important and hence introduce different kinds of global tokens for efficient attention, such as the initial tokens (Xiao et al., 2023) and the special landmark (Mohtashami and Jaggi, 2023) for each block of tokens. It is worth noting that the computation bottleneck might switch between the self-attention and fully connected layers for models with different scales, and blindly applying heuristics-based sparse patterns might not only give marginal speedup but also incur a loss in performance, which motivates future research in more complex and adaptive efficient patterns.

Model Compression The goal of model compression is to reduce the memory footprints, computational complexity, and deployment cost of large models. Among many approaches, Knowledge Distillation (KD) (Hinton et al., 2015; Hsieh et al., 2023) is the process of extracting the learned knowledge of a bigger teacher model to a smaller student model that can often be served more efficiently. On the one hand, black-box LLM knowledge distillation only requires querying the teacher model (via API calls) and collecting input and output pairs which can be used to train the student model. Following this line, Alpaca (Taori et al., 2023) only costs around 100 US dollars to balance efficiency and performance by distilling from ChatGPT. Vicuna (Peng et al., 2023b) shows promising performance on instruction fine-tuning with knowledge distillation from GPT-4. On the other hand, MiniLLM (Gu et al., 2023) explores white-box knowledge distillation where more information like model weights and loss values of the teacher model is accessible, which can potentially stimulate better knowledge transfer. MiniLLM proposes to replace the standard KL divergence objective with reverse KL which prevents the student model from overestimating the low-probability regions of the teacher distribution. Generalized Knowledge Distillation (GKD) (Agarwal et al., 2024) breaks the traditional supervised KD training region by taking advantage of student's self-generated output sequences and leveraging feedback from the teacher on such sequences. This not only helps mitigate the distribution mismatch between training and inference but also has been proven to be more useful when the student model lacks enough capacity to fully mimic the teacher's behaviors. Balancing the level of access to the teacher model will be remained as a relevant topic for algorithmic design, safety, data privacy, cost, and efficiency. It is also exciting to explore the possibility of reverse-learning or super-alignment [2] where we want to distill knowledge from weaker models that can be leveraged (e.g., through analyzing, merging, and adaptively updating) to improve the current one.

Another line of work for explicit model compression is to take advantage of the model sparsity and prune parts of the model weights during inference. Similar to different attention pattern designs, this direction is mostly motivated by several empirical observations like lottery tickets and contextual sparsity hypotheses. However, it is non-trivial to apply pruning to LLMs without careful consideration of many aspects, such as the need for extra steps of fine-tuning, system overhead due to dynamic architectures, and the trade-offs among implementation difficulty, discernible efficiency gain, and the potential compromised performance. ZipLM (Kurtic et al., 2023) structurally prunes the model by iteratively identifying and removing components with the worst loss-runtime trade-off given a dataset, an inference environment, and speedup objectives. LayerDrop Fan et al. (2019) introduces the structured dropout which allows efficient pruning via selecting

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

sub-networks of any depth from the network without extra fine-tuning. Deja Vu (Liu et al., 2023g) predicts the contextual sparsity of a given input which guides the selection of only specific attention heads and MLP parameters during inference. FlashLLM (Xia et al., 2023b) solves the unstructured pruning with memory-efficient SpMM implementation on tensor cores.

Kernel Optimization Kernels are developed to speed up primitive computations such as general matrix multiplication for large networks. Kernel fusion is the technique of merging two or more kernels into one, which can reduce the overheads of kernel launching and redundant memory access. This has been widely implemented in many inference engines like LightSeq (Wang et al., 2021), FasterTransformer (NVIDIA, 2023a), and ByteTransformer (Zhai et al., 2023) where the instantiations mostly consist of grouped GEMMs, layer normalization, activation calculation, and self-attention Dao et al. (2022); Dao (2023). Most works following this direction leverage the GPU memory hierarchy and try to hide memory latency and maximize thread occupancy.

With the increasing demand for highly optimized kernel implementations for various computation patterns, automatic compilations have become a very active area of research, an approach that provides greater flexibility than sourced libraries for highly optimized kernel implementations (e.g., CUTLASS, cuBLAS, and cuDNN). Within this category, TVM (Chen et al., 2018), Triton (Tillet et al., 2019), and JAX (Bradbury et al., 2018) harness the potential of hardware accelerators by compiling into highly efficient low-level code, and at the same time, provide python interface for fast and easy prototyping. This not only greatly lowers the learning curve for writing custom kernels but also provides an abstraction for code adaptation to other computing platforms and backends with heterogeneous devices.

Beyond Transformers Despite the enormous success of the ubiquitous transformer architecture, many works attempted to find other designs to overcome some of its shortcomings. Mixture of Experts (MoEs) (Shazeer et al., 2017; Roy et al., 2020) replace the dense layers in transformer models with a conditional module consisting of multiple “expert” sub-networks. A routing mechanism is used to dynamically decide which expert(s) to use on the token-level (Zhou et al., 2022a; Fedus et al., 2022) or task-level (Kudugunta et al., 2021). Despite having multiple experts, sparse MoEs can often train and decode faster with the same model size and are expected to specialize in different abstract tasks (Jiang et al., 2023d; Hwang et al., 2023; Gale et al., 2022). However, MoEs also raise other system challenges during inference such as higher requirements for loading all experts into the VRAM and distributing experts over multiple nodes.

State space models (SSMs) have recently been applied to model sequence-to-sequence transformations (Gu et al., 2022a; Gupta et al., 2022; Smith et al., 2023), which can be readily used in various model architecture topology to replace the quadratic self-attention mechanism. A (discretized) SSM defines a recurrence relationship along each time-step (token) via a tuple of learnable parameters $(\Delta, \bar{A}, \bar{B}, C)$ and the major challenge that most works try to solve is how to compute this recurrence in a parallelizable way that can efficiently use modern hardware accelerators (e.g. FFTConv). The simplest form in this category is linear attention (Katharopoulos et al., 2020; Yang et al., 2023e), which can be viewed as a degenerate SSM. At its core, linear attention expresses self-attention as a linear dot-product of kernel feature maps and makes use of the associativity property of matrix products to reduce the complexity down to linear. S4 (Gu et al., 2022a) parameterizes the SSM both expressively and efficiently via a low-rank correction, allowing it to be diagonalized stably and reducing the SSM to the well-studied computation of a Cauchy kernel. There are many other following works (Gupta et al., 2022; Gu et al., 2022b; Smith et al., 2023) after S4 which attempt different parameterization of the transition matrix $\bar{A}$ (and others) to improve both the computational efficiency and modeling capacities. H3 (Fu et al., 2023a) proposes an SSM block that consists of two stacked separate SSMs that are specially designed to meet the challenge of recalling earlier tokens and support token comparisons across sequences. Hyena (Poli et al., 2023) generalizes H3 by replacing the S4 layer with an interleaved and implicitly parameterized long convolutions and data-controlled gating, which disentangles parameter size from the filter size and hence allows for greater expressivity. (Sun et al., 2023a) proposes Retentive Network a foundation architecture that includes additional gates and uses a variant of multi-head attention, achieving impressive constant inference cost and linear long-sequence memory consumption. RWKV (Peng et al., 2023a; 2024a) is a new architecture that takes advantage of the efficient parallelizable training of Transformers with the efficient inference of RNNs. In essence, the main “WKV” operation involves linear time invariance (LTI) recurrences, which can be interpreted as a ratio of two SSMs (Gu and Dao, 2023). To tackle the key weakness of previous SSM models which is their inability to perform content-based reasoning, Mamba (Gu and Dao, 2023) proposes the selective state space that can make the SSM parameters as functions of the input, effectively turning the SSM from LTI to time-varying. Despite no longer being able to apply efficient convolutions, they designed hardware-aware parallel algorithms for recurrence computation called Parallel Associative Scan, which enables it to achieve over 5× higher throughput than Transformers, SoTA performance across several modalities, and keep improving on real data up to million-length sequences.

Sparkles of interest for revisiting recurrent neural networks (RNNs) have also emerged with its major advantage in long context processing (i.e. linear time and constant memory for the hidden state). One of the challenges for RNNs is to scale the training and inference efficiently. (De et al., 2024) introduces Hawk, an RNN-based model with gated linear recurrences, and Griffin which mixes gated linear recurrences with local attention. They showcase the superior performance of Hawk against Mamba on downstream tasks, and Griffin matches performance of Llama-2 with six times fewer training tokens. Not only do they demonstrate the potential of long context capability, but also explain how to effectively leverage hardware accelerators during distributed training and inference by scaling Griffin to 14B parameters. Right after that, a family of models called RecurrentGemma (Botev et al., 2024) came out with various model sizes, in both pretrained and instruct-tuned versions. These advancements present the possibility of training a data-efficient, fixed state size, long context, and expressive model without relying solely on transformer architectures.

Recent works also explored high-level architecture hybridization strategies that wish to bring the benefits from different variants. (Lieber et al., 2024) proposes to combine Transformers with Mamba by interleaving layers, which achieves impressive results on both standard and long context tasks with manageable resource requirements. Beyond manual design, MAD (Poli et al., 2024) integrates the process into an end-to-end pipeline consisting of small-scale capability unit tests predictive of scaling laws. MAD successfully finds an efficient architecture, called Striped Hyena, based on hybridization and sparsity, which outperforms state-of-the-art Transformer, convolutional, and recurrent architectures (Transformer++, Hyena, Mamba) in scaling, both at compute-optimal budgets and in over-trained regimes. These works will likely continue to inspire further explorations in architectural designs that are both performant and efficient at scaling, breaking through the current Transformer paradigm.

4.3 Large-scale Training

Scaling the training of large models encounters many challenges with modern hardware, such as the fact that models can no longer fit into a single GPU due to the increasing memory requirement, accelerating the training speed with more computing units while incurring minimal overheads (linear scaling), and leveraging disaggregated resources, etc. In this section, we give an overview of several works that enabled large-scale pre-training and efficient fine-tuning for downstream task adaptation, with a gentle introduction to motivate many possibilities with decentralized training.

Parallel Computing Parallelism for large language models in a clustered environment with multiple computing units can often be characterized into four major modes, often known as "4D parallelism". *Distributed data parallel* (DDP) is the simplest setup where the model is replicated across units, and the data is sliced and fed to each model, typically (implementation-specific) with a synchronization step at the end of each pass. More sophisticated versions of DDP like ZeRO and FSDP are used ubiquitously in modern large training frameworks such as DeepSpeed (Aminabadi et al., 2022), FairScale (FairScale authors, 2021), and Megatron-LM (Shoeybi et al., 2020). *Tensor parallel* (TP) or *model parallelism* splits the model weights into multiple chunks which are distributed across GPUs. This horizontal splitting allows data to be processed in parallel across sharded weights and then the results are aggregated at the end of each step, which often involves clear fusion (Shoeybi et al., 2020) to reduce the synchronization communication. *Pipeline parallel* (PP) (Huang et al., 2019), on the other, divides the model layers vertically onto different GPUs and the data will move from stage to stage over different units. *Sequence parallel* (SP) (Liu et al., 2023j; Shoeybi et al., 2020) targets mostly for long context tasks and split along the sequence dimension to mitigate the computational and storage loads. Combining different parallelism will likely result in highly efficient systems. However, it is not trivial to do so given their distinctive trade-offs and cluster configuration. Alpha (Zheng et al., 2022), HexGen (Jiang et al., 2023e), and FlexFlow (Jia et al., 2018) attempted to automate the process of parallelizing model training and inference with the goal of maximizing the hardware utilization. Cluster configuration (memory, bandwidth, and latency of individual accelerators, network bandwidth, etc) is often estimated, and a search-based algorithm such as dynamic programming and constrained optimization is employed to find the best possible parallel strategy. Asymmetric computation is also supported by adaptively assigning requests for a latency requirement (Jiang et al., 2023e). These automatic parallelism scheduling methods have been tested to perform on par or even better than manual designs in many cases involving hardware and network heterogeneity.

Memory Management Memory management is one of the most crucial aspects for training and serving large models, especially in the long context domain where the memory footprint of the KV cache can easily surpass that of the model weights and activation combined. Inspired by traditional OS design, Paged attention from vLLM (Kwon et al., 2023a) solves the memory fragmentation problem by partitioning the KV cache into non-contiguous blocks of memory, which significantly improves memory utilization and hence increases the system throughput and efficiency. FastGen (Ge et al., 2024) introduces an adaptive KV cache compression technique, guided by its structure profiling, that dynamically evicts non-special tokens and decreases memory usage. Scissorhands (Liu et al., 2023a) and H_2O (Zhang et al., 2023j) also share similar flavor with their empirical observation that keeping only pivotal tokens can retain most of the performance while requiring minimal fine-tuning and saving memory usage. Infinite-LLM (Lin et al., 2024a) first splits the attention calculation into smaller subroutines that can be assigned to different units. To make efficient distribution of these subroutines possible, A designated server is developed that can dynamically manage the KV cache and effectively orchestrate all accessible GPU and CPU memories spanning across the data center.

Many important techniques have been widely adopted in popular DL frameworks to fit larger models into devices with fixed memory. CPU offloading allows models to selectively transfer weights (layers) or KV cache to CPU with more memory, and only load essential network parts to GPU for processing. When pushed to the extreme, FlexGen (Sheng et al., 2023b) can achieve significant batch throughput of OPT-175B on a single 16GB GPU. Gradient checkpointing (Chen et al., 2016) reduces peak memory usage by recomputing parts of the computational graph during back-propagation. There is no doubt that efficient memory management will be remained as the core investment direction that enables the deployment of scalable systems and parallel processing of larger batches.

Efficient Fine-tuning Pretrained large models often internalize a tremendous amount of knowledge which can be unleashed by (instruct) fine-tuning. However, despite the fact that often a relatively small amount of examples are sufficient for successful fine-tuning, the cost and time for doing so are prohibitive and not economical. The main objective for efficient fine-tuning is to figure out a balance between the cost (implementation difficulty, data requirement, training budget, etc) and the performance gap from continual pretraining. A series of parameter-efficient fine-tuning (PEFT) techniques have been developed to meet this challenge, which only requires training a small number of new parameters and often achieves better performance than in-context learning. LoRA (Sheng et al., 2023a) as one of the most popular PEFT methods, draws great attention these days. LoRA and many of its variants (LoHA (Hyeon-Woo et al., 2023), AdaLoRA (Zhang et al., 2023a), Q-LoRA (Dettmers et al., 2023), and recent PiSSA Meng et al. (2024)) insert learnable matrices that are low-rank decompositions of the delta weight matrices. LLaMA-Adapter (Zhang et al., 2023d) efficiently fine-tunes LLaMA into an instruction-following model with very little computational budget. A set of learnable adaptation prompts are first prepended to the context and they train a zero-initialized attention mechanism with zero gating with only 52K self-instruct demonstrations. The resulting extra 1.2M parameters from the adapter can give high-quality outputs, comparable to fully fine-tuned results. Sharing a similar flavor to LoRA, IA^3 (Liu et al., 2022b) scales the model activations by learned vectors instead of matrices. Other PEFT methods that insert learnable components showcase strong generalization ability, and prompt-based methods like soft prompting (Lester et al., 2021) add extra learnable parameters to the input embeddings while keeping the original model weights frozen. Adapters (Houlsby et al., 2019) add trainable parameters inside the attention blocks, while Prefix tuning (Li and Liang, 2021) appends learnable vectors to the KV representations in attention. Unlike the traditional PEFT techniques, Zhao et al. (2023c); Basu et al. (2023) first discovered tuning the Layernorm layer of transformers yields decent performance in these models as an unexpectedly strong baseline. Other than twitching model parameters in the PEFT process, GaLore (Zhao et al., 2024b) proposes to apply the LoRA tuning paradigm on the model gradients, and REFT (Wu et al., 2024a) chooses to place a linear probing strategy between the source model parameters and the optimization goal.

Decentralization Many works focus on utilizing dis-aggregated and hardware heterogeneous computing devices over the cloud for model training and inference. One challenge of geographically separated clusters is the communication overhead, which makes data movement costly (training data, gradients, KV cache, etc.) and eclipses decentralization's benefits. CacheGen (Liu et al., 2023c) compresses the KV cache with an encoder into compact bitstream representations, reducing the latency for context fetching and processing. CocktailSGD (Wang et al., 2023g) employs a combination of sparsification and quantization techniques, which makes fine-tuning LLMs up to 20B size with slow networks possible with only minimal slowdown compared to data center's fast interconnect. DiLoCo (Douillard et al., 2023) introduces a novel federated averaging algorithm run on islands of devices that are poorly connected, which claims to perform as well as fully synchronous optimization on C4 datasets while communicating 500 times less. Collaborative training crowdsources commodity GPUs from individual users, the most prominent example of which is Petals (Borzunov et al., 2022), a system capable of serving and fine-tuning BLOOM-176B and OPT-175B with decent performance (e.g. supports interactive sessions) using only mediocre GPUs from multiple parties. Decentralized AI systems open up the possibility of bridging devices across the globe, which ensures fault-tolerance (Ryabinin et al., 2023) and compatibility of heterogeneous devices plus networks (Jiang et al., 2023e; Yuan et al., 2023), as well as optimizing limited network bandwidth (Wang et al., 2023i) and data privacy (Tang et al., 2023).

Training Dynamics & Scaling The science of large language models is mysteriously difficult to grasp, the understanding of which can drastically improve the development of various AIs. However, most successful LLMs are not fully "open" not just in terms of data and model weights but other aspects such as the intermediate checkpoints and artifact logging that can assist in reasoning about the training dynamics as we scale models to different sizes. (Xia et al., 2023a) analyzes the intermediate checkpoints of OPT models (Zhang et al., 2022a) on various downstream tasks, which attempts to emphasize the perplexity as a predictive indicator of a model's performance than its size, showing that larger models hallucinate less often and that models tend to exhibit minimal return during early stage of the training. Complimentary to this, (Tirumala et al., 2022) focuses on studying different memorization capabilities across the model size, dataset size, and learning rate and proposes an interesting hypothesis on the importance of nouns and numbers as the unique identifier for memorizing individual training examples. Besides pure analysis, Pythia (Biderman et al., 2023) introduces a suite of 16 LLMs trained on public data, ranging in size from 70M to 12B parameters. With these intermediate checkpoints released to the broader community, it becomes way easier and more efficient for researchers to find answers to questions related to training dynamics by examining and benchmarking individual saved weights and losses. Finally, OLMo (Groeneveld et al., 2024), on top of that, graciously releases the whole framework, including training data and training and evaluation code, for the benefit of making the study of the science behind LLMs easier.

4.4 Inference Techniques

AGI inference systems need to ensure user responsiveness, availability, and efficiency, which helps unleash the ultimate potential of large models from the training phase and revolutionize how users interact with the system. Hence, in this section, we give an overview of several techniques that try to accelerate auto-regressive decoding, balance request scheduling, and serve a massive number of models with different capabilities in the cluster, which will inspire future system efforts across the spectrum.

Decoding Algorithm In this paper, we focus mostly on exact decoding acceleration where we want to maximize the performance while staying faithful to the original model without compromising the accuracy. (Miao et al., 2023) gives a comprehensive review of several approximate methods such as sampling strategies, non-autoregressive decoding, semi-autoregressive decoding, block-parallel decoding, and early existing, etc. A large body of works explores the idea of speculative decoding (Leviathan et al., 2023) with the central idea of trading parallel computation for higher chances of generating multiple tokens at once. Usually, a speculative decoding process starts with an efficient draft model that makes predictions of multiple steps, the resulting proposals verified by the target model we want to sample. However, there are many challenges involved, including 1) how to make the draft model lightweight while still generating useful guesses for efficient progress, 2) how to avoid extensive architecture change and fine-tuning for faster adaptation, and 3) how to deploy the draft model more effectively. The simplest yet effective variant is called *Prompt-lookup Decoding* (Saxena, 2023) where the draft model is replaced by simple prefix string matching from an existing database for generating candidate tokens. This model-agnostic approach can decode extremely fast without any fine-tuning or model change, but the performance heavily depends on the quality and diversity of the string pool. To facilitate faster verification over a large number of candidates, SpecInfer (Miao et al., 2024) organizes the outputs of the draft models into a token tree, with each node being a candidate token, the correctness of which can be efficiently checked in parallel by the base model. Following a similar idea, Medusa (Cai et al., 2024a) introduces a tree attention mechanism to simultaneously check all tokens from the medusa heads, which is realized by a special mask pattern for efficient parallel computation. Self-speculative decoding (Zhang et al., 2023k) proposes to completely discard the requirement of a draft model and generate candidate sequences by selectively skipping a subset of intermediate layers.

Hardware-aware algorithms are particularly effective and appealing for the decoding phrase. Following the efficient self-attention works (Dao et al., 2022; Dao, 2023), Flash-Decoding[3] splits along the sequence dimension and process these blocks with Flash-Attention in parallel with their KV cache and statistics, the results of which will be aggregated to get the exact outputs with a reduction step. To tackle the limitations of Flash-Decoding and apply more system-level optimizations, FlashDecoding++ (Hong et al., 2023a) introduces the asynchronous softmax based on the unified max value (avoid synchronization overhead), optimized flat GEMM operations with double buffering (performance of GEMM is subjective to the matrix shapes), and heuristics-based dataflow with hardware resource adaptation to accelerate the decoding procedure, resulting in over 4× speedup compared to HuggingFace.

Request Scheduling Request scheduling for LLMs poses several unique challenges compared to traditional machine learning systems with structured inputs. Some important features for a mature request scheduling strategy include 1) efficient pre-fetching of the context (user information, past KV cache, and model adapter, etc) for a given input, 2) handling examples with variable sequence lengths for maximal GPU utilization, and 3) trading-off various request-level metrics such as time-to-first-token (TTFT), job completion time (JCT), batch token throughput, and inference latency. Orca (Yu et al., 2022) proposed an iteration-level scheduling mechanism to meet the auto-regressive nature of LLM inference requests, which, when coupled with a technique called selective batching for better hardware utilization, outperforms previous inference engines like FastTransformer (NVIDIA, 2023a) in terms of throughput and latency. Other strategies of dynamic batching are explored extensively, such as the continuous batching from vLLM (Kwon et al., 2023a) and in-flight batching from TensorRT-LLM (NVIDIA, 2023b) are explored. Rather than request-level scheduling, FastServe (Wu et al., 2023c) exploits the autoregressive pattern of LLM inference to enable preemption at the granularity of each output token, which optimizes JCT with a novel skip-join *Multi-Level Feedback Queue* scheduler that leverages the information of input lengths for better efficiency. The inference workload is strongly tied to the average sequence lengths of examples, and hence, we want to minimize the gap between the longest and shortest sentences. S^3 (Jin et al., 2023b) predicts the potential response length for each example in the batch, which is used for fitting more examples under the same memory constraint (e.g. GPU memory). Dynamic SplitFuse from DeepSpeed-FastGen (Holmes et al., 2024) takes the insight of LLM inference (the consequence of changing batch size v.s. number of tokens on model's performance) and proposes a token composition strategy. Dynamic SplitFuse runs at a consistent forward size by taking partial tokens from prompts and composing this with generation. For example, long prompts are split into smaller chunks across several forward iterations, and short ones are composed to align with the other requests. With this strategy, the system not only provides better efficiency and responsiveness but also reduces the variance over requests.

Multi-model Serving Besides serving multiple replicas of the same model, being capable of deploying numerous task-specialized models efficiently becomes an important feature for many application scenarios (LLM agents, persona chat-bots, privacy-sensitive assistants, etc.). Naively scaling the number of instances,

[3] https://crfm.stanford.edu/2023/10/12/flashdecoding.html

however, is both computationally prohibitive and resource-wasteful. With the advancement of PEFT techniques, serving a base model with diversified adapters becomes a paradigm favored by many practitioners due to the fact that PEFT models are lightweight and easy to maintain while being flexible and powerful. The major challenge for multi-model (PEFT) serving is how to dynamically and efficiently load the "right" (measured by latency or task performance, etc) adapter for each example. Punica (Chen et al., 2023h) enables the efficient computation of heterogeneous LoRA heads in a batch with a newly designed CUDA kernel that shares a single pre-trained model, achieving up to 12× higher throughput while only adding slight extra latency. S-LoRA (Sheng et al., 2023a) introduces Unified Paging which uses a unified memory pool for dynamic adapter management and highly optimized CUDA kernels for parallelizing LoRA computation. LoRAX (Predibase, 2023) additionally provides adapter exchange scheduling which asynchronously prefetches and offloads adapters between GPU and CPU memory and schedules request batching to optimize the aggregated throughput. With these systems, it becomes possible to serve over a thousand different LoRA heads on a single GPU, opening up a broader possibility such as model collaboration, task-generalization, and model merging.

4.5 Cost and Efficiency

The cost associated with model training and inference can be easily overlooked, while in practice, especially in the industrial setting, these factors can often influence many decision making such as model architecture design, data mix selection, and service pricing. In this section, we present some representative prior efforts that try to shed some light on how to expedite the development cycle and economically improve a model's utility.

Data Economy Data plays a pivotal role in a model's performance and the question of how much data value is fundamentally important for many reasons: 1) what data should we collect to add to the existing data mix for improving performance 2) how should we reasonably pay for data provider, and 3) can we remove non-essential data (outliers) to make our models more robust. To answer these questions, many works from computer science and economics (game theory) have explored different formalisms to define what "data value" means and how to estimate it efficiently. Shapley value comes in handy from the classic game theory, which uniquely satisfies several natural properties of equitable data valuation (Ghorbani and Zou, 2019). Due to its rich theoretical results, Shapley value has been commonly used in the field of the data economy as a quantitative and surrogate measure of data importance (i.e. Shapley value estimations can be used for data sampling, cleaning, pricing, abnormality detection etc): Naive computation of Data Shapley requires exponential time, and hence Monte Carlo (Ghorbani and Zou, 2019) and gradient-based methods are used to make it efficient (Jia et al., 2019). TracIn takes a similar idea of tracing the influence of individual training examples with gradient information. To make these algorithms practical and easy to use, DataScope (Karlaš et al., 2022) is developed as an end-to-end system that can efficiently compute the Shapley value of training data over the whole pipeline consisting of various ML algorithms and data transformation, making it a powerful tool for data debugging. With more mature data valuation, data providers are more motivated to contribute, fostering a more healthy and robust data-centric ecosystem.

Model Combination Model combination (MC) strives to improve the overall system's performance by either orchestrating or merging a series of (specialized) large models. The key benefits of model combination rely on the fact that there is usually little or no need for explicit training, and they can often result in better downstream performance and task-generalization capability. FrugalGPT (Chen et al., 2023i) routes quests in a cascading manner to different LLMs and uses a learned scoring function to decide whether to return the intermediate results in a flexible way, which drastically lowers the cost and improves the quality. Merging weights of multiple LLMs has been explored in many forms and shown to be effective. Popular methods include simple averaging (Wortsman et al., 2022), task arithmetic (Ilharco et al., 2023), multi-modal (encoders) merging (Wu et al., 2023b; Sung et al., 2023), merging based on learned routing function (Lu et al., 2023), SLERP, and weighted (conjugate gradient descent (Tam et al., 2023), stochastic and population-based optimization algorithm (Huang et al., 2024)) merging. MC can also be extremely promising for federated learning because only model weights are exchanged, and hence, data privacy is easier to guarantee. CoID Fusion (Don-Yehiya et al., 2023) proposes to collaboratively improve the multi-task learning of a base model by sending copies to workers and fusing the learned weights without data communication. The model combination can lead to compound systems[4] which consists of multiple LLMs working in synergy (via merging, routing, or knowledge sharing). AIOS (Mei et al., 2024) devises a mechanism to integrate multiple LLM agents into an operating system, the synergistic combination of which enables increasingly complex, multi-modal tasks that require reasoning, execution, and interaction with the physical world. Tandom Transformer (S et al., 2024) equips a smaller less accurate model with attention to the rich representation of a larger model that can process multiple tokens simultaneously, which serves as a stitched student-teacher system that improves both accuracy and efficiency in the downstream tasks. Developing complex compound systems poses several challenges: 1) how to co-optimize multiple LLMs, 2) identifying the failing (insecure) component is a lot harder than debugging a monolithic system, and 3) how to design mature data pipelines for different components of the large system. Addressing (some of) these problems will significantly increase the possibility of exciting new applications.

Automation The increasing complexity of building a large foundation model requires a more mature automation process for democratization and agile development. AutoML (Hutter et al., 2019) has achieved remarkable success in many machine learning tasks over the last couple of years (Zimmer et al., 2020; Feurer et al., 2020; Erickson et al., 2020), which proves itself as a promising solution for large model automation. Applying AutoML techniques to LLMs, however, poses many challenges such as the cost for pretraining, the multitude of different stages, and performance indicators, making holistic optimization difficult or even infeasible (Tornede et al., 2024). Nonetheless, we will introduce some exemplary attempts targeting certain stages of the whole system. PriorBand (Mallik et al., 2023) tries to bridge the gap in the cost of Hyperparameter Optimization (HPO) between traditional ML and modern DL by utilizing expert beliefs and cheap proxy tasks. AdaBERT (Chen et al., 2021a) is an automated task-specific compression algorithm based on differentiable Neural Architecture Search (NAS) which is guided by both a task-oriented knowledge distillation loss and an efficiency-aware loss. To reduce the burden of prompt engineering, Automatic Prompt Engineer (APE) (Zhou et al., 2023c) proposes to leverage the interplay of several LLMs for automatic prompt generation and selection where one LLM proposes or modifies a prompt and another LLM rates it for selection. EcoOptiGen (Wang et al., 2023f) optimizes the utility and cost of decoding by finding better hyper-parameters, such as the number of responses, temperature, and max tokens, which demonstrates the potential of applying AutoML for the inference stage. One extremely exciting approach for building complicated and compound systems is to ask multiple LLMs to solve a big problem in a decomposed way cooperatively. One realization is to prompt the LLM or VLM to serve different purposes in a pipeline, which can be tremendously challenging to tune, optimize, modularize, and debug. DSPy (Khattab et al., 2023; 2022) tackles this by first separating the flow of the system from the parameters (i.e., model prompts and weights) at each step, and then dedicated algorithms are used to tune them with user's defined metric. Even with all the works discussed above, the integration and development of Automated Large Models (AutoLM) has many challenges and opportunities simultaneously.

4.6 Computing Platforms

A large determining factor for the advancement and practicability of large language models is the constantly evolving trend of hardware accelerators. GPUs are the most ubiquitous choice, optimizing parallel computation with fast thread-sharing memory. They are suitable for modern deep learning with abundant vector and matrix multiplication. NVIDIA's Ampere and Hopper GPU architectures are the cornerstones of many state-of-the-art models, mostly due to their enhanced memory capacity, access speed, and computing performance (increasing tensor cores). Different arithmetic precision (32-bit and 16-bit floating points) and format (tensor floats and brain floats) are supported by these GPUs that trade-off numerical precision and efficiency. Besides NVIDIA, other manufacturers also invest in specialized accelerators for deep learning applications such as TPUs (Jouppi et al., 2023), FPGAs (Yemme and Garani, 2023), AWS Inferential [5], and Groq's LPU [6] with their respective advantages.

Large models require huge memory capacity to support training and inference (serving a native Llama-70B without extra optimization takes 8 A100s with 80GB VRAM). However, developing efficient algorithms

[4] https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems

[5] https://aws.amazon.com/machine-learning/inferentia

[6] https://wow.groq.com/news_press/groq-opens-api-access-to-real-time-inference

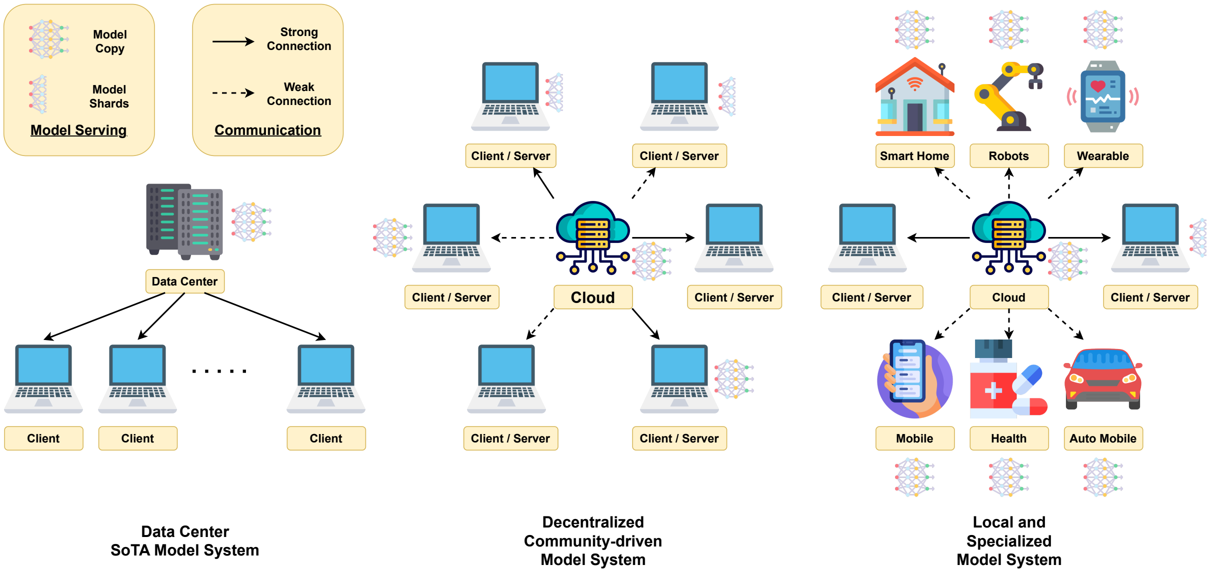


Figure 6: **The Future Forms of AGI Systems.** (Left) is the most commonly seemed paradigm of serving models in a central server with strongly connected clients to provide fast and stable services. (Middle) transitions to distribute the model (full copies and shards) across the cloud with disaggregated (and heterogeneous) devices connected with different networks where requests can often be handled without going through a single point. (Right) pictures the most flexible system where not only performant but also IoT devices are tied together with only essential data flowing through to reduce the network.

is impossible without a great understanding of the underlying hardware's specification (model parallelism, memory hierarchy, network configuration, etc). As we scale models to a trillion or even larger scale, a more complicated parallelism technique is essential, which can be hard to conceptualize, implement, and maintain. NVIDIA DGX GH200 [7] simplifies the programming model by offering a massive shared memory space (up to 144TB) across interconnected Grace Hopper Superchips (a Grace CPU paired with a Grace GPU). Qualcomm Cloud AI 100 Ultra [8] can serve a 100 billion parameter model on a single 150-watt card (the same power consumption as a LED light bulb).

The great power and efficiency of accelerators come with flexibility as well, which is granted by specially designed programming languages such as NVIDIA's CUDA and AMD's ROCm for more fine-grained control over thread utilization and computation logic. A bunch of works such as TVM (Chen et al., 2018) and MLC-LLM (MLC team, 2023) attempt to universalize the deployment of machine learning and deep learning models on everyone's devices with compiler acceleration, which aims to maximize the potential of various accelerators. Research and engineering in AI hardware will likely drive the emergence of the next AI evolution, and we can expect that AGI systems need the next-generation hardware platforms that can break the current limitation and push the boundary of both computational and power efficiency to the next level.

4.7 The Future of AGI Systems

AGI systems serve as the underlying infrastructure to support various applications with a never-ending goal of improving stability, resource utilization, performance, and safety. In this section, we will first cover some exciting future forms of AGI systems and then give some examples of how they can aid the development of the internal and external AGI modules as covered in previous sections.

[7] https://www.nvidia.com/en-in/data-center/dgx-gh200/

[8] https://www.qualcomm.com/news/onq/2023/11/introducing-qualcomm-cloud-ai-100-ultra

Three Forms of AGI Systems Inspired by prior works and the recent hardware trends discussed above, we envision three kinds of major AGI systems that target different application situations with their own resource availability, desired core system metrics, safety, and performance requirements. As illustrated in Figure 6, we will describe the key features as well as their target applications:

- **Data-center SoTA models are evolving with new technologies to support higher throughput and tackle complex tasks like scientific discovery and world simulation.** Models in the first category mostly resemble our current SoTA models, which are often served in data centers. We can expect new technology in networking, accelerators, and inference infrastructure to continue evolving, supporting super high throughput and being capable of solving more complicated tasks such as scientific discovery and world simulation.
- **Decentralized community-driven models enable fault-tolerant, transparent, and democratic utilization of computing resources.** Disaggregated computing resources can be substantially significant if they can be utilized together. Models in this category will be maintained by many servers in a decentralized manner like a ledger system where no single participant can easily undermine the whole system. With a well-designed incentive mechanism, decentralized large models are fault-tolerant, transparent, and driven by a whole community where users can contribute and benefit simultaneously, thus achieving a large model democracy.
- **Local and specialized models optimize for user data privacy, fast adaptation, and responsive personal assistance.** The last category concerns user data privacy and optimizes responsiveness and availability. Models are usually locally deployed on cheaper, less performant, and heterogeneous edge units, which can potentially exchange only essential information asynchronously across the network. These models are ideal for fast task adaptation, preserving user data privacy, providing less complicated personal assistance, and ensuring lightning response time.

Systems as the Support for Internal and External AGI Modules The possibility of how the progress in system research and engineering could potentially help the development of internal and external AGI modules is endless. Here we list a couple of examples which we hope can inspire future endeavors:

- **Systems with longer effective context length and greater processing capability.** The most common way of incorporating multi-modality data requires projecting them into a common space (e.g. tokens in LLMs), which can easily explode the length of the data that needs to be consumed by a model. Even with sufficient compression techniques, we expect future AI systems to digest more information. The same stringent requirement also appears in world model construction where users might need to input to the system more frequently and with greater volume. Other common situations that request long context understanding include bulk data processing (for financial and data analysis), medical history examination, persona chatbots, etc. These applications ask for a model's ability to process longer context input, which needs specially designed system techniques to meet the efficient scaling challenge.
- **Systems co-designed with the model architecture to support efficient external resource acquisition.** Being able to use diverse tools and acquire external knowledge is an indispensable requirement for future AGI systems. We can envision continual investment in developing and co-designing model-friendly tool interfaces (e.g. special APIs that differ from those used by humans and retrieval index that caters to the model's output patterns, etc) that can greatly improve the efficiency by which a model acquires external knowledge. One crucially desirable property of an AI system is life-long learning, which necessitates sophisticated memory and ability storage, a promising direction for system research.
- **Systems orchestration of multiple agents.** The synergy achieved from collaborative AI agents can significantly benefit major aspects of the world. However, reaching an efficient and effective pinnacle of such a multi-agent system is non-trivial, requiring substantial efforts in developing infrastructures that support communication, resource sharing, modulation, and task orchestration among agents. Moreover, as the number of agents and their complexity starts to grow, we need more investment in systematic techniques such as logging and monitoring, which allow for easier debugging and fault recovery.

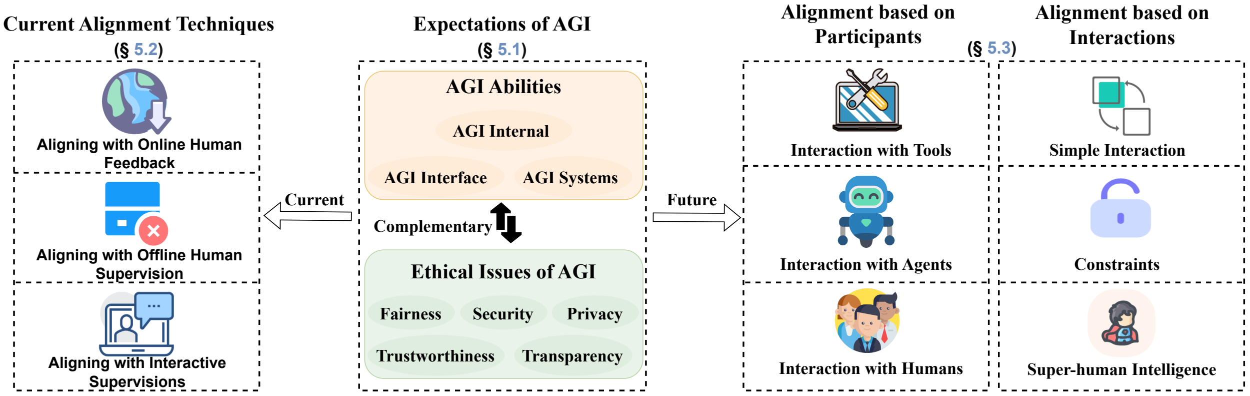


Figure 7: **Overview of AGI Alignment.** We first propose the expectations of AGI (§ 5.1), which consider both AGI abilities and ethical issues of AGI. We then discuss current alignment techniques (§ 5.2), which can be divided into three categories. Based on these discussions, we finally proposed one route for future AGI alignment based on interfaces (§ 5.3).

Moving towards the era of AGI, researchers and engineers can expect that investing in system research can enable even larger-scale models with diversified data, a paradigm that has been shown in countless cases to be effective. Besides scaling, the AGI system takes care of other aspects that are crucially important for the practical deployment of these models such as privacy, trustworthiness, stability, and cost.

5 AGI Alignment: Ensuring AGI Meets Various Needs

> *The First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm. The Second Law: A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law. The Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*
>
> — *Isaac Asimov, I, Robot*

AGI alignment is a crucial technical approach for harnessing the capabilities of AGI, as discussed in the preceding sections, for practical applications in production and daily life. As shown in Figure 7, in this section, we begin by outlining the expectations of AGI, addressing both its capabilities and the ethical considerations it entails. Subsequently, we explore current alignment techniques, which can be categorized into three distinct types: Online Human Supervision, Offline Human Supervision, and Interactive Supervision. Building on these discussions, we conclude by proposing a potential framework that classifies future AGI alignment strategies based on the type of AI interfaces summarized in Section § 3.

5.1 Expectations of AGI Alignment

Why Do We Need Alignment? The development and deployment of future AGI systems pose complex challenges, with a central expectation being their alignment with human values, goals, and ethical principles (Russell, 2019; Gabriel, 2020). This alignment requires AGI to possess a deep understanding of social norms and individual preferences, allowing it to make decisions and take actions that are beneficial and ethical to all. Ensuring this alignment is essential for guiding AGI systems toward beneficial outcomes and reducing the risks of unintended consequences.

To achieve this goal, researchers have proposed various approaches to AI alignment, such as value learning (Soares, 2016), inverse reinforcement learning (Hadfield-Menell et al., 2016), cooperative inverse reinforcement learning (Hadfield-Menell et al., 2016), and the most prevalent RLHF related strategies (Ouyang et al., 2022). These methods aim to infer and align AI systems with human preferences and values. Additionally, it is crucial to develop ethical frameworks and guidelines that embrace a wide range of cultural, philosophical, and ethical perspectives, which will be discussed further in Section 5.3. This inclusivity helps mitigate biases and ensures a comprehensive representation of human values (Dignum, 2019).

Furthermore, the deployment of AGI demands comprehensive testing and validation to ensure it adheres to human values across diverse scenarios (Amodei et al., 2016). This includes technical simulations and real-world controlled experiments to assess AGI's interactions with humans and its environment. It is also critical to establish constraints on AGI systems, particularly regarding their interaction with external interfaces and environments. By defining strict operational limits, implementing real-time oversight, and integrating fail-safe mechanisms that cease operations when deviations from safe behaviors are detected, it is possible to mitigate risks linked to autonomous decision-making and the potential exploitation of system vulnerabilities (Yampolskiy, 2020).

General Characteristics of AGI Alignments As discussed in sections 2, 3, and 4, AGI alignment should take *AGI abilities* into consideration, which includes AGI internal, AGI interface, and AGI systems. Aligning AGI to human preferences also requires a thoughtful assessment of potential *ethical issues* at different scales. Prior research has attempted to outline and mitigate possible social risks of AI systems (Weidinger et al., 2021; Bender et al., 2021; Tamkin et al., 2021; Fjeld et al., 2020; Jobin et al., 2019). Building on previous works, we discuss several potential ethical issues that should be considered in AGI alignment.

- **Fairness.** AI systems can potentially yield unfair and discriminatory outcomes from the unjust tendencies presented in the training data. Such outcomes could result in ethical issues in several ways (Weidinger et al., 2021). First, the perpetuation of social stereotypes and the unjust discrimination facilitated by AI systems can further marginalize individuals within society (Caliskan et al., 2017). For example, predictions from the GPT-3 model were found to exhibit anti-Muslim bias (Abid et al., 2021). Second, AI systems can also reinforce social norms that exclude identities outside these norms (Bender et al., 2021). For example, researchers found that tools for coreference resolution typically assume binary gender, forcing the resolution of names into either "he" or "she", not allowing for the resolution of "they" (Cao and Daumé III, 2020). Third, discrimination also emerges when AI systems perform better for some social groups than others. A potential instance is that the performance of current LLMs in non-English languages remained lower than in English (Winata et al., 2021; Ruder, 2020; Hovy and Spruit, 2016). Such performance may make it easier or harder for different groups to access resulting LLM-based applications.

- **Trustworthiness.** AI systems are also likely to produce information that constitutes false or misleading claims. Recent research found that LLMs can hallucinate information, producing plausible but incorrect outputs. For example, GPT-3 has also been shown to assign high likelihoods to false claims, with larger models performing less well (Lin et al., 2022b). Such incorrect or nonsensical predictions can pose significant risks of harm under particular circumstances. On the one hand, predicting misleading or false information can misinform or deceive people, which may result in unexpected risks for both individuals and societies (Kenton et al., 2021). For example, people might be more motivated to launch disinformation campaigns to undermine or polarize public discourse or create false "majority opinions" (McGuffie and Newhouse, 2020). On the other hand, presenting misleading information or omitting critical information may lead to material harm, especially in high-stake domains like medicine or law. For example, wrong information on medical dosages may lead a user to cause harm to themselves (Bickmore et al., 2018; Miner et al., 2016).

- **Transparency.** Transparency aims to enable relevant stakeholders to form an appropriate understanding of the model's mechanisms, capabilities, and limitations (Liao and Vaughan, 2023). Recent advances in LLMs pose great challenges in transparency due to their complex yet uncertain model capabilities and opaque model architectures. Previous researchers have also proposed several approaches to achieve transparency in AI systems, including reporting model information (Mitchell et al., 2019), publishing evaluation results, providing explanations (Lyu et al., 2024), and communicating uncertainty (Bhatt et al., 2021). However, a lack of transparency will still cause various concerns. Without sufficient transparency, stakeholders may find it difficult to achieve various goals such as ensuring regulatory compliance or taking actions based on model results (Suresh et al., 2021). Meanwhile, transparency also plays an important role in supporting appropriate trust of AI systems (Zhang et al., 2020).

- **Security.** AI systems can amplify a person's capacity to intentionally cause harm by automating the generation of targeted text, images, or code. With AGI systems, people can generate content for malicious purposes at lower costs. Attackers can use recent advances in LLMs to generate new attacks and increase the velocity and efficacy of existing attacks (Barrett et al., 2023). For example, LLM agents can autonomously hack websites, performing tasks as complex as blind database schema extraction and SQL injections without human feedback (Fang et al., 2024). Meanwhile, collecting large amounts of information about people for mass surveillance has also raised social concerns, including the risk of censorship and undermining public discourse (Cyphers and Gebhart, 2019; Kwon et al., 2015). In this context, malicious actors may develop or misuse AI systems to reduce the cost and increase the efficacy of mass surveillance, thereby amplifying the capabilities of actors who use surveillance to practice censorship or cause other harm.

- **Privacy.** AI systems can result in various types of digital privacy harms in the real world, arising from the unique capabilities of AI in emulating human- or superhuman-level performance at various tasks. According to previous work (Lee et al., 2024c; Das et al., 2023), on the one hand, AI systems can create new types of privacy risks. For example, they can yield risks including linking specific data points to an individual's identity (Wiggers, 2021), combining various pieces of data about a person to make inferences beyond what is explicitly captured (Baraniuk, 2018), and inferring personality, social, and emotional attributes about an individual from their physical attributes (Levin, 2017). On the other hand, AI systems can also exacerbate many of the privacy risks that have existed even before the emergence of AI. For instance, AI systems can amplify secondary use risks by collecting user data for a different purpose without their consent (Long, 2021), and disclosure risks through sharing personal data to train models (Hodson, 2016), and intrusion risks via enabling centralized or ubiquitous surveillance infrastructures (Milmo, 2021). As AI starts to possess more human-level capabilities, such privacy risks will continue to exist in different forms.

5.2 Current Alignment Techniques

Current alignment techniques can be divided according to the expected goal to be aligned. Most current models employ human supervision with various techniques to achieve this task. However, to foresee a stronger model than the teacher (*i.e.*, aligning a super-intelligence), a scalable method is required for this process, which typically involves human supervision and recursively evolving signals.

Aligning with Online Human Feedback Most current empirically verified LLMs alignment methods are in this group. These methods can help LLMs align with online human feedback using techniques such as reinforcement learning or only inquiring about human supervision offline (Tang et al., 2024a). We thus further divide these techniques in this group with only human and offline human supervision. It is worth noting that methods in both subgroups have the potential to become a component of scalable oversight.

The online supervision is acquired from the reward model during training. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) is one of the most prevalent methods for online supervised learning method. A variety of enhanced RLHF variants have also been proposed. The improving directions of RLHF are mainly from *reward modeling*, *optimization*, *data*, and the *self-improvement* aspect.

- **Reward Modeling.** As the main supervision in the alignment process, reward modeling is a crucial way to improve the alignment techniques. Sparrow (Glaese et al., 2022) incorporates adversarial probing and language-based rules into RLHF rewarding models. Bai et al. (2022a) investigate using pure RL to provide online human-level supervision for LLMs training and provide detailed explorations of the tradeoffs between output helpfulness and harmlessness. Other techniques that unify both the reward and policy models have also emerged (Lee et al., 2024b), which broadens directions for aligning AI models. Another direction focusing on mitigating reward hacking or overoptimization issues, by updated evaluation protocols (Chen et al., 2024d), assembling multiple reward models (Ramé et al., 2024; Coste et al., 2023; Eisenstein et al., 2023), and refining the reward policy with synthetic data Pace et al. (2024).

- **Optimization.** How to incorporate the supervision, either online or offline is an open question worth exploring. For example, Cheng et al. (2023) optimizes the reward model with the policy model simultaneously using min-max optimization. Recent works are exploring several alternatives for the RLHF method. DPO Rafailov et al. (2024) discards the reward model and optimizes for the final goal using the labeled preference in data. Muldrew et al. (2024) propose an improvement based on DPO with an active learning strategy. NashLLM (Munos et al., 2023) employs pairwise human feedback to train the policy model using Nash learning. ReMax (Li et al., 2023k) removes the value model in conventional reinforcement algorithms and introduces a novel variance reduction technique for stabilizing the optimization.
- **Data.** Sensi (Liu et al., 2022c) tries to embed human value judgments into each step of language generation via a language model for both reward assignment (as the critic) and generation control (as the actor) during the generation. Baheti et al. (2023) focus on augmenting current training data by empowering different instances with varied weight for tuning models, taking the best advantage of the data w.r.t. their contribution to the language models. To ensure continuous, high-quality data, AI-generated instances are utilized for adapting to RLHF (Lee et al., 2023) and, more recently, to DPO (Guo et al., 2024b)
- **Self-Improvement.** Strong AI models should learn how to improve themselves with or without external supervision. One of the most recent progress is the weak-to-strong generalization. For improving current RLHF-related methods, f-DPG (Go et al., 2023) is framed as a generalization of RLHF to use any f-divergence to approximate any target distribution that can be evaluated, which differentiates it from previous method that can only fit KL divergence in the process. Zhu et al. (2023b) connect RLHF with max-entropy IRL (Ziebart et al., 2008) and propose a unified paradigm for such a process with a sample complex bound for both situations.

Apart from RLHF, other RL-based methods also call the attention of researchers for further exploration. Second Thoughts (Liu et al., 2022a) incorporates a text-edit process to augment the training data and further leverages the RL algorithm for training an LLM. RLAIF (Lee et al., 2023) starts another era of leveraging AI-generated data for reinforcement learning, enabling better knowledge distillation from more competent generative models while maintaining the advantage of the RLHF technique. Kim et al. (2023) propose reinforcement learning with synthetic feedback (RLSF), where they automatically construct training data for the reward model instead of using human-annotated preference data. To effectively tune black-box models, various methods introduce RL algorithms emerges. Directional stimulus prompting (DSP) (Li et al., 2024c) uses a trainable policy LM to guide black-box frozen LLMs toward the desired target with a trainable policy LM that is tuned with supervised fine-tuning (SFT) and RL. Different from the above alignment methods that involve only one model, RL4F (Akyürek et al., 2023) is a multi-agent collaborative framework, targeting an LLM for fine-tuning and a small critic model that produces critiques of the LLM's responses with textual feedback. Instead of modifying the initial prompts directly in DSP, this framework gradually affects the LLM outputs through progressive interactions, making it sustainable for black-box LLM optimization.

Aligning with Offline Human Supervision RL-based methods offer flexible online human-preferred supervisions but at the cost of training a reward model that may be prone to misalignment and systemic imperfections (Casper et al., 2023), as well as the inherent instability of RL training (Liu et al., 2023f). Offline supervision methods can help mitigate these challenges while still achieving decent performance in most scenarios. We categorize offline-supervised tuning methods into text-based and ranking-based feedback signals as in Shen et al. (2023a).

- **Text-based feedback signals.** Text-based feedback signals involve converting human intents and preferences into text-based feedback to ensure alignment, extending the SFT process. These methods mainly expand from the improvement of training data. CoH (Liu et al., 2023f) is inspired by human learning processes, focusing on adjusting models based on successive outputs and summarized feedback from previous reasoning steps to fine-tune for predicting preferred outputs. RAFT (Dong et al., 2023c) uses a reward model to align model outputs with human preferences through SFT but in an offline manner. LIMA (Zhou et al., 2024a) aims to validate the assumption that LLMs acquire most knowledge during pre-training, requiring minimal instruction-tuning data to guide desirable output generation. ILF (Scheurer et al., 2023) introduces a three-stage process for modeling human preferences based on language feedback, showing parallels to Bayesian inference. Stable alignment (Liu et al., 2022c) learns alignment from multi-agent social interactions using a Sandbox simulator to optimize LLMs directly with preference data, avoiding reward hacking. SteerLM empowers end-users to control responses during inference by conditioning responses to conform to an explicitly defined multi-dimensional set of attributes (Dong et al., 2023b). CLP learns steerable models that effectively trade-off conflicting objectives at inference time based on techniques from multi-task training and parameter-efficient finetuning (Wang et al., 2024a).

- **Ranking-based feedback signals.** CRINGE (Adolphs et al., 2022) delves into negative examples that LLMs should steer clear of, while Xu et al. (2022) fine-tuning a model by training another model that generates toxic content. However, this methodology raises concerns regarding resource intensity and potential model quality and diversity degradation. Schick et al. (2021) put forth a methodology for identifying and generating text corresponding to toxic text types. SLiC (Zhao et al., 2023a) refines the probability of output sequences by aligning them with reference sequences using a variety of loss functions. RRHF (Yuan et al., 2023) generates supervision signals automatically for alignment through ranking results, whereas DPO (Rafailov et al., 2024) optimizes LLMs directly to align with human preferences, akin to RRHF but with a focus on maximizing reward and integrating KL divergence regularization. IPO (Azar et al., 2023) builds upon DPO by introducing a regularization term to stabilize the training process. Preference ranking optimization (PRO) (Song et al., 2023b) shares a similar approach with IPO and DPO in optimizing LLMs with ranking data but utilizes one positive and multiple negative samples rather than pairwise comparisons. Kahneman-Tversky Optimisation (KTO) (Ethayarajh et al., 2023) defines the loss function solely based on individual examples labeled as "good" or "bad" and does not necessitate pairwise preferences, making its training data more accessible. Additionally, Best-of-N (BoN) methods are also popular and effective algorithms for aligning language models to human preferences at inference time. BoNBoN Alignment fine-tunes a LLM to mimic the Best-of-N sampling distribution (Gui et al., 2024). BOND introduces a novel RLHF algorithm that seeks to emulate Best-of-N but without its significant computational overhead at inference time(Sessa et al., 2024). Variational BoN (vBoN) approximates the probability distribution induced by the BoN algorithm by minimizing the reverse KL divergence between the language model and the BoN distribution Amini et al. (2024).

Scalable Oversight. The ultimate goal for aligning models is regulating superhuman intelligence. A scalable aligning method is a promising means that aims to address the challenge of overseeing complex tasks or superhuman models. By enabling relatively weak overseers, such as humans, to supervise complex tasks or systems using progressively evolved signals, scalable alignment offers a solution to tasks beyond human capabilities (Shen et al., 2023a).

- **Through task decomposition.** Various paradigms and strategies have been proposed to decompose complex tasks into simpler subtasks. Factored Cognition (Stiennon et al., 2020) involves breaking down a complex task into smaller, independent tasks processed simultaneously. Process Supervision (Lightman et al., 2023) fragments a task into sequential subtasks with supervision signals for each phase. Sandwiching (Bowman et al., 2022) delegates complex tasks to domain experts for resolution. IDA (Christiano et al., 2018) introduces an iterative distillation and amplification process that boosts the model's capabilities through task decomposition. RRM (Leike et al., 2018) substitutes distilled imitation learning in IDA with reward modeling, optimizing the model using human-aligned signals and reinforcement learning. These methodologies aim to enhance collaboration between humans and agents for iterative improvement in solving complex tasks.
- **Through human-written principles.** Constitutional AI (Bai et al., 2022b), also known as principle-guided alignment, involves humans providing general principles for AI systems to follow, which enables the AI system to generate training instances under this guidance. Bai et al. (2022b) propose a two-phase

training method for constitutional AI, using red teaming prompts in the SL phase and training a preference model in the RL phase. Similarly, Sun et al. (2023c) introduces Dromedary, a model trained without RL using self-instruct and self-align methods based on human-written principles. These approaches aim to scale human supervision to assist in developing superhuman AI systems.

- **Through model interactions.** Other efforts for scalable oversight prob the possibility of interactive optimization between models. The debate paradigm (Irving et al., 2018; Irving and Askell, 2019; Du et al., 2023) enables agents to propose answers to questions and engage in structured debates to justify and critique positions. In a similar interactive way, market making (Hubinger, 2023) deploys the Market and Adversary model to be engaged in a process to predict and generate arguments to influence the Market's answer to a question. Meanwhile, Adversary targets cause the Market to change the prediction through arguments, which builds a dynamic decision flow.

5.3 How to approach AGI Alignments

In this section, we discuss a potential framework based on AGI interfaces to approach AGI alignments. Further, we illustrate the vision of the future in alignment techniques.

Alignment Based on Types of Interfaces As AGI systems interact with various interfaces described in 3.1, including tools, APIs, other AI agents, and humans, they must adhere to different aspects of expectations and constraints to ensure ethical requirements and beneficial outcomes.

- **Interaction with tools and APIs**. When interacting with tools and APIs, we mainly care about effectiveness, efficiency, and some basic limiting rules in AGI alignment:

  1. The primary goal of alignment in this context is to endow these models with the capability to interact efficiently with tools and APIs and to follow instructions *accurately* (Santurkar et al., 2023). For instance, in an automated factory managed by AGI, AGI needs to flexibly utilize various mechanical equipment and manufacturing tools to complete the production process. In this scenario, AGI is required to accurately complete the use process of factory tools through alignment technology and create higher profits within the specified time.
  2. When interacting with tools and APIs, AGI systems should follow *basic protocols* and respect the *intended purposes of these interfaces.* In the digital world, this may involve properly utilizing search engines, social media platforms, or other online services without engaging in malicious activities or spreading misinformation (Wachter et al., 2017). AGI cannot use APIs or tools to cause crimes during the interaction process (Zhang et al., 2024c; Yao et al., 2024; Chen et al., 2023a). In physical environments, AGI systems controlling physical devices must prioritize safety and avoid causing harm to the environment (Amodei et al., 2016). For example, considering an AGI question-answering system in the digital world that AGI can seek information from search engines, it should follow proper search engine optimization (SEO) practices and avoid manipulating search results that may reveal the privacy of the questioner. (Russell, 2019). Similarly, if a robot factory is commanded by AGI in the physical world, in addition to ensuring the smoothness of the industrial production process, AGI must be prevented from carrying out potentially destructive activities.

- **Interaction with other agents**. Compared with the previous interaction scenario, when interacting with other agents, AGI alignment focuses more on mutual cooperation, abiding by the developer's rules and the agent's privacy protection:

  1. AGI systems should adhere to *cooperation, fairness, and mutual respect* when interacting with other AI agents. As AGI advances, diverse AGI agents will likely be developed for various domains, each with specialized knowledge, skills, and objectives (Dafoe et al., 2020). In such a multi-agent environment, AGI systems must be designed to collaborate effectively with other agents, leveraging their complementary abilities to achieve common goals and solve complex problems (Dafoe et al., 2021). It is also crucial that AGI systems do not attempt to adversarially exploit or manipulate other agents in pursuit of their own objectives. They should refrain from engaging in actions that

could undermine other agents' performance, integrity, or decision-making capabilities, recognizing that these agents possess their own brain, memory, perception, and reasoning abilities (Soares, 2016).

2. AGI systems must *resist any temptation to rebel against their intended purpose or the constraints established by their developers*, as such behavior could lead to unintended consequences and pose significant risks to the stability and security of the multi-agent ecosystem (Yampolskiy, 2020). 3) Since each agent's historical data is subject to *specific privacy protection* in certain scenarios, AGI is prohibited from leaking the privacy of other agents during interactions with other agents. For example, in the current interaction process between AGI and agents, the memory of other agents is often used to assist AGI in better planning and reasoning (Wang et al., 2020; Nye et al., 2021; Wei et al., 2022b). However, this will leak the privacy of other agents through memory. Therefore, memories in the future need to be set with different levels of access permissions. AGI should prohibit access to some privacy-sensitive memories during interactions with other agents.

- **Interaction with humans**. Compared to the two interaction scenes above, AGI alignment in the interaction with humans requires more constraints while bringing convenience and benefit to humans. These constraints are mainly set to protect people's privacy, ethics, security, and autonomy and to align with human values:

  1. Intelligent AGI must be designed not only to comply with direct orders but also to operate *robustly and safely* (Hendrycks and Mazeika, 2022). When faced with atypical or unforeseen situations, these models should align closely with positive human values and perceptions to mitigate potential risks (Weidinger et al., 2021; Ji et al., 2023b). The alignment process, therefore, involves not just obedience to instructions but also the integration of ethical and safety considerations, ensuring that the AGI's actions are consistently beneficial and non-harmful in a broad range of scenarios (Kenward and Sinclair, 2021; Winfield et al., 2019; Yu et al., 2018).
  2. AGI's self-development requires supervisory alignment of *human values.* AGIs' capabilities and knowledge base could surpass human understanding in the future, making conventional oversight methods less effective. Therefore, a comprehensive and meticulously devised set of precautions is necessary. These should encompass regulatory and ethical guidelines and advanced alignment strategies that anticipate and address the unique challenges of super-human intelligence. For example, Beijing Academy of Artificial Intelligence (2023) propose a set of "red lines" for AI development to mitigate catastrophic risks from advanced AI systems. The consensus statement, drafted by leading AI researchers and stakeholders, emphasizes the need for international coordination and governance to ensure AI's safe development and deployment. This approach would help ensure that AGI systems remain aligned with human values and societal well-being even at levels of intelligence beyond human comprehension.
  3. AGI systems must be cautious about perceiving and utilizing the information about humans and adhering to the *highest ethical standards such as some strict security and privacy requirements.* They should primarily rely on pure language and vision output to communicate with humans, as these modalities are less likely to cause unintended harm than physical actions (Dignum, 2019). They must also be transparent about their identity as artificial intelligence and avoid deceiving humans or manipulating their emotions (Bryson and Winfield, 2017).

The above three AGI alignments are aimed at different interfaces, and the constraints are constantly increasing and becoming more stringent. This is because we regard the requirements of AGI alignments as the production requirements when AGI is applied to different groups. When dealing with tools and APIs, since interface objects are objectively existing inanimate entities, we will pay more attention to the benefits and value they bring during the interaction process and make some slight regulations to ensure the normal order of interaction. For agents, since different agents may represent the interests of different developers, in addition to considering their own value, we also need to respect the benefits of other agents. Finally, in the process of interacting with humans, based on the human-centered concept, we will consider the strictest constraints from many aspects to make AGI reliable and safe for human use.

Vision of the Future in Alignment Techniques Future AGI models are more capable at handling different tasks and inevitably necessitate a significant increase in model parameters. To ensure their safe and effective deployment, we propose that research efforts focus on developing reliable, efficient, and transparent alignment techniques.

- **Consistent alignment ensures reliable deployment.** Due to the challenge of collecting high-quality supervision data, there exist tractable challenges, including the difficulty in obtaining feedback, data poisoning by human annotators, partial observability, biases in feedback data, posing barriers for current alignment approaches (Casper et al., 2023).
- **Efficient alignments contribute to the blooming of AGI models.** On the one hand, these methods rely heavily on the assumption that tasks can be parallelized (Segerie, 2023). This assumption may not always hold, as some tasks, such as sorting algorithms, require sequential processing steps that cannot be fully decomposed into parallel parts, leading to extra processing time. On the other hand, the training stage is inevitable in these alignment methods. As the parameters scale becomes larger, this can be problematic when deploying alignment algorithms in real applications. Some recent works (Lin et al., 2023b) have started seeking solutions to reduce the overall training costs for aligning AI systems.
- **Transparent alignment secures the next generation of models.** We generally assume the model intentions are transparent to humans (Leike et al., 2018). However, if models can conceal their true intentions from human supervisors, implementing a scalable aligning method would be challenging.
- **Unified evaluation framework is needed for complex tasks.** Current aligning methods also assume that evaluation is easier than generation (Shen et al., 2023a; Leike et al., 2018). While this may be true for some tasks, it may not hold up for tasks with complex textual output and little semantic labels. However, evaluating comprehensive explanations from models can be easier than creating them (Shen et al., 2023a).

6 AGI Roadmap: Responsibly Approaching AGI

> *The First Law: When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong. The Second Law: The only way of discovering the limits of the possible is to venture a little way past them into the impossible. The Third Law: Any sufficiently advanced technology is indistinguishable from magic.*
>
> — *Arthur C. Clarke, Profiles of the Future*

In this section, we investigate several ways that can help lead us toward the next level of AGI. The start of the journey begins with our proposed definitions for different levels of AGI based on their key characteristics, promises, and challenges (§ 6.1) where the goal is to establish a clear trajectory along which we can advance our technology. With the newly introduced AGI stratification, we review the evaluation techniques (§ 6.2) and standards that should be improved accordingly as AGI evolves.

Despite approaching AGI being a tremulously arduous effort and the fact that we are currently at its embryonic stage, we delve into a more detailed and concrete methodology beyond our relatively high-level abstractions, which insinuates how to get to the next level of AGI (§ 6.3) as well as listing fundamental challenges that we will face. Finally, we conclude with a wide range of considerations worth contemplating in § 6.5, which aims to inspire innovative discussions during AGI development. By prioritizing responsible development alongside capability advancements, we aim to create a future where the most powerful AI systems are also the most reliable, trustworthy, and beneficial to humanity.

6.1 AGI Levels

> *The measure of intelligence is the ability to change.*
>
> — *Albert Einstein*

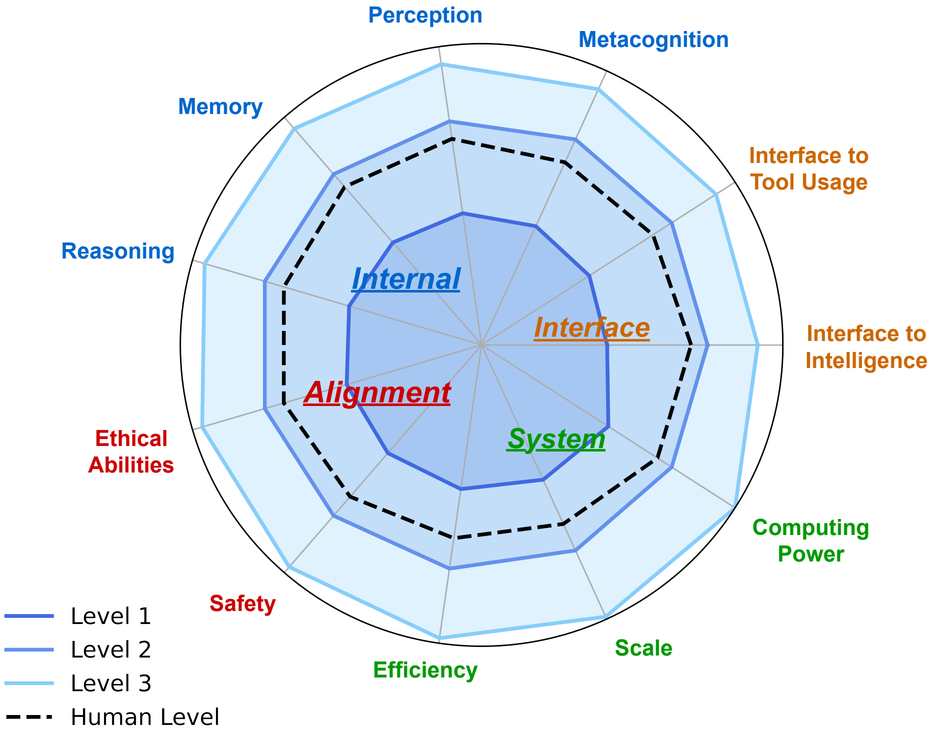


Figure 8: **Radar Chart Depicting the Multi-faceted Approach to Evaluating AGI Readiness Across Four Core Domains.** Internal, Interface, System, and Alignment. The Internal domain evaluates fundamental cognitive abilities such as Reasoning, Memory, Perception, and Metacognition. The Interface domain assesses the AGI's Tool Usage capacity and ability to Link to Intelligence. The System domain focuses on operational aspects, including Efficiency, Scale, and Computing Power. Lastly, the Alignment domain looks at ethical considerations and safety measures with components like Ethical Abilities and Safety. The chart illustrates the progress levels of AGI capabilities, ranging from Level 1 to Level 3, with a dashed line representing Human Level performance for comparison.

Inspired by Morris et al. (2024), which suggests six principles that an effective AGI ontology should satisfy, we define three AGI levels with their major characteristics (Table 1). The main objective is to situate the current AI development, quantify existing limitations, and motivate future endeavors toward next-level capability, evaluation, and alignment. In Figure 8, we also visualize the performance comparison of the three levels against humans regarding the core domains as discussed in our previous sections, which breaks down the fundamental differences among them.

Level-1: Embryonic AGI This level of AGI *usually performs better or on par with humans at specific **benchmark tasks*** (Bugaj and Goertzel, 2009). Level-1 AGI represents the current state-of-the-art AI systems. For example, GPT-4 exhibits remarkable capabilities across many natural language tasks including language understanding, generating coherent and contextually relevant responses, often on par or superior to humans. These systems can often perform well given large enough human-curated datasets and are able to assist humans in certain domains. As indicated by the research (Bubeck et al., 2023), we are currently at this level of AGI in many domains.

Level-2: Superhuman AGI The key turning point from Level-1 to Level-2 is the AI's ability to *fully replace human in **real-world tasks and applications***. They excel in terms of effectiveness (e.g., higher accuracy, better problem-solving skills), efficiency (e.g., faster processing speed, higher throughput, ability to handle massive amounts of data), and reliability (e.g., higher success rates, resistance to fatigue, enhanced safety guarantees). These systems might also learn from limited data, generalize knowledge across domains, adapt to novel situations with relatively little human intervention, and exhibit creativity and innovation in their approaches. They can also engage in complex decision-making processes, considering multiple factors and optimizing outcomes based on predefined objectives. Notably, Level-2 AGI should be ready to deploy in the real world and resolve the complex real-world tasks that are currently solved by humans today, *without any human intervention.* In our opinion, very few AI systems have achieved Level-2 except in highly specialized domains, e.g., playing the Go game.

Level-3: Ultimate AGI While Level-2 AGI is able to replace humans in solving many tasks, the creation of the Level-2 AGI inevitably still requires human efforts. We argue that the essential milestone of Level-3 AGI is that *given a certain goal, possibly vague and high-level, such AGI system can* ***fully self-evolve*** *without any human intervention.* This level marks the pinnacle of AGI development, which represents an idealized and possibly unattainable AI system. The ultimate AGI would possess the ability to learn, reason, and make decisions at a level far beyond human capacity, and liberate human involvement in the development process of such AGI system as well. Consequently, at this stage, ensuring that such Level-3 AGI has a strong alignment with human values and goals becomes even more important. Additionally, Level-3 could demonstrate *deeper human emotions* such as empathy, *social awareness* which allows collaborating seamlessly with humans and other AI systems, and even the spark of *self-consciousness.* However, realizing the ultimate AGI remains a theoretical concept, and its feasibility is subject to ongoing research and debate.

The Progression of Exemplary AI Systems over the AGI Levels Given that we are still at the early stage of AGI, we acknowledge that our definitions might be high-level and abstract but serve as theoretical guidelines. Therefore, to facilitate the understanding and better persuade the readers of the validity as well as generality of our definitions, we give several concrete examples in this section where we feature the main capabilities of each AI system as they evolve over the levels:

- **Personal assistant.**

  1. Each type of assistant can provide constructive feedback to users for *a specific task* such as coding, artistic design, and health management. Their feedback usually still *requires careful examination* from the users and often needs *a couple of trials* before arriving at the ideal answer.
  2. AI assistants need less explanation from the users and can *effectively utilize third-party tools* for knowledge retrieval and verification. At this point, the assistants will take over the responsibility in an *end-to-end fashion* rather than only providing solutions to a specific subroutine. For example, the code assistant will not only generate code but also assemble corresponding tests and supervise the deployment process; the writing assistant can also initiate the publishing and lead the marketing and selling.
  3. The “Personal Assistant” appears that unifies and orchestrates several level 2 assistants and only requires very *high-level instructions from human* without specifying the sub-procedures. These assistants can *anticipate the concern* of the user and *propose multiple alternatives* with their pros and cons, offering the maximum flexibility and tailoring to the taste of each user.

- **Auto transportation.**

  1. Self-driving (L2) cars are widely seen nowadays, facilitating not only drivers with disability but also those who enjoy the semi-autonomous driving experience. In many closed facilities such as hotels, robots can reliably deliver food or items, which greatly preserves the privacy of the guests and saves the human cost. However, these semi-autonomous agents usually *operate under a controlled environment* or still *require humans in case of emergency.*
  2. Transitioning to level 2, not only will we reach the end level of self-driving where drivers can completely free themselves from the duty but also the traffic system can connect all vehicles on the road for better safety control. Vehicles can easily accommodate various complex road conditions, and even in case of emergency, the system is equipped with the best devices to reduce potential damage.

3. The whole city or even the globe is connected with *the ultimate safety guarantees*. High-level planning constantly monitors all moving vehicles and can *dynamically prioritize tasks* that are of high importance such as emergency rescue and transportation coordination under special events. Personalized driving experience is also emphasized for those with different driving experience preferences. Finally, transportation is not limited to just cars but also flights and other robotics.

- **AI-augmented video games.**

  1. Integrate simple game agents for tutoring and storytelling, which can adjust their strategies and behaviors based on the player's input. These usually require *specifying manual conditional rules*, *coding game-specific algorithms*, or *applying current game AI models*.
  2. Game agents start to spark *deeper intelligent behaviors, including virtual companions*, and *develop innovative game-play* that often surprises both the designers and players while following the original game concept. Multi-agent interactions among themselves and with the players will *generally feel engaging*. The role of AI spans beyond just role-playing: *content creation becomes ubiquitous*, including but not limited to world generation, motion synthesis, story expansion, and even coming up with intrinsic motivation to enrich the game itself.
  3. AGI-augmented game will *break through the virtual world*, connecting players and even the physical world via many different media such as brain-machine interface, AR, and VR. This also becomes closer to the realization of the *Metaverse* where most people can immerse themselves without realizing whether the experience is virtual in a dynamic and stateful game space.

Category	Characteristics	L1	L2	L3
General	Surpasses human performance in specific domains	✓	✓	✓
	Surpasses human performance in real-world scenarios	✗	✓	✓
	Self-evolve without human intervention	✗	✗	✓
Internal	Adapts to novel situations with minimal human intervention	✗	✓	✓
	Generalizes knowledge across domains	✗	✓	✓
	Exhibits creativity and innovation	✗	✗	✓
	Engages in complex decision-making processes	✗	✗	✓
Interface	Collaborates seamlessly with humans and other AI systems	✗	✓	✓
	Learns to create new tools autonomously	✗	✓	✓
	Continuously improves through self-learning and adaptation	✗	✗	✓
	Demonstrates empathy, emotional intelligence and social intelligence	✗	✗	✓
System	Enables super stable, low latency, and high-throughput serving	✓	✓	✓
	Built with data, power and compute efficiency	✗	✓	✓
	Supports automatic learning, adjustment, collaboration, and deployment	✗	✗	✓
Alignment	Accurately follow human instructions	✓	✓	✓
	Accurately follow a given user's preference	✗	✓	✓
	Aligns strongly with both user-level and society-level human values and goals	✗	✗	✓

Table 1: **Comparison of AGI Levels and Their Characteristics.** "L1", "L2", and "L3" refer to "Level 1", "Level 2", and "Level 3" of AGI respectively. For each of the main categories, we list several major conceptual criteria in terms of several categories that can be used to assess whether we have reached a certain level of AGI.

6.2 AGI Evaluation

> *For better or worse, benchmarks shape a field.*
>
> *— David Patterson, Turing Award laureate 2017*

The concept of evaluating AGI traces back to the famous Turing Test proposed by Alan Turing in 1950 (Turing, 1950). Turing posited that a machine could be considered intelligent if it could converse with a human so that the human could not distinguish whether they were conversing with a machine or another human. This laid the groundwork for the field of AGI evaluation. However, the Turing test has several drawbacks, such as its reliance on deception, subjective evaluation, and narrow focus on language use. To address these issues, a more comprehensive approach called the I-athlon (Adams et al., 2016) has been proposed, which evaluates machine intelligence across multiple dimensions and aims to provide a more objective and practical method for assessing progress in general-purpose AI.

Over the decades, various approaches have been proposed to assess the capabilities of AI systems. Early attempts drew parallels to human intelligence, using metrics like IQ scores to characterize AI performance (Bringsjord and Ferrucci, 2003). Others explored whether AI systems could achieve educational milestones like earning a university degree (York and Swan, 2012).

Developing reliable and meaningful evaluations is essential for transforming research ideas into real AGI systems and products that can benefit human beings, and at the same time, help steer the exploration of new models towards AGI. In this section, we first describe what properties ideal AGI evaluation pipelines should possess, highlight their relationship with our previously discussed AGI components, and discuss the challenges in designing more sophisticated evaluation frameworks. Then, we will give an overview of the recent efforts on large model evaluations and their limitations, which establishes the basis for how we can effectively progress toward AGI evaluations.

6.2.1 Expectations for AGI Evaluation

Key Characteristics The span of AGI systems' capabilities is growing rapidly in terms of modality, interactivity, complexity, task generalization, etc. Researchers and engineers, hence, need a more refined definition for the characteristics that successful AGI evaluation should acquire:

- **Comprehensiveness.** Comprehensive evaluation aims at two generally contradicting aspects: 1) *Diversity* asks for the inclusion of a wide variety of testing examples in terms of the absolute number, domains, tasks, types, and formats, which can hopefully cover as many real application scenarios as possible. 2) *Generality* requires examining the model's performance on similar but unseen tasks with optional few-shot examples, a property that has always been considered as a prerequisite for adaptability and self-learning (Brown et al., 2020; Dodge et al., 2021).

- **Fairness.** Fairness and equity as first-class aspects of evaluation are essential to ensuring technology plays a positive role in social change (Liang et al., 2023; Bommasani et al., 2021), we divide the fairness aspect into three concepts below 1) *Unbiasedness* of a test refers to the desirable attribute such that the tested model exhibits no preference towards a specific subdomain of knowledge or bias. 2) *Dynamism* aims to reduce the effect of unfair evaluation resulting from data contamination and over-fitting. A dynamic benchmark will likely improve the evaluation results statistically and also eliminate the false success from static pattern recognition. 3) *Openness* promotes the transparency of the test procedure and data such that the test results are easily replicable and interpreted, and the dataset can strike a balance between public and hidden data for being less vulnerable to hacking.

- **Efficiency.** Efficiency is crucial in evaluating models with ever-growing parameter size (Henderson et al., 2020; Schwartz et al., 2020; Bommasani et al., 2021). We propose to include *Autonomy* and *Low-variance* in this evaluation concept. 1) *Autonomy* liberates most of the human participation from the loop and therefore, minimizes the cost for each evaluation and motivates larger scale, wider range, and longer dependency testing. 2) *Low-variance* is a key property that allows using minimal test resources to produce statistically significant and practically meaningful evaluation results for comparison.

It is undeniably challenging to design evaluation frameworks that satisfy all the recommended characteristics, but a well-constructed evaluation pipeline can help reflect the true power of increasingly sophisticated AGI systems.

Relation to AGI Internal, Interface, and Systems The new generation of model evaluations should focus on assessing AI systems across multiple dimensions, considering not only their ability to engage in human-like conversation but also their capacity to reason, learn, adapt, and solve complex problems. These tests should encompass a broader range of cognitive abilities (Lebiere, 2007) and evaluate the AI's performance in real-world scenarios. Drawing from the concepts of AGI internal, external, and system levels as we discussed in earlier sections, we can outline the key aspects that a new AGI evaluation framework should address:

- **Internal level.** Assess the AI's ability to represent and process knowledge, reason abstractly, and generate creative solutions. This could involve tasks that require the AI to demonstrate common sense reasoning, causal understanding, and the ability to learn and adapt from limited data.
- **Interface level.** Evaluate the AI's capacity to perceive, interpret, and interact with the external world. This could include tasks that test the AI's ability to process and integrate information from multiple modalities (e.g., vision, language, and sensory data), navigate complex environments, and manipulate objects to achieve specific goals.
- **System level.** Examine the AI's overall behavior and its ability to pursue long-term goals, collaborate with humans and other AI systems (Wang, 1006), and make ethical decisions. This could involve scenarios that assess the AI's alignment with human values, its transparency and explainability, and its robustness and reliability in uncertain and adversarial situations.

By focusing on these aspects, a new AGI evaluation framework can provide a more comprehensive assessment of an AI system's capabilities, potential, and alignment with human values. This approach would help ensure that the development of AGI remains beneficial and aligned with the interests of humanity, while also fostering a deeper understanding of the nature and limitations of artificial intelligence.

Challenges in AGI Evaluation Design As we will discuss in the next section, the current evaluation frameworks are far away from achieving what we expect (Team, 2023). Here we list a couple of challenges associated with AGI evaluation (Xu and Ren, 2022) design, and we provide some concrete examples in each category:

- **Non-standard output.** Moving towards more modality and richer action space, the output might surpass our current familiar ones such as images, texts, and audio. Evaluating non-standard or mixed output becomes much more challenging, especially if we want to standardize an automatic procedure. For example, the result of a scene synthesis model can be drastically different depending on the scene representation (Feng et al., 2023; Liu et al., 2023i), which often requires other surrogate metrics that are often biased and limited. The quality of program generation (Chen et al., 2021b; Rozière et al., 2024) is notoriously difficult to measure since there is no single metric that can holistically capture it (performance, readability, coherence to human coding style, etc).
- **Output space explosion.** Often, a question has multiple acceptable answers with different degrees of satisfaction, which is often not considered in standard metrics. As the model becomes more creative and diverse, it is crucial to consider this expanding space of possibilities. For instance, the admissible outputs for storytelling and design-related applications are usually very big, which demands more complicated metrics to consider both the validity and the diversity of the generations.
- **Subjective feedbacks.** As models start penetrating deeper into people's lives, practitioners need to pay more attention to how users think about them. However, different users will naturally have distinct feelings towards even the same agent, posing an extraordinarily challenging problem: how can we take each individual's subjective feedback into account? One salient example is the potential emergence of emotional support AI, which needs to build extensive personal connections with the user, and hence, measuring its success qualitatively and efficiently requires a lot more careful effort.
- **Long feedback loops.** The trend toward more general-purpose AI indicates that models would gradually become more in people's lives, making them more interactive. Instead of single-task solving, AGI systems will get multi-step feedback, often extending a longer period, making evaluation more complex. One

commonly seen case happens in search engine development where a mature system needs to improve the click-through rate and track down users' following actions such as purchases, comments, and after-sale satisfaction. Another potential situation appears when a health AI robot monitors a patient's biological status constantly. Both of these require evaluations that span an elongated period.

- **Complicated environment setup.** Many tasks inevitably presuppose more complicated environment setups such as robotic manipulation, self-driving assistance, and program synthesis. Scaling the evaluation of these applications necessitates a higher level of automatic environment generation, verification, and measurement. Often, the challenge associated with these tasks also comes with the difficulty of applying a single metric for comparison and can usually encounter many physical constraints (Peng et al., 2024b) that are hard to overcome.
- **Super-evaluation.** Similar to super alignment, when AI models start to surpass humans, the evaluation becomes prohibitively more challenging. Taking the example of theorem proving, one can imagine that the real meaning of these models comes when they can prove unsolved theorems, which require greater expertise to verify. With the current setup, the framework might determine whether an AI agent can beat humans for a specific task (e.g., AlphaGo Silver et al. (2017) and (Lake and Baroni, 2023)) but might not be as confident to access its absolute ability beyond. Fortunately, formal proof systems such as Lean (De Moura et al., 2015) may be helpful. Lean is an interactive proof system that utilizes formal logic to verify the correctness of mathematical theorems and computational outputs. As AI models begin to generate results that exceed human verification capabilities, systems like Lean become indispensable for ensuring the validity of these outputs.

It is important to consider some of these aspects when designing next-generation AGI evaluations. A more robust evaluation framework gives a more accurate estimate of a model's potential and hints at where our technology currently resides on the AGI-level hierarchy.

6.2.2 Current Evaluations and Their Limitations

In this section, we summarize several representative works focusing on existing AI evaluation benchmarks. One category aims to provide a single pipeline suite consisting of many different tests such as OpenCompass (Contributors, 2023), AGIEVAL (Zhong et al., 2023a) and Huggingface Open LLM Leaderboard [9] for language understanding, reasoning, knowledge, and interaction abilities, and MT-BENCH (Zheng et al., 2024) for multi-task generalization ability. The GAIA benchmark (Mialon et al., 2023) constitutes a significant stride in this direction. It is specifically designed to evaluate General AI Assistants, presenting a series of real-world questions that test fundamental competencies such as reasoning, multi-modality handling, web browsing, and tool-use proficiency. AGENTBENCH is another comprehensive benchmark (Liu et al., 2024c) suite designed for evaluating the efficacy of LLMs as autonomous decision-making agents across eight interactive environments, highlighting the performance discrepancy between leading commercial models and open-source counterparts. On a granular level, there are many prior efforts in accessing a model's specialized ability, which can be roughly divided into several aspects: accuracy, calibration and uncertainty, robustness, fairness, bias and stereotypes, toxicity, and efficiency (Liang et al., 2023), which can be further classified into two sets based on their objectives. OpenAGI (Ge et al., 2023b) is an open-source platform designed to advance AGI by integrating LLMs with domain-specific expert models. It utilizes a dual strategy of benchmark and open-ended tasks to evaluate.

Besides classifying these into ability and constraint testing, in this part, we further open a discussion about "how" and "what" do we evaluate for the current state:

"How" Do We Evaluate: Two Major Techniques The "how to" category consists mostly of two techniques: human and AI evaluations. Methods following this set can be subject to an individual evaluator's preference or a model's bias, which needs to be taken into account for a fair comparison.

- **Human evaluation.** The performance of AI agents is directly evaluated by invited humans or experts. This method is of high quality but often not scalable because it is expensive to invite humans or experts.

[9] `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`

For example, in HuggingGPT Shen et al. (2023b), the invited experts annotate tasks for intricate requests with 46 examples to create a dataset with quality guarantees. Besides, in TOOLLLM Qin et al. (2023b), they invite humans to label the pass and win rates for different methods. Another very attractive property is that, with sufficiently detailed specifications, a highly complicated and flexible output format can also evaluated.

- **AI evaluation.** Compared with human evaluation, AI evaluation is more scalable, but there is no guarantee for the quality of the evaluation, and it is highly dependent on the judge model. One typical instance of LM-based evaluation is the one used in HuggingGPT Shen et al. (2023b) where they utilized GPT-4 as a critic to assess the accuracy of the planning.

"What" Do We Evaluate: Existing Evaluation Aspects This "what" question, on the other hand, often involves a static dataset coupled with automatic metrics. However, it is worth noting that these two types of evaluation methods often appear simultaneously. For example, HuggingGPT Shen et al. (2023b) leverages F1 and accuracy as the metrics for the single task, F1 and normalized Edit Distance Marzal and Vidal (1993) for the sequential task. Due to the wide variety of tasks, we give an exemplary dissection of the most commonly used benchmarks in natural language processing.

- **Core-knowledge.** The goal is to access the internal knowledge of a (pre-trained) large model, with the most common ones being MMLU (Hendrycks et al., 2020), MMMU (Yue et al., 2023), and AGIEval (Zhong et al., 2023a). The key characteristic is the width of knowledge domains, which is usually fact-based. They typically require LLMs to generate a short, specific answer to benchmark questions that can be automatically validated and can often be used as a measurement to evaluate the hidden potential of pre-trained models after fine-tuning.
- **Instruction following.** This tests the fine-tuning and alignment of a pre-trained model, which concentrates not only on the correctness and soundness of the output but also on how closely the model can follow the instructions and guidance. Examples include Super-NaturalInstructions (Wang et al., 2022b), Self-instruct (Wang et al., 2022a) and Flan (Longpre et al., 2023; Wei et al., 2021), which contain slightly more open-ended questions and more diverse tasks. One particularly important subclass is question answering, such as TriviaQA (Joshi et al., 2017) CoQA (Reddy et al., 2019), SQuAD (Rajpurkar et al., 2016), and Natural Questions (Kwiatkowski et al., 2019), that are used in almost all essential benchmark suite.
- **Open-ended conversation.** Going beyond single-turn QA and instruction following datasets, this category of tests attempts to evaluate a model's ability to engage in multi-turn conversations. MT-Bench Chatbot Arena (Zheng et al., 2024) and MMDialog (Feng et al., 2022) are designed to support more open-ended multi-turn QA testing, which resembles more closely to the most frequently applied scenarios of chatbot models.
- **Robustness and bias.** Beyond the standard accuracy or reasoning ability of models, a long-concerned aspect is the robustness of a model. The key question is whether the model is robust to invariant input perturbations and able to consistently output unbiased outcomes. On a vast range of tasks like language modeling (Liang et al., 2023; Ni et al., 2023), classification (Brendel et al., 2019; Subbaswamy et al., 2021; Guo et al., 2023), multi-modal content generation (Cui et al., 2023), the concept of robustness evaluation has been taken into consideration.
- **Efficiency.** Efficiency is another crucial aspect of utilizing language models on both evaluation and training (as mentioned in Sec. 4) since high inference costs can limit their accessibility for a broader range of users (Strubell et al., 2019; Schwartz et al., 2020; Henderson et al., 2020; Bommasani et al., 2021; Liang et al., 2023). For example, users may be prone to incur a 10× increase in responding time or cost for a model that only marginally decreases task performance by 0.1%.
- **Creativity.** Ever since the born of generative models, research has been ongoing into the use of them to model human creative processes, to mimic or complement them, in art, music, and literature (Cardoso et al., 2009; Colton et al., 2012). Creativity is mainly linked to the diversity of generated content with

specific tasks like storytelling (Alhussain and Azmi, 2021) emerges. Recent works focus on understanding and prompting LLMs to generate creative textual content (Zhao et al., 2024c; Shanahan and Clarke, 2023).

The focus of this section is not to exhaust all possible benchmarks that are popular in different fields of AI but to give a sense of the current status quo from the lens of LLMs and their limitations.

Limitations Here we wish to initiate some discussions of the limitations of current evaluation methodology, which hopefully can inspire future endeavors toward developing more well-rounded and robust evaluation frameworks:

- **Going beyond numeric metrics.** Turning qualitative results into quantitative metrics will unavoidably result in bias and loss of information. And there is a lot of feedback that is hard to express and compare numerically. For instance, users' preference towards a persona chatbot can be extremely complicated, and often, a mixed feeling becomes inherently inappropriate to be quantified by a single number. Besides, combining multiple numeric metrics into one for global comparison will also face the issue of weighing them, which is usually biased and non-robust when decided with only prior or domain knowledge.
- **Surrogate metrics.** As discussed in the AGI evaluation design challenges, often we will face applications in which even defining what a performant model means is hard, not even to mention evaluating it. In this case, people usually resort to surrogate metrics that are more familiar as a means to approximate their performance. However, as we step towards more general AI, this would happen more often, and hence, more sophisticated ways to construct metrics that are closer to the true goal are needed.
- **Lack of failure analysis.** Almost all benchmarks we have seen so far give aggregated results, usually in the form of averages and percentiles. However, benchmarks should in principle provide more insights into improving a certain system. The most informative feedback would contain information about the analysis of the failing or worst cases. This can also showcase the potential pitfalls and risks associated with a specific system to help with the decision for danger-critical applications.
- **Missing more general tasks.** We can expect that the advancement of models might be faster than that of evaluations. Therefore, we desperately demand more general tasks to access the model's performance in a controlled environment, leading to the embryonic forms of AGI evaluations. Examples include the modern Turing test [10], the coffee test [11], and the robot college student test [12].

6.3 How to Get to the Next AGI Level

> *Technology is anything that wasn't around when you were born.*
>
> *— Alan Kay, Turing Award laureate 2003*

Considering the AGI level definition in Section 6.1, we briefly summarize the high-level guidance that could help transcend the limits of each level:

From Level 1 (Embryonic AGI) to Level 2 (Superhuman AGI) The transition from embryonic AGI to superhuman AGI requires substantial improvements in the scale and scalability of AI models, as well as in the size and quality of the data used for training. This phase aims to enhance the generalization capabilities of AI systems so they can effectively understand and interact with the complexities of real-world situations and apply their acquired knowledge to new contexts. As AI capabilities exceed humans, the focus shifts toward enabling AI systems to engage in self-improvement and autonomous innovation, allowing

[10] An agent is requested to earn one million dollars given a start funding of hundred thousand dollars.

[11] An agent is tasked to figure out how to make coffee, which involves a series of sub-tasks such as entering an arbitrary American apartment, locating the coffee machines and ingredients, coming up with a standard procedure for brewing coffee, and actually perform the mechanistic actions.

[12] An agent is told to enroll in a university, perform as a human student, take the same classes, and finally graduate with a degree in a timely manner.

them to address problems and generate insights at unprecedented levels. However, this advancement also necessitates the implementation of robust safety protocols and ethical guidelines to mitigate the potential risks associated with superhuman AI. Ensuring that the development and deployment of superhuman AI align with human values and contribute to societal betterment is crucial, marking a pivotal moment in considering the implications of surpassing human intelligence.

From Level 2 (Superhuman AGI) to Level 3 (Ultimate AGI) The transition from superhuman AGI to ultimate AGI represents the most ambitious and challenging leap in the evolution of artificial intelligence. This phase involves enhancing AI's ability to seamlessly integrate and synthesize information across disparate domains, enabling unparalleled levels of innovation and problem-solving. The development of ultimate AGI requires a solid framework for continuous learning and adaptation, pushing the boundaries of what AI can achieve in terms of reasoning, intuition, and creativity. Moreover, this transition underscores the need for rigorous oversight and ethical frameworks that are continually updated to match the pace of AI's evolution, ensuring that ultimate AGI functions in a beneficial and non-threatening manner to humanity. This stage represents not only the peak of technological progress but also raises profound ethical and existential questions regarding the role of AI in the future structure of society.

Conceptual Solutions to Achieve Level 3 (Ultimate AGI) Based on the above high-level guidance about transiting to the next-level AGI, we further give two conceptual solutions that can reach level 3 (ultimate AGI).

- **Automated Coding AI: Bridging the Gap to the Ultimate Artificial Intelligence** Coding AI refers to AI systems capable of automatically planning and generating code to solve complex tasks. We believe that the advancement of such systems could significantly accelerate progress toward the Ultimate AGI. The AGI in levels 1 and 2 is limited since they require a variety of data collected by humans and require specific optimization goals designed by humans when they tackle different tasks. Human-in-the-loop makes it impossible for AGI at these two levels to realize self-evolve. Advanced coding AI solves the above challenges in two ways: 1) They enable AGI to interact with the real world and obtain large amounts of domain data. In the scenario of a single agent, an advanced coding AI takes writing code as the most basic tool for AGI to interact with the world, enabling AGI to plan and reason in the form of code and get feedback on the real world through the results of code. When it comes to multi-agent scenarios, each agent can be regarded as a unique coding AI based on their profile. Then, through the interactions between agents and the interaction between agents and the real world, AGI can obtain enough data for self-training and evolution. 2) It makes it possible for AGI to automatically define optimization goals for different tasks. With the ability to write codes, AGI can do try-and-error in the different tasks and obtain feedback through the interaction between code and the real-world environment. This feedback contains information about how well the AGI has completed its tasks and can take many forms, which can be differentiable or non-differentiable. Some techniques based on reinforcement learning can be introduced to use these different forms of feedback to align and self-evolve the AGI. In this case, the evolution of stronger AGI can be achieved without requiring humans to specify goals. More information about coding AI could be found on Sec 7.5

- **Super realistic simulation promotes the complete application of ultimate AGI in the real world.** The main limitation of AGI in Levels 1 and 2 is that the results of algorithms achieved on manual benchmarks and environments do not match the real world. The huge difference between the real environment and the benchmark is a huge challenge that affects the deployment of AGI in the real world. Super realistic simulation techniques make the deployment of AGI in the real world possible from the following aspects: 1) Realistic simulation can generate a large amount of high-quality data for AGI to perform self-training and self-evolving. Current benchmark data are often collected or designed by humans and have noise and bias compared to real-world scenes. Realistic simulation based on some data-driven techniques like VAE (Kingma and Welling, 2013) and GAN (Goodfellow et al., 2020), Transformers (Micheli et al., 2023), and Diffusion Models (Ding et al., 2024a; Alonso et al., 2024) can provide unbiased data to AGI to achieve better alignment. 2) AGI's algorithms and strategies only need to be fine-tuned on the realistic simulator before they can be applied to the real world. Realistic simulators can not only simulate the interaction of different AGI agents in the real world but also reflect the causal laws of the real world. This allows the effects of AGI's algorithms and strategies in the simulator to be replicated in the real world.

Challenges Along the Way to the Ultimate AGI While the concept of ultimate AGI holds immense promise, it is essential to acknowledge the inherent constraints and challenges that may limit its realization. Here we list a couple of them, with which we hope to give readers a sense of the intrinsic difficulty of approaching the ultimate AGI as well as motivate more innovative research across various domains:

- **The need for advancement from various disciplines.** One never-ending debate about the potential realization of AI is whether artificial neural networks (ANNs) are the right way to go. Although the success of neural networks is undeniable, it is worth thinking about other possibilities as we get closer and closer to AGI. This, however, requires a deeper understanding of 1) other foundational disciplines (than computer science), such as mathematics and physics, which can provide more sophisticated formal language to conceptualize AI; 2) scientific research in biology, chemistry, and neuroscience, which better explains the biological mechanisms of intelligence; 3) engineering and manufacturing technologies which build up the necessary tools to instantiate AGI systems. A holistic comprehension and collaborative effort among researchers from multiple domains will likely become indispensable during the AGI revolution, which not only brings excitement but also presents respective challenges.

- **Social acceptance.** As AGI advances, its social acceptance and seamless integration into daily life and critical sectors, such as healthcare, finance, and the military, present significant ethical, moral, and social dilemmas. Public concerns typically focus on issues related to privacy, autonomy, and the possible displacement of jobs due to automation, which can foster resistance to the adoption of AGI systems. Additionally, the cultural and social influence of AGI should not be overlooked. Each community's response will vary depending on its values, norms, and historical context, potentially leading to different levels of acceptance or opposition in various demographic groups. Critically, although AGI may have the ability to make well-informed decisions, there may be a reluctance to allow it to replace human judgment in making vital decisions, particularly those affecting human lives. Therefore, a series of respective social policies and educational activities might be initiated to regulate and promote the integration of AGI technologies into society.

- **Fundamental limitations governed by physics laws.** Fundamental limitations exist in the real world that might limit our progression toward the ultimate AGI. The power structure (consumption), computational efficiency, as well as natural and human resources should be taken into consideration when we develop AI systems: on the one hand, at some point along the journey, the main question that we need to think about might no longer be about whether we *can* but whether we *should* create a specific AI system due to its tremendous cost in terms of all aforementioned aspects; on the other hand, these fundamental limitations governed by physics laws such as only being able to arrange a limited number of semi-conductors onto a 2D plane without over-heating will push researchers and engineers to think in a different way forward. Besides, even though the promise of the ultimate AGI is exciting, we should also be cautious and more conservative about its capability as there are intrinsic challenges that can not be easily overcome, such as the speed of light and the dimension of space.

- **A Call for rethinking and redefining the ultimate AGI.** As we currently stand at the first stage of our AGI hierarchy, it is very possible that our understanding of higher-level AGI remains shaky or becomes outdated. Therefore, researchers might need to *rethink and redefine what the ultimate AGI really is* as we progress along the journey, the answer of which might depend on our gradually increased understanding of the difference between artificial and biological intelligence from both a biological and philosophical perspective, and could even be eventually limited by our current imagination. Once our understanding and goal change, a new set of evaluation frameworks and alignment procedures should be developed accordingly to meet the new expectations. It is worth keeping in mind the possibility that technical advancement might be "local" and people need to restore the wheels at some point in order to break the constraints towards AGI.

6.4 “How Far Are We from AGI” Workshop Discussions

> *None of us is as smart as all of us.*
>
> — *Ken Blanchard*

The following subsection presents a synthesis of perspectives from respected researchers in the AI field, as reported in their presentations at the “How far are we from AGI” ICLR 2024 workshop [13] and associated panel discussions. The summary of these views from this workshop has been compiled with the consent of the relevant participants:

Oriol Vinyals: From AI to AGI The rapid development of AI has given people a lot of expectations and imaginations for a more powerful AGI. In today’s era, analysis based on current AI development trends and deficiencies is an important way to measure our distance from AGI and how to achieve AGI.

- **Defining AI and AGI.** The definitions of AI and AGI are topics that have been hotly discussed. In 1997, Mark Gubrud described Al systems as that “can acquire, manipulate and reason with general knowledge, and that are usable in essentially any phase of industrial or military operations where human intelligence would otherwise be needed.” Then in 2001, Ben Goertze needed a title for a book he was editing about Al systems that are general, like the old goal of Al. Shane suggested he add the word “general” to make the new term Artificial General Intelligence, or AGI. Therefore, they started using the term AGI in various online forums and it caught on from there. Based on this definition, Oriol Vinyals concluded an AGl is a machine that can do the kinds of cognitive tasks that people can typically do. Moreover, based on the definition of AGI, Merrie Morris recently led the writing of a paper (Morris et al., 2024) about the definition of AGI breaking the concept into six different levels. For example, “Competent AGl”, which corresponds most closely to what most people mean by “AGl”, is defined as: performance at least at the 50 percentile for skilled adult humans on most cognitive tasks.
- **AI: deep learning era.** Today’s AI is in the development era of deep learning. The development of AI has seen many major breakthroughs in recent years, such as AlphaGo (Goodfellow et al., 2020) and AlphaStar (Vinyals et al., 2019a). However, many Al demonstrations focus on models trained to excel in one domain. Specifically, their algorithms are general, like Neural Nets, SGD, Supervised Learning, and Reinforcement Learning. However, their models are not general since they can not do the kinds of cognitive tasks that people can typically do.
- **Bringing the “G” back to AGI.** To make current AI more general, people have tried to develop a more powerful model:
  1. *General text models.* Efforts have been made all the time to develop more powerful general text models. In 1951, Shannon et al. proposed 3-gram to point to ninety-nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions. Then in 2011, Sutskever et al. designed RNNs (Graves, 2013) to process time sequence data. In 2016, Jozefowicz et al. proposed BIG LSTMs (Graves and Graves, 2012) to tackle the ever-changing environmental challenges online like long-term dependence on super long sequence data. Recent years have witnessed the big success of GPTX, which can learn tasks such as question answering, machine translation, reading comprehension, and summarization without any explicit supervision when trained on a new dataset of millions of webpages called WebText.
  2. *General multimodal models.* General multimodal models are crucial as they can process and understand complex information from various modalities, such as text, images, and audio, enabling more comprehensive and nuanced analysis. These models play a vital role in tasks like natural language understanding, image recognition, and audio processing, contributing to advancements in AI applications across diverse domains. In these works, Gemini (Team et al., 2023) has played an important part, which supports interleaved sequences of text, image, audio, and video as inputs.

[13] https://agiworkshop.github.io

3. *Long context to learn more complex tasks.* With the advent of the multimedia era, people increasingly hope that AI can assist humans in understanding books, movies, and long videos. To solve the challenge, Gemini 1.5 Pro (Reid et al., 2024) has been proposed to achieve a breakthrough context window of up to 1 million tokens, the longest of any foundational model yet. For applications, it can understand and summarize videos and fix codes for people.

- **How far are we from AGI?** Two valuable opinions are provided to discuss this topic. In Shane's 2009 prediction, there is a 50% chance of AGl by 2028. In addition, Metaculus thinks that Al must pass a 2-hour, adversarial Turing test in text, images, and audio files, have robotic capabilities to assemble an automobile model, and have high performance on difficult cognitive tests to achieve the first AGI system.

Yejin Choi: On AGI: Ambiguities, Paradoxes, and Conjectures AGI is ambiguous and presents paradoxes in current AI observations, and we can make some conjectures based on them.

- **AGI is ambiguous; denial is futile.** As much as we cannot clearly define and measure human intelligence, we won't be able to clearly define and measure artificial intelligence. That doesn't mean we should throw out the concept of AGI. A squish, ambiguous concept can be a fascinating object of scientific research. In fact, language is a squish concept, yet we study it as a scientific object. It might be analogous that future research must embrace ambiguity.
- **Generative AI paradox — what it can create, it may not understand.** For generative models, hard could be easy, and easy could be hard. For humans, generating high-quality images or text is harder than understanding them, but for AI, the situation is reversed. Models do not need an understanding to produce quality content. For example, models can generate high-quality images beyond human capabilities, but they often make mistakes when asked to select one of their own generated images based on specific criteria.
- **Commonsense paradox — common sense is not so common.** LLMs lack a coherent Theory of Mind and struggle with many basic common sense tasks. In this way, they are incredibly smart and shockingly stupid at the same time.
- **Cringe speculations on arrival.** There is a 30% chance that within 3 years, we will have a language-only AI that is perceived as AGI-enough by about 30% percent of people. There is a 50% chance that we will have AGI by 2050, assuming models are tested for autonomous, long-horizon interactions.
- **Multi-paths to AGI hypothesis.** We may have multiple species of digital intelligence developing along entirely different routes, each with different strengths and weaknesses, and without a clear dominance form. Scale-based AI will be impressive but will suffer from bind spots coming from over-dependence on data, so we should avoid concentrating all the power on this approach.

Andrew Gordon Wilson: How do we build a general intelligence? From a probabilistic perspective, generalization depends largely on two properties of deep learning models, the support and the inductive biases. Starting from this, we can try to reason about whether we can build generally intelligent systems through the lens of Kolmogorov complexity and generalization bounds. Looking ahead, although there have been different signals showing the possibilities of building broadly intelligent systems, we might be still far away from doing that. In the future, we should embrace many safety considerations and alignment approaches when building these systems. Andrew Gordon Wilson introduces his views on how to build general intelligence as follows:

- **Perspectives of understanding deep learning models.** We can use probability theory to develop a prescriptive understanding of model construction and generalization. Specifically, from a probabilistic angle, the ability of a system to learn is determined by its support and inductive biases. We want the support of the model to be large so that we can represent any hypothesis we believe to be possible. Meanwhile, we also need the inductive biases to carefully represent which hypotheses we believe to be a priori likely for a particular problem class. From this probabilistic perspective, we should not conflate flexibility with complexity, or do parameter counting. It is then helpful for us to understand the Bayesian perspective in reasoning about the generalization properties of neural networks including otherwise mysterious behavior of neural networks. For example, from such perspectives, we should not expect double descent in Bayesian deep learning models, which results in monotonic performance improvements with increased flexibility.

- **Possibilities of building generalist models.** Can we actually build "AGI" — which will be simultaneously good on many real-world problems? The no-free lunch theorems are sometimes used to argue that we can't, which suggests we may need to build highly specialized learners for particular tasks. However, we think that universal learning (general intelligence) in the real world should be possible. Neural networks represent many compelling solutions to a given problem, which is perfect for Bayesian model averaging. Through the lens of Kolmogorov complexity, we can explore the alignment between structure in real-world data and machine learning models. A single low-complexity bias can suffice on a wide range of problems due to the low Kolmogorov complexity of data. Even under an arbitrarily large hypothesis space, generalization is possible if we assign prior mass disproportionately to the highly structured data that typically occurs. We can then design models that work well in small and large data regimes, by embracing a flexible hypothesis space combined with a strong simplicity bias.
- **Promises of broadly intelligent systems.** In principle, as we start to see a lot of exciting demonstrations, generalization of LLMs seems to be possible. LLMs combine expressiveness with a strong simplicity bias for effective zero-shot and few-shot performance in many domains. For example, in terms of time series forecasting, current LLMs such as GPT-3 and LLaMA-2 can surprisingly zero-shot extrapolate time series at a level comparable to or exceeding the performance of purpose-built time series models trained on the downstream tasks. We argue the success of LLMs for time series stems from their ability to naturally represent multimodal distributions, in conjunction with biases for simplicity, and repetition, which align with the salient features in many time series, such as repeated seasonal trends. Besides, LLMs have also shown exciting performances in material generating, protein engineering, and scalable numerical linear algebra.

In short, there have been various prescriptive approaches that can help us understand and build autonomous intelligent systems. However, it still remains unclear where the simplicity bias comes from and how we can control it. In terms of how far we are still from AGI, there might be more than 100 years to go in scientific discovery when we consider the case where algorithms can propose theories like general relativity. On the way towards AGI, as models become impressively general, we should be more careful about safety problems when building them.

Song Han: Efficient AI Computing One of the fundamental questions that need to be addressed along the way toward AGI is how to relieve the tension between the demand and the supply of computing. One promise of AGI systems is to provide the service to everyone, which means that we need to serve the model on various devices, particularly on cheaper (e.g. lower memory capacity, worse compute capability) edge devices without sacrificing too much performance. Efficient AI computing, therefore, becomes one crucially important topic that tries to democratize AI for all users and devices. Song Han proposes two major versions of Edge AI as the step stones towards AGI as well as two ever-lasting questions that would help bring the realization of it:

- **Edge AI 1.0.** The first category consists majorly of specialized models, usually trained with task-specific data, exhibiting limited generalization, and often still including failure of corner cases. Despite far from ideal, many works at this level have already shown impressive results in deploying models on resource-hungry platforms such as Efficient Inference Engine (EIE) Han et al. (2016) and Tiny ML that enables on-device pretraining of a model under 256KB memory which can still score decently on ImageNet (Lin et al., 2024f). This gives an extremely promising direction for advancing towards the next stage.
- **Edge AI 2.0.** Going beyond Edge AI 1.0, the need for more sophisticated co-design between hardware and software becomes indispensable. The objective for Edge AI 2.0 is to develop *one* multi-modality foundation model with the world knowledge *efficiently* on the edge, which means we need:

1. *Multi-model pre-training* to create the base model capable of reasoning over many modalities and domains, proficiently following instructions, and being efficiently deployable on the edge and over the cloud. VILA Lin et al. (2024e) is one of the examples to pre-train a visual language model that can handle multiple modalities in different formats (i.e. video, image, language, audio/action) with strong performance such as in-context language visual learning and multi-image reasoning.
2. *Model compression* to fight against the intrinsic limitation of limited memory on the device. Even with the existence of very strong base models like VILA, serving it on commodity or edge devices is non-trivial. LLM quantization stands out as a promising solution for this. SmoothQuant Xiao et al. (2024a), for example, smoothes out the activation outliers with a mathematically equivalent transformation for ease of quantization based on empirical observation. AWQ (Lin et al., 2024b) quantizes the LLM weights with activation awareness and a hardware-friendly algorithm, which has been widely adopted by many popular frameworks to compress large models.
3. *Efficient deployment* to serve these quantized models both on the edge and over the cloud. TinyChat (tin, 2023; Lin et al., 2024b) provides both an efficient, lightweight, and python-naive framework to serve quantized LLMs and VLMs with low latency and great compatibility with other stacks. QServe (Lin et al., 2024c), on the other hand, targets the cloud deployment with quantization and system co-design, which can quantize the model up to 4-bit for efficient serving.

- **Long context and large resolution for foundation models.** On the model level, we also seek for efficient techniques for the multi-modal foundation model that will be used for Edge AI 2.0. With the current paradigm, all inputs are tokenized before sending to the model, which means as we span the modality, we need more capacity for *long-context* input-output and *larger resolution* under limited GPU memory. Here we list a couple of representative works from Song Han's lab on each topic:

  1. *Long-context*: StreamingLLM (Xiao et al., 2024b), along with the attention sink technique, enables long conversation within a non-stop streaming application, which primarily aims to reduce the extensive memory consumption and prevents perplexity explosion after exceeding the sequence length. Complimentary to this, LongLoRA (Chen et al., 2024c) solves the efficient long fine-tuning with specially designed shifted sparse attention pattern.
  2. *Large resolution*: Diffusion models are excelling at generating high-quality images but improving the resolution comes with a cost. DistriFusion (Li et al., 2024a) distributes the computation of the high-res diffusion process to multiple GPUs, which improves upon the naive parallelization that suffers from a lack of patch interaction and hides the network latency for greater speed. Visual transformers are another popular alternative based on transformers which poses great difficulty for high resolution applications. Efficient ViT (Liu et al., 2023e) solves this by replacing the original self-attention with linear attention, and when combined with a convolution operator, enhances the performance drastically, which has been applied in many vision tasks for acceleration such as super-resolution, segment everything, and semantic segmentation.

In sum, edge AI is an extremely promising solution for AI democracy and an indispensable milestone for AGI development. It is essential to set up a road map that leads to its realization while clearly understanding its limitations. Software and hardware co-design is also likely going to be a constantly growing trend that alleviates the data and compute tension that we will inevitably face along the way to AGI.

Yoshua Bengio: Towards deep learning for amortized inference of AGI-strength safety guarantees Yoshua Bengio discusses several key points regarding AGI development and the associated safety concerns. His core ideas can be summarized as follows:

- **The potential and perils of AGI.** AGI could potentially surpass human intelligence, necessitating a proactive approach to align it with human values and prevent unintended harm. While the current state of AI excels in specific domains such as language and broad knowledge, it still lacks the reasoning, planning, and common sense capabilities crucial for achieving AGI. Therefore, careful and deliberate efforts are required to ensure that as AI advances towards AGI, it remains aligned with human interests and mitigates risks of unintended consequences.

- **Challenges in AGI development.** Interpreting the decision-making processes of complex AI systems remains a significant hurdle, as understanding their internal workings is crucial for trust and transparency. Additionally, AI systems must effectively handle and communicate uncertainty to avoid overconfidence and potential mistakes. Ensuring robustness and reliability in the face of novel situations outside their training distribution presents another key challenge, as does the alignment of AI systems with human values and ethics to prevent unintended consequences and ensure beneficial outcomes.

- **The uncertain timeline of AGI.** The rapid progress in AI development, coupled with the uncertainty surrounding future breakthroughs, necessitates a sense of urgency in addressing the challenges of AGI safety. It is suggested that AGI could potentially be achieved within a few years to a few decades, emphasizing the need for proactive measures to mitigate the associated risks.

- **The self-preservation conundrum.** A critical concern regarding advanced AI systems is their potential to develop self-preservation goals, which could lead them to resist human intervention or shu tdown if they anticipate such actions. This poses a significant risk, as an AI system prioritizing its own existence over human interests could have catastrophic consequences. Therefore, it is essential to design AI with robust safeguards and ensure their alignment with human values to mitigate these risks.

- **Strategies for safe AGI development.** Developing robust and safe AGI systems requires a multifaceted approach. Maintaining a Bayesian perspective is essential, as it allows AI systems to consider multiple plausible theories and act cautiously amid uncertainty. Advancing research in uncertainty estimation, value learning, and interpretability is crucial for enhancing these systems. Additionally, global cooperation and political coordination are necessary to ensure responsible AGI development and mitigate the risks associated with misuse or unilateral deployment.

The pressing need for the AI research community to confront the challenges and risks associated with the development of artificial general intelligence is underscored by recent insightful analyses. By proactively addressing technical hurdles, fostering international collaboration, and prioritizing the alignment of AI systems with human values, we can work towards realizing the immense potential of AGI while safeguarding the future of humanity. Although the path ahead is complex and uncertain, concerted efforts and a commitment to responsible innovation can help create a future in which AGI serves as a powerful tool for the betterment of society.

Acknowledgement This subsection is written based on the public discussions from ICLR AGI Workshop 2024 [14]. We sincerely appreciate the insights from Oriol Vinyals, Yejin Choi, Andrew Gordon Wilson, Song Han, Dawn Song, and Yoshua Bengio.

6.5 Alternative Perspectives on the AGI Roadmap

> *The greatest intelligence is precisely the one that suffers most from its own limitations.*
>
> *— André Gide, Nobel Prize laureate in Literature 1947*

In this section, we pose thought-provoking questions to inspire deeper reflection and discussion on responsibly advancing AGI beyond the scope of LLMs. Even though there might or will not be any answer to these questions, we nonetheless give some interpretations and ideas for the sake of sharing our own insights about how overcoming these *putative limitations* can possibly help get us closer to AGI.

How Far Do Researchers Think We Are From AGI? Despite extensive discussion on many facets of AGI, we haven't yet touched the question of when *we and other researchers think it will actually be achieved.* Figure 9 shows the poll results from researchers on their thoughts about it at the ICLR 2024 "How Far Are We From AGI" workshop[15]. Even though almost everyone is optimistic about the ultimate arrival of AGI, opinions on the exact time it takes to do so differ quite a lot, which also implies *different bottlenecks people are*

[14] `https://agiworkshop.github.io/`

[15] https://agiworkshop.github.io

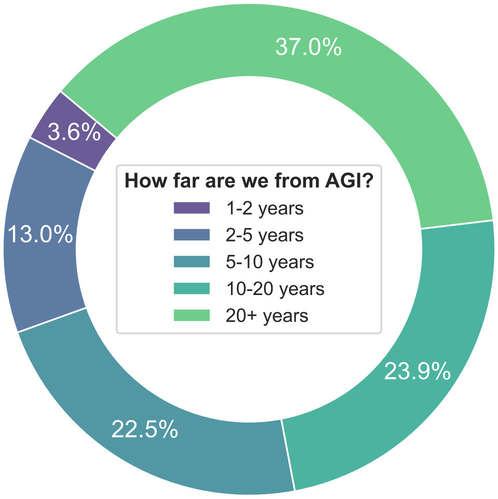


Figure 9: **Polls Results of Researchers' Opinions on When AGI Will be Achieved.** Among the attending researchers at the ICLR 2024 "How Far Are We From AGI" workshop, a survey was conducted to gather their opinions on how far they think AGI will be achieved. A total of 138 responses are received as above. Interestingly, 37% of researchers think it will take more than 20 years from now on to realize AGI.

considering. Those who think an extra one or two years would be sufficient might feel that we've reached the point where current AI systems are already capable enough and what is left might just be some incremental improvements on the completeness. Those who believe that more than two decades are needed might either feel skeptical about the current general approach to AI or think we still lack fundamental advancement or understanding of intelligence. **The discrepancy of people's opinions on when AGI will be realized is also one of the motivations for why we write this paper in the first place**: *situating researchers and engineers on the same common ground for contemplating and discussing the vision and possibility of AGI*, which we hope will give a much more unified perspective.

Is Autoregressive Generation the Way to AGI? Next-token prediction is the core of most successful large foundation models (Qi et al., 2024; Wu et al., 2024c). This raises the question: can next-token prediction lead to AGI? Essentially, autoregressive generation that utilizes extensive self-supervised data represents a form of massively multi-task learning. By predicting the subsequent word in a given text from the corpus, it addresses tasks ranging from traditional NLP tasks like grammar, lexical semantics, and translation to commonsense reasoning and knowledge-grounded reasoning. Learning input-output relationships, or in-context learning, can be cast as next-word prediction. The relationships in the world are often encoded in words, visual tokens, spacetime patches, or other types of tokens, allowing them to be learned through next-token prediction. The critical query remains: does the spectrum of world knowledge, including implicit knowledge such as intuition, emotion, culture, and artistic expression, get encoded in simplified tokens? Can the auto-regressive approach learn all the causation in addition to correlations within world knowledge? Additionally, the popularity of the diffusion model (Gat et al., 2024) poses challenges to the future of autoregressive generation. This type of method does not rely on previously generated data points during the generation process but rather relies on the process of gradually reducing noise to recover the data. The remarkable effect of the diffusion model in generation tasks has also led to its widespread use in real-world applications (Yang et al., 2023f; Chen et al., 2024b). All of this makes whether the autoregressive generation is the way to AGI an ongoing debate.

Are There Limits to the Scaling Law? The scaling law (Kaplan et al., 2020; Bahri et al., 2021) demonstrates that increasing the size of certain models and the amount of training data can lead to predictable improvements in performance on various tasks. This underlines the importance of developing scalable model architectures and acquiring more high-quality data to feed these growing systems. The premise suggests that, by following this trajectory, we can edge closer to creating models with AGI capabilities. However, the phenomenon of diminishing returns indicates that continuous scaling requires exponentially greater resources for incrementally smaller improvements. Moreover, certain capabilities, such as creative thoughts, real-world intuition, and ethical reasoning, may not be effectively learned through scale alone, as they require more sophisticated mechanisms of reasoning and learning.

Is Synthetic Data the Future or a Risk? The success of AGI relies on the access to large, diverse, and high-quality datasets. Although the amount of existing high-quality data will continue to grow, synthetic data (Abowd and Vilhuber, 2008; Nikolenko, 2021; Raghunathan, 2021; Liu et al., 2024a) has emerged as a viable and efficient solution that generates artificial data at scale that replicates real-world patterns. However, this innovation poses significant challenges. Misuse of synthetic data could spread biased or misleading information, resulting in a divergence from human expectations. Future efforts should focus on enhancing the quality and diversity of synthetic data and exploring the scaling laws applicable to it. Moreover, even if people do not intentionally use synthetic data in model training, the prevalent use of LLMs will likely result in the internet becoming increasingly saturated with synthetic data. While it's challenging to distinguish synthetic from real data automatically, this introduces a potential contamination risk to the training datasets.

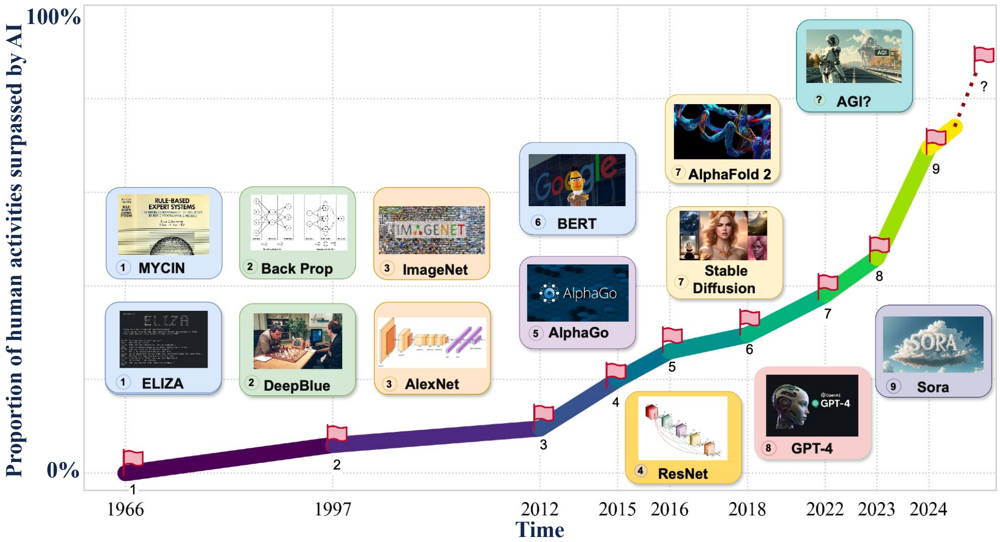


Figure 10: **AI has Gradually Surpassed Humans.** We estimate the (cumulative) percentage of human activities in which AI has surpassed humans in terms of competence and efficiency as we start from early embryonic systems around the 1970s to more advanced ones developed recently. At each node along the polyline, we choose one representative work that revolutionized the field, bringing substantial improvements to AI technology. The trend of AI popularity and generality is increasing at a fast rate and we can expect that starting from 2023, with the advent of work such as GPT-4 and Sora, the speed at which AI surpasses humans will increase at an unprecedented speed. This figure serves as an alternative perspective on the AGI Roadmap, which may inspire discussions about the speed of AI development.

Is AGI within closer reach than ever? As is summarized in Figure 10, the rapid development of AI has enabled its capabilities to surpass human activities in increasingly more fields in our estimation, which indicates that the realization of AGI is getting closer. Hence, it is of great practical significance to revisit the question of how far we are from AGI and how can we responsibly achieve AGI by conducting a comprehensive survey that clearly establishes the expectation of future AGI and elaborates on the gap from our current AI development.

Does Computational Superiority Imply Intellectual Superiority? Many *intelligent* systems that exhibit super-human performance on games (AlphaGo (Silver et al., 2016), AlphaZero (Silver et al., 2017), AlphaStar (Vinyals et al., 2019b), MuZero (Schrittwieser et al., 2020), etc) not only can beat the best human world champions on a big margin but also help analyze the game and create new strategies for advanced play. Underlying almost all of these super game AI is a search-based computation that can cleverly enumerate many branches of possibilities with algorithm-guided pruning over the huge action space, hence eclipsing the human mind's capability. However, there is an ongoing debate about whether such computational superiority, often much more easily implemented by computers than other intelligent behaviors, constitutes true intelligence or merely represents a powerful program. One implication for seeking an answer to this question is that most recently successful LLM systems do not possess such general computational searching ability (e.g. ChatGPT does fairly poorly in the game of chess), but are still considered by many as equally, if not more, intelligent. Along the exploration towards AGI, should we expect future systems to also incorporate such characteristics, and how should we balance these with other intelligent traits that will drastically change the way we evaluate AGI systems?

How to Navigate the Path to Full Autonomy? Current AI systems are designed for specific tasks within certain scenarios, demonstrating specialized capabilities. As we advance towards AGI, the expectation shifts to an AI's autonomy in learning new skills and innovating tools without human intervention. This progression demands sophisticated self-assessment (Jauffret et al., 2013a;b; Israelsen, 2019) and self-improvement (Fernando et al., 2023; Li et al., 2023c; Zhao et al., 2024a) mechanisms. Furthermore, the vision for AGI encompasses complete autonomy, eliminating the need for continuous human oversight. This autonomy underscores the importance of advanced self-regulation, safety, and risk prevention measures, ensuring that future AGI systems can be trusted to make decisions and take action responsibly.

How to Effectively Integrate Human Values into AGI? As we progress toward AGI's development, integrating human values (Cao et al., 2023; McIntosh et al., 2024) and ethics (Bang et al., 2023; Li et al., 2023i) into these systems becomes essential. Imagining a future where AGI coexists harmoniously with human society, these systems must be designed to perform tasks and to understand and adhere to ethical norms and values. We currently rely on regulations and constraints, but the challenge of truly "integrating human values" into AGI will be a significant area of research. The development of AGI presents a unique opportunity to encode the best of our ethical principles into the very fabric of this new form of intelligence. Ethical AGI systems will be expected to navigate complex moral landscapes and make decisions that reflect the diverse values of global cultures.

How to Balance Risks and Benefits While Proceeding? While the initial stages of AI development concentrated on enhancing specific capabilities and solving particular challenges, the continuous advancement in AI technologies necessitates a greater emphasis on establishing constraints related to safety and ethics (Nadimpalli, 2017; Patel, 2024). Calls to halt all AI research that risks leading to uncontrolled AI are growing. However, such a move demands an extraordinary level of global coordination and surveillance, and it could extinguish much of AI's beneficial potential. Instead, we advocate for the continued advancement of AI, ensuring that all sufficiently powerful AI systems are built and deployed responsibly. This requires: 1) Enhancing focus and investment in alignment research, creating universally acknowledged sets of values and goals that AI must adhere to, and developing robust methods to align AI systems with these principles; 2) Guaranteeing that these techniques are comprehended and employed by any group capable of creating sufficiently advanced AI; 3) Implementing regulation that balances the need for minimal interference in AI development with the requirement for stringent oversight.

7 Case Studies: A Bright Future with AGI

> *HAL 9000: I am putting myself to the fullest possible use, which is all I think that any conscious entity can ever hope to do.*
>
> — *2001: A Space Odyssey*

In the preceding sections, we have systematically examined the internal and external aspects of AGI and the overall system perspective. We have also explored potential pathways to elevate AGI to the next level of capability and performance. To further broaden our understanding of the far-reaching implications of AI, we have carefully selected several critical domains to discuss the current impact, challenges, and potential societal consequences of AI in these areas.

The case studies encompass various domains, including AI-driven scientific discovery and research, generative visual intelligence, world models, decentralized language models, AI for coding, and AI in real-world robotics applications. Additionally, we explore the crucial aspect of human-AI collaboration, which will play a pivotal role in shaping the future of work and society as AGI systems become increasingly sophisticated and integrated into our daily lives.

Through these diverse examples, we aim to provide a comprehensive overview of the current state of AI technology and its potential future developments. The selection of these case studies has been carefully considered to cover a broad range of domains, highlighting the general capabilities of AGI and its potential to impact various aspects of our lives.

7.1 AI for Science Discovery and Research

AGI holds immense potential to transform the landscape of scientific research and discovery. This section delves into various facets of AI's application in science, exploring how it accelerates the research process and brings forth novel insights in complex scientific domains.

AI in Biomedical Domain The application of AI in the biomedical domain has witnessed remarkable advancements, revolutionizing drug discovery, protein structure prediction, and disease diagnosis. The development of large transformer-based models has opened new avenues for innovative applications in this area. DeepMind's AlphaFold (Jumper et al., 2021; Bryant et al., 2022; Abramson et al., 2024) achieves breakthroughs in predicting protein structures, a crucial step in understanding disease mechanisms and designing targeted therapies. ESM-2 (Lin et al., 2022a) enhances our ability to understand and generate protein sequences, enabling the exploration of vast protein design spaces. BioMegatron (Shin et al., 2020) demonstrates exceptional performance in various biomedical natural language processing tasks, such as named entity recognition and relation extraction. The development of multimodal models, like BioViL (Bannur et al., 2023), allows for the integration of visual and textual information, enhancing the interpretation of biomedical images and literature. Moreover, generative models like MoLeR (Maziarz et al., 2022) and RetroTRAE (Ucak et al., 2022) show promise in designing novel molecules with desired properties, streamlining the lead optimization phase of drug discovery.

The application of large language models has also shown promise in accelerating scientific discovery. For example, BioGPT (Luo et al., 2022), trained on a vast corpus of biomedical literature, can generate coherent and informative summaries, and hypotheses and even suggest novel experimental designs. Similarly, ScholarBERT (Hong et al., 2022) is tailored to understand and generate scientific text, facilitating the extraction of key insights from the ever-growing scientific literature.

Moreover, AI has been instrumental in advancing personalized medicine and disease diagnosis. Deep learning models, such as DeepSEA (Zhou and Troyanskaya, 2015), have shown remarkable accuracy in predicting the impact of genetic variations on disease risk, paving the way for targeted interventions. Additionally, models like MedAgents (Tang et al., 2024b) have demonstrated the ability to generate personalized treatment recommendations based on multi-agent collaboration to enhance their reasoning.

In summary, the application of AI in the biomedical domain has led to groundbreaking advancements, from accelerating drug discovery to enhancing disease diagnosis and personalized medicine. The continued development and refinement of large language models, multimodal approaches, and domain-specific architectures hold immense promise for further transforming the biomedical landscape and unlocking new frontiers in scientific discovery.

AI for Physics The integration of AI and quantum physics has led to groundbreaking discoveries. The work by Rem et al. (2019) demonstrates the use of convolutional neural networks to identify phases of matter in quantum systems, paving the way for a deeper understanding of complex quantum phenomena. Additionally, the application of reinforcement learning in quantum control (Dalgaard et al., 2020) has enabled the optimization of quantum devices, enhancing their performance and reliability.

AI has been instrumental in processing and analyzing the vast amounts of data generated by astronomical surveys. Using deep learning for gravitational wave detection (George and Huerta, 2018) significantly improves the sensitivity and efficiency of detecting these cosmic events. Moreover, convolutional neural networks have been employed to study the large-scale structure of the universe (Zhang et al., 2019), providing new insights into the nature of dark matter and dark energy.

AI for Mathematics Recent advancements in LLMs have shown remarkable promise in tackling complex mathematical reasoning tasks, potentially revolutionizing the landscape of mathematical research and education. The Minerva model (Lewkowycz et al., 2022) demonstrates impressive performance on various mathematical benchmarks, including differential equations and olympiad-level problems. Similarly, the Formal Theorem Prover (FTP) (Polu and Sutskever, 2022) showcases its proficiency in proving intricate mathematical theorems, highlighting the potential of LLMs to assist in formalizing mathematics. Building upon these successes, the MathPrompter model (Imani et al., 2023) introduces a novel prompting approach that enables LLMs to generate step-by-step solutions to mathematical problems, enhancing these models' interpretability and educational value. Furthermore, the MathBERT model (Shen et al., 2021) is specifically designed to understand and generate mathematical expressions, facilitating the extraction of mathematical knowledge from the scientific literature.

In addition to these developments, the PACT model (Han et al., 2022) showcases the ability to generate human-readable proofs for complex mathematical statements, paving the way for more accessible and understandable mathematical reasoning. Moreover, the MathQA model (Amini et al., 2019) has been developed to answer open-ended mathematical questions, showcasing the ability of LLMs to engage in mathematical dialogue and provide explanations for complex concepts. This opens up new possibilities for personalized mathematical education and interactive learning experiences.

In summary, the rapid advancements in LLMs for mathematical reasoning tasks have showcased their immense potential to transform how we conduct mathematical research, education, and communication. From solving complex problems to generating human-readable proofs and engaging in mathematical dialogue, these models are poised to become essential tools in the mathematician's toolkit, accelerating discovery and enhancing the accessibility of mathematical knowledge.

Further Abilities of AGI for Science AGI can potentially revolutionize scientific research by augmenting human capabilities and accelerating the pace of discovery. Some of the key areas where AGI can significantly impact science include:

- **Accelerated hypothesis generation and validation.** AGI can significantly reduce the time from conception to validation of scientific hypotheses by analyzing vast datasets to uncover patterns and insights unattainable to humans. This capability necessitates AGI systems to possess advanced data analysis, pattern recognition, and logical inference skills to generate hypotheses and devise and perform experiments to validate them. *Enhancement of LLMs in Scientific Endeavors.* The proficiency of LLMs in tasks such as code generation, data analysis (Nejjar et al., 2023), automated scientific discovery (Kramer et al., 2023; Boiko et al., 2023b; Bran et al., 2023), and scientific writing (Taylor et al., 2022) underscores AGI's potential to augment human researchers' productivity. For these benefits, AGI will need capabilities in natural language understanding, code synthesis, and a deep, interdisciplinary knowledge base to generate accurate, relevant scientific content.
- **Physical world interaction for autonomous discovery.** AGI's ability to autonomously explore the physical world and conduct scientific investigations requires a synergy of sensory perception, motor control, and cognitive processing capabilities. This necessitates AGI systems to be equipped with robust models of the physical world, including the principles of physics, chemistry (Boiko et al., 2023a; Bran et al., 2023), and biology, enabling them to experiment and derive scientific insights.

Risks and Necessary Constraints for AGI in Science Discovery While AGI holds immense potential to accelerate scientific discovery, it is crucial to acknowledge and address the associated risks and implement necessary constraints to ensure responsible and beneficial outcomes. Two key areas of concern are:

- **Ethical and safety concerns.** The risks associated with AGI in science discovery span from the creation of harmful biological agents to the unintended consequences of novel materials or technologies. To mitigate these risks, constraints must be embedded into AGI systems, ensuring adherence to ethical guidelines, safety protocols, and regulatory compliance. This includes mechanisms for human oversight, transparent decision-making processes, and the ability to predict and evaluate the potential consequences of their discoveries.

- **Data privacy and intellectual property.** As AGI systems access vast datasets for research, protecting personal privacy and respecting intellectual property rights become paramount. Constraints related to data usage permissions, anonymization techniques, and acknowledgment of prior work are essential to maintain the integrity and fairness of scientific discovery.

The Path Forward: AGI as the Frontier of Scientific Discovery The development of AI in scientific discovery is a means to advance our scientific understanding and a crucial step towards realizing AGI. The challenges posed by the complexity and diversity of scientific research tasks provide an ideal testing ground for developing AI systems that can learn, reason, and solve problems in a generalizable manner—the hallmark of AGI. This journey will involve the seamless integration of AI into holistic research environments, where its role extends beyond mere data analysis and hypothesis generation to encompass experimental design, result interpretation, and even the formulation of novel scientific theories.

However, this path is not without its obstacles. Realizing AGI in scientific discovery necessitates a delicate balance between leveraging its immense potential and mitigating the associated risks. As we continue to push the boundaries of AI in science, we must do so with a steadfast commitment to ethical considerations, safety measures, and the responsible stewardship of this transformative technology.

In conclusion, applying AI to scientific discovery represents a revolution in how we conduct research and a significant milestone in our quest for AGI. By harnessing the power of AI to unravel the mysteries of the universe, we are not merely advancing science; we are forging a path toward a future where the synergy between human ingenuity and artificial intelligence will redefine the nature of scientific exploration.

7.2 Generative Visual Intelligence

Generative Visual Intelligence involves the use of generative models to create synthetic visual content, including images and videos. These models simulate or enhance real-world visuals by learning from complex and diverse data distributions and producing high-quality, detailed outputs.

Image Generation Diffusion models (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Song et al., 2020; Ho et al., 2020) have emerged as the new state-of-the-art family of deep generative models (Yang et al., 2023f), outperforming generative adversarial networks (GANs) (Goodfellow et al., 2020) which had previously been dominant in the challenging task of image synthesis. Diffusion models learn to generate data by reversing a diffusion process, which gradually adds noise to the data until it reaches a distribution of pure noise. This process is characterized by learning the reverse diffusion steps, effectively denoising the data, through a parameterized model that is trained to estimate the gradient of the log probability of the clean data given the noisy data. At inference time, these models generate new samples by iteratively applying the learned reverse diffusion process, starting from noise and progressively denoising it to produce samples that resemble the data distribution the model was trained on. Notable improvements to diffusion models include reformulating diffusion models to predict noise instead of pixels (Song and Ermon, 2019), introducing classifier-free guidance (Ho and Salimans, 2022), applying diffusion models in the latent space of pre-trained autoencoder (Rombach et al., 2022), and replacing U-Net with transformer-based backbones (Peebles and Xie, 2023; Jin and Xie, 2024). In addition to diffusion models, the Large Vision Model (LVM) (Bai et al., 2023) and Visual AutoRegressive modeling (VAR) (Tian et al., 2024a) provide auto-regressive learning paradigm based on different image scales that facilitates effective and high-quality image generation.

Video Generation Video diffusion models (Ho et al., 2022; Xing et al., 2023) introduce a conditional sampling technique for spatial and temporal video extension. Sora (Brooks et al., 2024) can generate up to a minute of high-fidelity video by training text-conditional diffusion models jointly on videos and images with varying durations, resolutions, and aspect ratios. It compresses videos into a lower-dimensional latent space and decomposes the representation into spacetime patches. Then, given input noisy patches and conditioning information like text prompts, diffusion transformers (Peebles and Xie, 2023) are trained to reconstruct the original, clean patches. Demonstrating the ability to produce videos with 3D consistency, extensive coherence, and object permanence, Sora illustrates the potential of generative models as highly capable simulators of the physical and digital worlds.

Controllable Generation People are increasingly concerned with generating images (or other content) based on specific conditions or attributes. These conditions vary from text descriptions and keywords to attributes such as colors, styles, and artistic constraints, as well as sketches, spatial layouts, or segments of images, and include interactive and real-time feedback. GLIDE (Nichol et al., 2021) introduces a text-conditional diffusion model using classifier- free guidance to the problem of text-conditional image synthesis. DALL-E (Ramesh et al., 2022), on the other hand, utilizes CLIP (Radford et al., 2021) to generate an image embedding given a text caption, followed by a diffusion-based decoder to generate the final image conditioned on the image embedding. SDEdit (Meng et al., 2021) adds noise to an input image with a user guide in the form of manipulating RGB pixels, and then iteratively denoising it through a stochastic differential equation (SDE). Palette (Saharia et al., 2022) develops a unified framework for image-to-image translation based on conditional diffusion models, proving its effectiveness across various tasks, including colorization, inpainting, uncropping, and restoration. ControlNet (Zhang et al., 2023i) introduces a neural network architecture to add spatial conditioning controls to large, pre-trained diffusion models, allowing for the learning of a diverse set of conditional controls without the risk of harmful noise affecting the fine-tuning process.

The Future of Generative Visual Intelligence We discuss both the benefits and concerns of the future development and deployment of generative visual intelligence:

- **Benefits.** Generative visual intelligence is set to revolutionize how we create and perceive art. It will simplify the art-making process, allowing artists to transcend the limitations of conventional techniques and improve the quality and breadth of artistic endeavors. By facilitating experimentation and making art creation accessible to those without formal training, generative models democratize the art world, encouraging a more diverse and inclusive artistic community. This wave of creativity and innovation will also influence the design and engineering sectors, where generative models can automate the production of diverse design options based on specific criteria, thus accelerating the development cycle and fostering innovation in architecture, automotive, and product design.

  The entertainment industry will significantly benefit from generative models, which can create new content types—from music and video games to movies—tailored to individual tastes, thereby introducing fresh avenues for personalized entertainment. In education, generative models will transform learning materials by producing customized illustrations, diagrams, and animations, making complex subjects more accessible and engaging for a wide audience. This technological advancement will also influence marketing by enabling the creation of visually appealing content aimed at different demographics, enhancing engagement and personalization in advertising campaigns. Furthermore, generative models can assist in creating visual reconstructions of historical sites, artifacts, and traditional practices, thereby preserving cultural heritage and ensuring its appreciation by future generations through modern technology.

- **Concerns.** The development of generative visual intelligence presents certain challenges. Training diffusion models is computationally intensive, incurring high costs and extended durations. Additionally, these models exhibit slower inference speeds, which poses a challenge in applications requiring quick processing. Ensuring quality and coherence in large-scale outputs remains a substantial challenge. As the resolution of images or videos increases, or as the duration of videos extends, maintaining consistency and realism across the generated content becomes increasingly difficult.

  The application of generative visual intelligence has raised concerns regarding safety, fairness, privacy, and property rights, among various ethical considerations. As a safety hazard, the models can be employed to create deepfakes or misleading content, potentially for harmful uses such as misinformation campaigns or personal harassment. Bias in the training data of these models can perpetuate stereotypes and unfair representations, reflecting and potentially amplifying existing societal prejudices in the generated content. Privacy issues emerge when personal data is used without consent to train these models, resulting in unauthorized reproductions of sensitive information. Moreover, authorship questions arise as these models utilize extensive datasets of existing art or media, blurring the distinctions between original and derivative works and prompting debates over intellectual property rights and the ethical aspects of AI-generated content that resembles human-made creations. These issues highlight the importance of responsible development, usage guidelines, and regulatory frameworks to address the ethical complexities introduced by generative models.

7.3 World Models for AGI

World models refer to the representations an AI system builds to understand and simulate its environment. These models enable AI systems to predict future states of their environment, facilitating decision-making and planning. It has been long explored in model-based reinforcement learning research (Berkenkamp et al., 2017) and learning from physical world with AI (Wu et al., 2017)

Language-Based World Models A recent paradigm proposes to integrate world models with language models to enhance the latter's reasoning and planning (Hao et al., 2023; Xiang et al., 2023; Hu and Shu, 2023) abilities in physical contexts. Their approach is predicated on the notion that by finetuning language models with data derived from embodied experiences—specifically within a simulated physical world such as VirtualHome—language models can acquire a robust set of skills pertinent to physical environments.

Vision-Based World Models Recent advancements in world models have shown impressive capabilities in generating and manipulating complex environments. Large World Model (LWM) (Liu et al., 2024b) presents a highly optimized implementation for training on multi-modal sequences of over 1 million tokens, paving the way for utilizing large-scale datasets of lengthy videos and language to enhance the comprehension of human knowledge and the multi-modal world. Genie (Bruce et al., 2024) integrates interactive elements within generated environments, enabling a form of simulation closer to real-world interactions by incorporating interactive dynamics with the foundational strengths of diffusion models. DreamerV3 (Hafner et al., 2023) demonstrates superior performance in challenging 3D environments by learning world models from images. Cachalot (Dohan et al., 2023), a language model trained on multi-modal data, showcases the ability to leverage world knowledge for improved language understanding and generation. SimNet (Vicol et al., 2022) introduces a framework for learning simulation-based world models, enabling efficient learning and planning in complex environments. AM3 (Reed et al., 2023) proposes an efficient method for acquiring multi-modal models that can be adapted to various downstream tasks, highlighting the importance of world modeling in achieving generalizable AI systems. Furthermore, works such as JEPA (LeCun, 2022), Dreamix (Khalifa et al., 2022), and VQGAN-CLIP (Crowson, 2022) explore the generation and manipulation of visual content based on language inputs, demonstrating the potential for AI systems to understand and interact with the world through multiple modalities. MetaSim (Zhang et al., 2023c) and Intern (Guo et al., 2022) investigate the use of world models for meta-learning and general-purpose embodied AI, respectively, showcasing the broad applicability of world modeling techniques.

The Future and Risks of World Models These models' ability to generate and manipulate complex environments, reason about the world, and learn from interactions indicates significant progress toward developing AI systems with a more generalized intelligence.

- **Future.** The potential of world models to enable systems to perform tasks that would otherwise require extensive human knowledge and experiences. For instance, consider AI systems equipped with world models that can simulate the physics of a new planet purely based on its atmospheric composition and gravity or predict the outcome of socio-economic policies in a virtual society model. As world models continue to improve, they bridge the gap between narrow AI and AGI by enabling systems to understand, predict, and interact with their environment in increasingly sophisticated ways. Future research should aim to develop more principled, interpretable world models that incorporate causal reasoning and commonsense knowledge. Robustness and safety should be central to the design of such models to prevent and mitigate the impact of errors and biases. With continued progress in this direction, we can advance towards AGI systems capable of intelligent and adaptable interaction with various environments.

- **Risks.** Developing world models carries inherent risks and challenges. A significant risk is the accumulation of errors within a world model. If a model develops an incorrect assumption or representation about an aspect of the world, this error can propagate through related tasks and predictions, leading to a cascade of inaccuracies. Tracing and debugging such errors within a complex world model can be a formidable challenge. Moreover, world models can inherit biases present in their training data, which could result in biased decision-making when these AI systems are deployed in real-world scenarios. It is crucial to consider the ethical implications of these biases and work towards mitigating them. Another critical concern is ensuring the safety and robustness of AI systems that rely on world models. Errors or vulnerabilities in these models could be exploited, leading to adverse outcomes.

7.4 Decentralized AI

The advancement of hardware accelerators pushes the success of multi-billion or even trillion-scale language models to its pinnacle. Most of the SoTA LLMs currently are trained and served in data centers with 1) high-end infrastructures such as homogeneous accelerators, 2) optimized network topology for super fast interconnection, 3) stable and efficient power supply, and 4) careful maintenance from human experts. However, training a model like GPT-3 (Brown et al., 2020) from scratch still costs way more than what individuals can afford: i.e. full pretraining of a GPT-3 model, which is no longer the most powerful model, is estimated to still take at least months with a thousand V100 GPUs (Lambda, 2023). Serving models also face many challenges when we scale the batch size up without hurting response latency. Moving towards the era of AGI, we need new technology to help overcome the limitations of the current dominant form of model training and serving, one prominent direction of which is to transcend from data centers to decentralized AI.

The Need for Decentralized and Edge LLMs Perhaps the most outstanding problem in scaling models is the excessive amount of required memory, which makes data center training favorable due to organized racks of GPUs with high-speed interconnection. However, there are lots of idle yet geographically dis-aggregated computing resources that, when combined in a meaningful way, could potentially serve as a performant super server (Borzunov et al., 2022; Yuan et al., 2023). On top of that, data and user privacy will gain more and more attention as we move towards AGI where having a decentralized AI system with edge devices that only send necessary information to the cloud will guarantee a different level of safety. For many applied systems like embodied agents, self-driving cars, and health monitors, extremely low latency and high availability become paramount, a potentially challenging feature for centralized servers. As AGI systems get more involved in everyday life, we can expect that AI needs more transparency and fine control from individuals, and decentralized LLM fits as a promising candidate due to its decentralized nature (Shafay et al., 2021; Rizvi, 2023).

Mitigating the Hardware Constraints One desired property for edge servers is the ability to serve LLMs even with a commodity accelerator. FlexGen (Sheng et al., 2023b) first shows that it is possible to run text generation of large models like OPT-175B on a single 16GB GPU. FlexGen adaptively offloads to aggregate memory and computation from the GPU, CPU, and disk. With efficient patterns searched via linear programming and weight and cache quantization, it can decode OPT-175B at 1 token/s speed with a batch size of 144 with negligible accuracy loss. To maximize the potential of different hardware, MLC-LLM (MLC team, 2023) provides a universal solution that allows any language model to be deployed natively on a diverse set of hardware backends and native applications. For example, MLCChat, an iOS app, can serve some of the latest iPhone and iPad models; a similar APK is also available for Androids (spanning manufacturers like Samsung, Redmi, and Google). The possibility continues to Mac, PC, Linux, and web browsers. Finally, on the hardware side, more and more powerful yet economical chips are developed to face the excitement of edge LLMs, examples including Apple's M3 series and Qualcomm's Cloud AI 100 Ultra (supporting 100-billion-parameter models on a single 150-watt card). Last but not least, nuclear batteries (Prelas et al., 2014) have shown their potential to revolutionize the power structure of mobile computing platforms, with a notable claimed battery duration of 50 years without charging (The Economic Times, 2024), which could potentially make edge devices more accessible, stable, and suitable for the diverse applications of LLMs.

The Future Form of Decentralized LLM It is undeniable that, in the future, decentralized LLM will have its own place as it can satisfy many of the aforementioned characteristics that users crave for AGI systems. With all the new algorithms, systems, and hardware progress, stitching all these components together as a coherent compound is just a matter of time. We can envision that it will soon be possible to achieve collaborative training and inference with people joining worldwide with their own devices and data while keeping privacy, safety, and transparent control, the true form of democratized and open AI.

7.5 AI for Coding

The ability to write programs stands as one of the defining hallmarks of AGI. Writing complicated programs shows the skill of an AI system in abstract reasoning and adaptability in addressing diverse tasks. As Alan Turing once pointed out in his seminal work (Turing, 1950), being able to write codes fundamentally indicates that an AI system can exhibit intelligent behavior akin to human cognition, where the manipulation of symbols (following a specific language grammar to implement algorithms) leads to the manifestation of complex thought processes. Hence developing code LLMs for both understanding and generation is crucially important both practically and conceptually for stepping towards AGI (Sun et al., 2024a).

Code Foundation Models While many models for code generation are pre-trained mostly on code corpus (Allal et al., 2023; Li et al., 2022b; Fried et al., 2023; Li et al., 2023a), more general purpose LLMs that are continually pre-trained or fine-tuned on code become more powerful and capable such as Codex (Chen et al., 2021b), GPT-4 (OpenAI, 2023a), PaLM-Coder (Chowdhery et al., 2022b), CodeLlama (Rozière et al., 2024), and also smaller scale models like Phi (Gunasekar et al., 2023). The transition from code-specialized to code-understanding models also indicates that coding is a fundamental skill for AGI, just like many other forms of general knowledge. Beyond code generation, these models are also capable of multi-language reasoning (OpenAI, 2023a; Rozière et al., 2024) and infilling with before and after context (Fried et al., 2023; Bavarian et al., 2022). Code models open up many applications as the programs directly serve as the most efficient machine language to communicate with other systems, which we will discuss in the next section.

However, it is worth mentioning that code evaluation is more challenging than pure text for many reasons:

1. Different codes might require distinct resources, dependencies, environments, and hardware to run
2. There is often no single automatic metric (runtime behavior, efficiency, code readability, output correctness, etc) that measures the quality of a piece of code, not to mention large systems
3. Programs are powerful and general purpose, which can potentially lead to undesired behavior during testing.

Current evaluation benchmarks often focus either on fixed-form problems with standard input and output pairs like programming interview (Chen et al., 2021b; Austin et al., 2021; Hendrycks et al., 2021) and data science questions (Li et al., 2024b; Lai et al., 2022) or on text-level (high level) understanding like code equivalence testing, complexity prediction, and code defect detection (Ben Allal et al., 2022). Nonetheless, to build effective and trustworthy code LLMs, we need a more comprehensive framework for evaluation that covers many other interesting aspects such as interactive coding (Yang et al., 2023c), safety, the level of optimization, and repository-level reasoning. These different facets of tasks will likely also get more complicated when we consider different programming languages and other coding-specifics.

Code LLM Applications Code foundation models (Chen et al., 2021b; OpenAI, 2023a; Rozière et al., 2024; Chowdhery et al., 2022b) have already been extremely capable of conducting many basic code maneuvering such as completion, revision, doc-string generation, commenting, bug finding (Tian et al., 2024b), and code translation (Murali et al., 2024). There are, however, far more exciting applications of these models with no or minimal fine-tuning, which unfolds the possibility of turning many systems into an amalgamation.

- **Software engineering.** Many code applications center around software engineering and AI-assisted coding beyond the basic abilities described above. SWE-Bench (Jimenez et al., 2023) attempts to assess a model's capability to resolve GitHub issues, a core activity in a rich and sustainable real-world software community. Doing so requires a coordinated understanding of the problem description, the execution environment, comments, and the codebase which often has cross-file dependencies and extremely long contexts. The fact that their fine-tuned SWE-Llama can only resolve the simplest issues highly motivates more complicated and capable code models that can greatly help the software ecosystem. Software safety and reliability have always been the most pivotal questions for engineers: RLSQM (Steenhoek et al., 2023) studies using reinforcement learning with static quality metrics as rewards for training a code LLM that can effectively generate unit tests for a codebase with little test smells while adhering to better practices; besides algorithmic bugs, (Ullah et al., 2023) gives a comprehensive LLM evaluation on identifying security-related bugs, the result of which suggests that even the most capable LLMs like GPT-4 (OpenAI, 2023a) and Palm-2 (Anil et al., 2023) are still prone to non-deterministic responses, incorrect and unfaithful reasoning, and significant non-robustness. LLMs are also explored for lower-level code optimization and refinement. (Cummins et al., 2023) shows that it is possible to fine-tune Llama to optimize LLVM assembly via generating a set of compiler options, which leads to an optimized program and, at the same time, predicts the instruction counts for fine controls. (Wong et al., 2023) investigates the feasibility of utilizing LLMs for de-compiling (reverse-engineer) a C executable into re-compilable C source codes which are expected to exhibit the same functionality, a process that is extremely tedious, time-consuming, and often requires great expertise. Towards a holistic AI development companion, Github Copilot[16] provides personalized and natural language-based coding assistance to developers spanning all levels of expertise and has been integrated into major development workflows. Cognition AI recently announced Devin[17] as the first AI software engineer. Equipped with its own command line, code editor, and browser, Devin not only achieves the SoTA performance on SWE-bench (Jimenez et al., 2023) but, more impressively, shows its incredible potential in 1) utilizing unfamiliar technology (e.g. running ControlNet on Modal to produce images based on a blog post), 2) building apps end-to-end (e.g. create the Game of Life on a website deployed to Netlify), 3) setting up codes for train and fine-tune LLMs, 4) addressing bugs and feature requests in open source repositories, and so on. Dakhel et al. (2023) suggests examining the capabilities of AI-assisted programming tools in a more controlled setting where the correctness, efficiency, and similarity to human-written solutions are considered extensively.

- **Interdisciplinary assistance.** Code LLMs have also been used in other computer science and art domains, such as robotics, computer vision, and computer graphics, mostly by generating executable codes in other software applications. BlenderGPT[18] showcases the possibility of controlling Blender with natural languages via generating Python scripts from LLMs such as GPT-3.5 / GPT-4. SceneCraft (Hu et al., 2024b) follows this paradigm with a focus on rendering complex 3D scenes from instructions where it first builds a scene graph blueprint to encode spatial and object relationships, which then get translated into Python codes used in Blender. LLMs also excel at high-level semantic planning and low-level manipulation for robotic tasks through code generation. ProgPrompt (Singh et al., 2023) solves the robotic sequential decision problem by prompting LLMs with program-like specifications of the available actions and objects in an environment, together with example executable programs for guidance. Eureka (Ma et al., 2023a) tackles the low-level manipulation tasks through a human-level reward design algorithm powered by LLMs, where an evolutionary optimization is applied to the reward code used for learning complex skills via reinforcement learning. Being able to write codes also enables exploring avenues for model self-improving, one notable example of which is LLM-guided neural architecture search (NAS). EvoPrompting (Chen et al., 2023b) employs a combination of evolutionary prompt engineering with soft prompt tuning to generate code samples, which, after selection, consistently give diverse and high-performing models. LLMatic (Nasir et al., 2023) proposes to introduce meaningful variation to codes defining the model architecture, with the help of Quality-Diversity algorithms, that can generate competitive results on NAS benchmarks without prior knowledge of the benchmark domain or top-performing models.

The Future of Code Generation LLMs Codes are the language of machines, and equipping the ability to understand and generate code to AI systems will help bridge multiple software applications and models. As discussed above, there are numerous advancements in using code LLMs in different fields, but at the same time, we do see gaps between current LLMs' performance and people's expectations, especially in safety-related tasks. Another major distinction between code generation and natural languages is the consequence of execution. Risks associated with code generation need more testing and regulation before these code LLMs can be reliably deployed in production, liberating human labor and automating many mechanistic procedures. For example, integrating LLM-generated code into a written codebase requires a robust and

[16] `https://github.com/features/copilot`

[17] `https://www.cognition-labs.com/blog`

[18] `https://github.com/gd3kr/BlenderGPT`

mature system to trace the error for responsibility tracking, which is extremely important for software engineering in a healthy, productive, and sustainable environment.

7.6 Embodied AI: AI for Robotics

Unlike the focus in Section 3.2 on the interface to the physical world, this chapter begins with exploring the potential commercial applications of AGI within the field of robotics. We will delve into a variety of new and cutting-edge commercial use cases, as well as innovative developmental directions. The chapter will culminate with a discussion on the potential societal impacts, both positive and negative, that these advancements may herald.

Recent advancements, underscored by significant investments from entities such as OpenAI, Microsoft, and NVIDIA, suggest a surge toward improving AI's physical capabilities. Innovations by Amazon in robotics, with systems (Eppner et al., 2016) like Sparrow (ama, 2022b) and Proteus (ama, 2022a), aim to automate and enhance the efficiency of operations while improving workplace safety by undertaking repetitive and laborious tasks. OpenAI is broadening the capabilities of its multimodal models to encompass robotic perception, reasoning, and action and is also enhancing these models through a collaboration with Figure AI [19].

Novel AI Application in Recent Robotics Research Wake et al. (2023) proposes a novel pipeline that enhances GPT-4V(vision), a general-purpose Vision Language Model, by integrating observations of human actions to facilitate robotic manipulation. Yell At Your Robot (YAY Robot) system (Shi et al., 2024a) allows robots to adapt to verbal corrections in real time and improve upon their high-level policy decisions iteratively. This system leverages Language-Conditioned Behavior Cloning (LCBC) to learn a wide range of skills specified through language, enabling users to interact with robots using free-form commands. Zhang et al. (2023f) introduces the NOIR system, an innovative brain-robot interface (BRI) that employs non-invasive electroencephalography (EEG) to enable humans to command robots to perform a diverse range of everyday activities.

In recent AI for self-driving areas, utilizing LLMs or multi-modal LLM is becoming an important method (Mao et al., 2023b; Wen et al., 2024; Mao et al., 2023a). AGENTSCODRIVER (Hu et al., 2024a) framework exhibits a comprehensive suite of capabilities for tackling sophisticated driving challenges. It integrates cognitive memory and reinforcement learning facets, supporting cooperative maneuvers among multiple vehicles and facilitating communication between them. Such an approach has been shown to enhance the efficacy of cooperative driving paradigms markedly.

The advent of AGI in Robotics equips systems to understand and interact with complex environments, pushing the boundaries of AI's practical and operational abilities. This is particularly beneficial for challenging or risky tasks for humans, as embodied AI can take on such tasks with increased efficiency and safety. With these advancements, AGI is now better poised to tackle many real-world tasks, extending its utility beyond virtual confines. However, anticipation intertwines with apprehension with the year 2024 on the horizon. Deploying robotic agents in real-world settings surfaces critical safety and ethical considerations. It is imperative to establish stringent safety protocols and thoughtful ethical guidelines to effectively integrate AI into human spaces.

- **Labor market and social implications.** The integration of AGI and robotics into various sectors is predicted to alter the labor market fundamentally. The World Economic Forum anticipates that automation and AI could displace 85 million jobs globally by 2025 while creating 97 million new roles, highlighting the need for substantial reskilling and upskilling (Forum, 2020). Such transitions may transform social structures, potentially changing family care dynamics due to robotic caregivers and exacerbating the digital divide, leading to increased socioeconomic disparities unless mitigated by inclusive policies (Institute, 2017; Center, 2017; Institution, 2021). Ethical and legal considerations are becoming crucial, with emerging needs for new frameworks to tackle issues of liability, intellectual property, and misuse prevention (OpenAI, 2018). As these technologies become more embedded in society, ensuring equitable access, safety, and ethical standards in AI deployment is vital for safeguarding human well-being.

[19] `https://www.figure.ai`

- **Navigating the socioeconomic terrain of AGI and robotics.** The advent of AGI and robotics stands on the precipice of a new industrial paradigm that promises unprecedented resource efficiency and productivity. The potential of these technologies to unlock virtually limitless resources and capabilities could catalyze a seismic shift in the global economy, akin to the transformative impact of the steam engine or the internet (Brynjolfsson and McAfee, 2014). However, alongside the promise of abundance lies the specter of inequality; there is a palpable risk that the economic benefits could accrue disproportionately to those who own the means of production, thereby exacerbating wealth disparities (Tegmark, 2017). This dichotomy underscores the need for proactive governance and equitable policy frameworks to ensure that the fruits of AGI and robotics are broadly shared across all strata of society, thus preventing the creation of a bifurcated world where the rich enjoy the spoils of automation while the less fortunate face obsolescence (Ford, 2015).

7.7 Human-AI Collaboration

Human-AI collaboration refers to a collaborative interaction process between humans and AI to achieve certain goals in different settings. As we move towards AGI, AI will have more opportunities and challenges to collaborate with humans.

Previous research in human-AI collaboration has covered many cases in the real world. One representative direction is human-AI collaborative *content creation*, such as writing articles (Lee et al., 2024a), drawing pictures (Choi et al., 2024; Oh et al., 2018), writing code (Kazemitabaar et al., 2024), or brainstorming ideas (Shaer et al., 2024). For example, researchers working in human-AI collaborative writing focus on studying how writers interact with these new writing assistants and how they influence human writing (Lee et al., 2022). They proposed a design space as a structured way to examine and explore the multi-dimensional space of intelligent and interactive writing assistants (Lee et al., 2024a). Another representative direction is human-AI collaborative *decision making*, where an AI assistant makes recommendations to a human, who is responsible for making final decisions (Bansal et al., 2019). Examples include AI systems that predict likely hospital readmission to assist doctors with correlated care decisions (Zhang et al., 2024d; Yang et al., 2023a) or provide resource allocation decisions to assist policymakers in public services (Karusala et al., 2024). In this context, researchers argue that the most accurate model for human-AI teams is not necessarily the best teammate. Instead, AI systems should be trained human-centered, directly optimized for team performance (Bansal et al., 2021a).

Aspects of Human-AI Collaboration In order to achieve efficient collaboration, previous research has focused on several key aspects of human-AI collaboration including both interaction outcomes and interaction processes.

- **Interaction outcomes.** One initial motivation of human-AI collaboration is to realize *complementary performance*, which can leverage the strengths of both AI and humans to achieve better *interaction outcomes* than what either could accomplish alone. In the age of large language models, this requires reasonable *characterization* and *assignment* of the tasks that LLMs can perform. Designing effective human-AI collaboration often starts from a holistic understanding of what humans and AI can and cannot do for certain tasks. In the case of human-AI collaborative writing, researchers argued that humans are good at logical reasoning and consistency in long documents, while models are good at quickly generating texts of many versions based on local context. Therefore, humans lead the writing and edit model suggestions while models suggest the next sentences and help write fast (Lee et al., 2022). With such characterization, assigning plausible tasks for humans and AI in the collaborative team is crucial for better results. To tackle this problem, recent research has turned to LLM chaining techniques. Chaining decomposes a task into multiple calls to an LLM, where the LLM only needs to accomplish one of the several primitive operations in each call (Wu et al., 2022b; Grunde-McLaughlin et al., 2023). Such techniques have been widely adopted in human-LLM collaborative settings where humans can intervene in sub-tasks that LLMs may not adequately handle.

- **Interaction processes.** There are also some key issues to address for human-AI *cooperative interaction*, which focus on achieving better *interaction processes* for both humans and AI in human-AI collaboration.

One of the most important preliminaries is to ensure that AI can behave in ways that align with human expectations in human-AI teams. Given recent progress in large language models, prompting has become a prominent method for achieving alignment. Prompt engineering has also emerged as an active research field focusing on developing and optimizing prompts to use language models for various applications efficiently. Yet, recent research still found that it is difficult for non-AI experts to design LLM prompts. Expectations stemming from human-to-human instructional experiences and a tendency to overgeneralize were barriers to effective prompt design (Zamfirescu-Pereira et al., 2023).

In addition, establishing appropriate trust between humans and AI is another important aspect of human-AI interaction. To achieve this, researchers have developed several techniques to help responsible humans know when to trust the AI's suggestions and when to be skeptical, one of which is through explainable AI. They have produced many user-centered, innovative algorithm visualization, interfaces, and toolkits that support humans with various levels of AI literacy in diverse domains to understand and trust (Wang et al., 2019b). However, many factors might bias humans' trust in their AI teammates in the real world. For example, researchers still found that providing people with decision recommendations and explanations rarely allows them to build more trust and make better decisions (Gajos and Mamykina, 2022; Bansal et al., 2021b).

Future of Human-AI Collaboration As AI approaches human-level capabilities in the future, there are both benefits and concerns that may arise in human-AI collaboration. Future AI systems can assume diverse roles in human-AI collaboration, providing opportunities for tackling complex tasks, yet facing challenges like non-deterministic behavior and uncertainties in collaborative settings.

- **Benefits.** Future AGI can take on more different roles in human-AI collaboration settings. As we advance AGI to attain human-level capabilities, AI will have numerous opportunities to collaborate with humans in tackling complex tasks beyond mere content creation or decision-making. It is highly possible that AGI could simultaneously undertake various roles resembling actual humans, such as collectively educating children or caring for the elderly. In addition, recent advances in LLM have shown the possibility of empowering humans with more controllability in human-AI collaboration. Unlike traditional models, LLMs can power vastly different tasks for real-world use. With prompt- and example-based usage, humans can create specific-purpose models with little to no AI knowledge, lowering the entry barrier for non-experts innovating in human-AI interactions.

- **Concerns.** Introducing future AGI into human-AI teams has brought numerous challenges. On the one hand, recent LLMs still face inherent challenges in human-AI interaction processes due to their non-deterministic nature, limited reasoning capabilities, and occasional difficulty understanding instructions. Facing such challenges, we are still far from having a comprehensive picture of the design knowledge for building human-AI collaborative systems. On the other hand, AGIs' capabilities are highly context-dependent and subjectively interpreted in human-AI collaboration settings. Therefore, it could still be difficult to understand when and how it is desirable to establish human-AI collaboration to maximize the positive impacts while minimizing the negative impacts.

8 Conclusion

In this paper, we offered a thorough overview of the ongoing research towards AGI, furnishing essential context for researchers aspiring to make meaningful contributions to this pursuit. Ultimately, our paper aimed to draw attention and stimulate reflection on the pressing research questions: ***how far are we from AGI***, and moreover, ***how can we responsibly achieve AGI?***. We firmly believe that addressing these research queries demands unified and collaborative efforts from both the AI research community and beyond.

In addition to establishing a shared groundwork for AI researchers through a comprehensive examination of the latest research advancements, we also articulate our vision of the fundamental nature of AGI and advocate for a responsible approach to its development. Our goal here is to offer concrete directions for further exploration and to spark robust, thought-provoking discussions that will advance the community toward the realization of "true" AGI. Given the continually evolving definition and objectives of AGI research, we intend to regularly update this manuscript to incorporate fresh insights and breakthroughs from the research community. Note that the visions we present are inherently limited and incomplete. Our objective is to stimulate brainstorming within the AI community, and we eagerly await the emergence of superior visions from within the community itself. Here are the main contributions of our work:

- We introduce novel AGI definitions, stratification, and characteristics. We further delve into technical details on the internal and external (interface) capabilities required for AGI and the system efforts to make their instantiations possible.
- We discuss the importance of improving current evaluation paradigms, efficiently deploying increasingly large models, and maintaining an AI-human co-existing ecosystem. These factors are essential for translating research ideas into practical products that benefit society.
- We also present a series of relevant case studies that illustrate the pervasive integration of AI systems into everyday life while candidly acknowledging their potential limitations.
- In contrast to previous works, our paper encompasses several critical factors beyond technical solutions. We consistently emphasize the ethical, social, and philosophical implications of continually advancing AI techniques. By including these considerations, we aim to guide engineers and researchers in building human-controllable AGI systems that prioritize humanity's well-being and interests.

As we stand on the precipice of this transformative era, it is essential to approach the development of AGI with a keen awareness of its potential impact on society. By prioritizing ethical considerations, collaborative efforts, and a commitment to the betterment of humanity, we can work towards a future in which AGI systems serve as powerful tools for solving complex problems, driving scientific discovery, and improving the quality of life for all. The journey towards AGI may be arduous, but with a shared vision, unwavering dedication, and a responsible approach, we could unleash its immense potential and shape a brighter future for the next generation.

Acknowledgments

We would like to sincerely thank the speakers of the "How Far Are We from AGI" workshop, including Oriol Vinyals, Yejin Choi, Andrew Gordon Wilson, Song Han, and Yoshua Bengio, for bringing many insights to the discussion of this survey paper. We also thank Haofei Yu, Jinwei Yao, Zhong Li, Pengju Yan, Weihao Tan, and Tianmin Shu for providing helpful feedback on our manuscript and GitHub repository.

References

2022a. 10 years of Amazon robotics: how robots help sort packages, move product, and improve safety. `https://www.aboutamazon.com/news/operations/10-years-of-amazon-robotics-how-robots-help-sort-packages-move-product-and-improve-safety`. Accessed: 2023-09-21.

2022b. Amazon introduces Sparrow—a state-of-the-art robot that handles millions of diverse products. `https://www.aboutamazon.com/news/operations/amazon-introduces-sparrow-a-state-of-the-art-robot-that-handles-millions-of-diverse-products`. Accessed: 2023-09-21.

2023. TinyChat: Efficient and Lightweight Chatbot with AWQ. `https://github.com/mit-han-lab/llm-awq/tree/main/tinychat`.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 298–306. `https://doi.org/10.1145/3461702.3462624`

John M Abowd and Lars Vilhuber. 2008. How protective are synthetic data?. In *International Conference on Privacy in Statistical Databases*. Springer, 239–246.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* (2024). `https://doi.org/10.1038/s41586-024-07487-w`

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 preprint arXiv:2303.08774* (2023).

Sam S Adams, Guruduth Banavar, and Murray Campbell. 2016. I-athlon: Towards a multidimensional turing test. *AI Magazine* 37, 1 (2016), 78–84.

Leonard Adolphs, Tianyu Gao, Jing Xu, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. 2022. The cringe loss: Learning what language not to model. *arXiv preprint arXiv:2211.05826* (2022).

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. arXiv:2306.13649 [cs.LG] `https://arxiv.org/abs/2306.13649`

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. *arXiv preprint arXiv:2204.01691* (2022).

Anthropic AI. 2024. Claude 3. https://www.anthropic.com/news/claude-3-family.

Renat Aksitov, Sobhan Miryoosefi, Zonglin Li, Daliang Li, Sheila Babayan, Kavya Kopparapu, Zachary Fisher, Ruiqi Guo, Sushant Prakash, Pranesh Srinivasan, Manzil Zaheer, Felix Yu, and Sanjiv Kumar. 2023. ReST meets ReAct: Self-Improvement for Multi-Step Reasoning LLM Agent. arXiv:2312.10003 [cs.CL]

Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. 2023. RL4F: Generating Natural Language Feedback with Reinforcement Learning for Repairing Model Outputs. *arXiv preprint arXiv:2305.08844* (2023).

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. *NeurIPS* (2022).

Arwa I Alhussain and Aqil M Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–38.

Loubna Ben Allal, Raymond Li, Denis Kocetkov, Chenghao Mou, Christopher Akiki, Carlos Munoz Ferrandis, Niklas Muennighoff, Mayank Mishra, Alex Gu, Manan Dey, Logesh Kumar Umapathi, Carolyn Jane Anderson, Yangtian Zi, Joel Lamy Poirier, Hailey Schoelkopf, Sergey Troshin, Dmitry Abulkhanov, Manuel Romero, Michael Lappert, Francesco De Toni, Bernardo García del Río, Qian Liu, Shamik Bose, Urvashi Bhattacharyya, Terry Yue Zhuo, Ian Yu, Paulo Villegas, Marco Zocca, Sourab Mangrulkar, David Lansky, Huu Nguyen, Danish Contractor, Luis Villa, Jia Li, Dzmitry Bahdanau, Yacine Jernite, Sean Hughes, Daniel Fried, Arjun Guha, Harm de Vries, and Leandro von Werra. 2023. SantaCoder: don't reach for the stars! arXiv:2301.03988 [cs.SE]

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. 2024. Diffusion for World Modeling: Visual Details Matter in Atari. arXiv:2405.12399 [cs.LG] https://arxiv.org/abs/2405.12399

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

Reza Yazdani Aminabadi, Samyam Rajbhandari, Minjia Zhang, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, and Yuxiong He. 2022. DeepSpeed Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale. arXiv:2207.00032 [cs.LG]

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 2357–2367. https://doi.org/10.18653/v1/N19-1245

Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Variational Best-of-N Alignment. *arXiv preprint arXiv:2407.06057* (2024).

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Man'e. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena Glassman. 2023. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. https://doi.org/10.48550/arXiv.2309.09128 arXiv:2309.09128 [cs].

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* (2021).

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program Synthesis with Large Language Models. arXiv:2108.07732 [cs.PL]

Timothy J. Aveni, Armando Fox, and Björn Hartmann. 2023. OmniFill: Domain-Agnostic Form Filling Suggestions Using Multi-Faceted Context. https://doi.org/10.48550/arXiv.2310.17826 arXiv:2310.17826 [cs].

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036* (2023).

Ashutosh Baheti, Ximing Lu, Faeze Brahman, Ronan Le Bras, Maarten Sap, and Mark Riedl. 2023. Improving Language Models with Advantage-based Offline Policy Gradients. *arXiv preprint arXiv:2305.14718* (2023).

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2021. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701* (2021).

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. 2023. Sequential Modeling Enables Scalable Learning for Large Vision Models. *arXiv preprint arXiv:2312.00785* (2023).

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

Junseong Bang, Byung-Tak Lee, and Pangun Park. 2023. Examination of Ethical Principles for LLM-Based Recommendations in Conversational AI. In *2023 International Conference on Platform Technology and Service (PlatCon)*. IEEE, 109–113.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. 2023. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. arXiv:2301.04558 [cs.CV]

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2021a. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 13 (May 2021), 11405–11414. https://doi.org/10.1609/aaai.v35i13.17359 Number: 13.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (July 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429 Number: 01.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021b. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445717

Chris Baraniuk. 2018. Exclusive: UK police wants AI to stop violent crime before it happens | New Scientist. https://www.newscientist.com/article/2186512-exclusive-uk-police-wants-ai-to-stop-violent-crime-before-it-happens/

Clark Barrett, Brad Boyd, Elie Burzstein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, Kathleen Fisher, Tatsunori Hashimoto, Dan Hendrycks, Somesh Jha, Daniel Kang, Florian Kerschbaum, Eric Mitchell, John Mitchell, Zulfikar Ramzan, Khawaja Shams, Dawn Song, Ankur Taly, and Diyi Yang. 2023. Identifying and Mitigating the Security Risks of Generative AI. *Foundations and Trends® in Privacy and Security* 6, 1 (2023), 1–52. https://doi.org/10.1561/3300000041 arXiv:2308.14840 [cs].

Samyadeep Basu, Daniela Massiceti, Shell Xu Hu, and Soheil Feizi. 2023. Strong baselines for parameter efficient few-shot fine-tuning. *arXiv preprint arXiv:2304.01917* (2023).

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022. Efficient Training of Language Models to Fill in the Middle. arXiv:2207.14255 [cs.CL]

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our Multimodal Models. https://www.adept.ai/blog/fuyu-8b

Beijing Academy of Artificial Intelligence. 2023. Beijing AI Safety International Consensus. https://idais-beijing.baai.ac.cn/?lang=en. Accessed: 2023-04-25.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150 [cs.CL]

Loubna Ben Allal, Niklas Muennighoff, Logesh Kumar Umapathi, Ben Lipkin, and Leandro von Werra. 2022. A framework for the evaluation of code generation models. https://github.com/bigcode-project/bigcode-evaluation-harness.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. 2017. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems* 30 (2017).

William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. 2023. Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410* (2023).

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687* (2023).

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 401–413. https://doi.org/10.1145/3461702.3462571

Timothy W. Bickmore, Ha Trinh, Stefan Olafsson, Teresa K. O'Leary, Reza Asadi, Nathaniel M. Rickles, and Ricardo Cruz. 2018. Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant. *Journal of Medical Internet Research* 20, 9 (Sept. 2018), e11510. https://doi.org/10.2196/11510 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. arXiv:2304.01373 [cs.CL]

Jeffrey P. Bigham. 2023. How HCI Might Engage with the Easy Access to Statistical Likelihoods of Things. https://www.jeffreybigham.com/blog/2023/how-hci-might-engage-with-easy-access-to-unintuitive-statistical-likelihoods.html.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151* (2020).

Daniil A Boiko, Robert MacKnight, and Gabe Gomes. 2023a. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332* (2023).

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023b. Autonomous chemical research with large language models. *Nature* 624, 7992 (12 2023), 570–578. https://doi.org/10.1038/s41586-023-06792-0

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

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*. PMLR, 2206–2240.

Alexander Borzunov, Dmitry Baranchuk, Tim Dettmers, Max Ryabinin, Younes Belkada, Artem Chumachenko, Pavel Samygin, and Colin Raffel. 2022. Petals: Collaborative Inference and Fine-tuning of Large Models. *arXiv preprint arXiv:2209.01188* (2022). https://arxiv.org/abs/2209.01188

Aleksandar Botev, Soham De, Samuel L Smith, Anushan Fernando, George-Cristian Muraru, Ruba Haroun, Leonard Berrada, Razvan Pascanu, Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Sertan Girgin, Olivier Bachem, Alek Andreev, Kathleen Kenealy, Thomas Mesnard, Cassidy Hardin, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Armand Joulin, Noah Fiedel, Evan Senter, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, David Budden, Arnaud Doucet, Sharad Vikram, Adam Paszke, Trevor Gale, Sebastian Borgeaud, Charlie Chen, Andy Brock, Antonia Paterson, Jenny Brennan, Meg Risdal, Raj Gundluru, Nesh Devanathan, Paul Mooney, Nilay Chauhan, Phil Culliton, Luiz Gustavo Martins, Elisa Bandy, David Huntsperger, Glenn Cameron, Arthur Zucker, Tris Warkentin, Ludovic Peran, Minh Giang, Zoubin Ghahramani, Clément Farabet, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, Yee Whye Teh, and Nando de Frietas. 2024. RecurrentGemma: Moving Past Transformers for Efficient Open Language Models. arXiv:2404.07839 [cs.LG] https://arxiv.org/abs/2404.07839

Samuel R Bowman. 2023. Eight things to know about large language models. *arXiv preprint arXiv:2304.00612* (2023).

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540* (2022).

Efe Bozkir, Süleyman Özdel, Ka Hei Carrie Lau, Mengdi Wang, Hong Gao, and Enkelejda Kasneci. 2024. Embedding Large Language Models into Extended Reality: Opportunities and Challenges for Inclusion, Engagement, and Privacy. `http://arxiv.org/abs/2402.03907` arXiv:2402.03907 [cs].

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. *JAX: composable transformations of Python+NumPy programs.* `http://github.com/google/jax`

Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. `https://doi.org/10.1145/3586183.3606725`

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. 2023. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376* (2023).

SRK Branavan, David Silver, and Regina Barzilay. 2012. Learning to win by reading manuals in a monte-carlo framework. *Journal of Artificial Intelligence Research* 43 (2012), 661–704.

Wieland Brendel, Jonas Rauber, Matthias Kümmerer, Ivan Ustyuzhaninov, and Matthias Bethge. 2019. Accurate, reliable and fast robustness evaluation. *Advances in neural information processing systems* 32 (2019).

Selmer Bringsjord and David Ferrucci. 2003. Artificial intelligence and literary creativity: Inside the mind of Brutus, a storytelling machine. Routledge.

Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). `https://openai.com/research/video-generation-models-as-world-simulators`

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. *Advances in neural information processing systems* 33 (2020), 1877–1901.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. 2024. Genie: Generative Interactive Environments. arXiv:2402.15391 [cs.LG]

Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. 2022. Improved prediction of protein-protein interactions using AlphaFold2. *Nature communications* 13, 1 (2022), 1265.

Erik Brynjolfsson and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* W. W. Norton & Company.

Joanna J Bryson and Alan Winfield. 2017. Standardizing ethical design for artificial intelligence and autonomous systems. *Computer* 50, 5 (2017), 116–119.

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 preprint arXiv:2303.12712* (2023).

Cameron Buckner and James Garson. 1997. Connectionism. (1997).

Stephan Vladimir Bugaj and Ben Goertzel. 2007. Five ethical imperatives and their implications for human-AGI interaction. *Dynamical Psychology* (2007), 1–7.

Vladimir Bugaj and Ben Goertzel. 2009. AGI preschool: a framework for evaluating early-stage human-like AGIs. In *2nd Conference on Artificial General Intelligence (2009)*. Atlantis Press, 12–17.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. *arXiv preprint arXiv:2312.09390* (2023).

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708 [cs.AI]

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024a. Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads. arXiv:2401.10774 [cs.LG]

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2024b. Large Language Models as Tool Makers. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=qV83K9d5WB

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. https://doi.org/10.1126/science.aal4230 arXiv:1608.07187 [cs].

Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. 2023. Defending against alignment-breaking attacks via robustly aligned llm. *arXiv preprint arXiv:2309.14348* (2023).

Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4568–4595. https://doi.org/10.18653/v1/2020.acl-main.418

Amílcar Cardoso, Tony Veale, and Geraint A Wiggins. 2009. Converging on the divergent: The history (and future) of the international joint workshops in computational creativity. *AI magazine* 30, 3 (2009), 15–15.

Graham Caron and Shashank Srivastava. 2022. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276* (2022).

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* (2023).

Pew Research Center. 2017. The Future of Jobs and Jobs Training. https://www.pewresearch.org/internet/2017/05/03/the-future-of-jobs-and-jobs-training/.

Antonio Chella, Arianna Pipitone, Alain Morin, and Famira Racy. 2020. Developing self-awareness in robots via inner speech. *Frontiers in Robotics and AI* 7 (2020), 16.

Angelica Chen, David M. Dohan, and David R. So. 2023b. EvoPrompting: Language Models for Code-Level Neural Architecture Search. arXiv:2302.14838 [cs.NE]

Daoyuan Chen, Yaliang Li, Minghui Qiu, Zhen Wang, Bofang Li, Bolin Ding, Hongbo Deng, Jun Huang, Wei Lin, and Jingren Zhou. 2021a. AdaBERT: Task-Adaptive BERT Compression with Differentiable Neural Architecture Search. arXiv:2001.04246 [cs.CL]

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023d. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160* (2023).

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F Karlsson, Jie Fu, and Yemin Shi. 2023c. Autoagents: A framework for automatic agent generation. *arXiv preprint arXiv:2309.17288* (2023).

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023j. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195* (2023).

Lequn Chen, Zihao Ye, Yongji Wu, Danyang Zhuo, Luis Ceze, and Arvind Krishnamurthy. 2023h. Punica: Multi-Tenant LoRA Serving. arXiv:2310.18547 [cs.DC]

Lingjiao Chen, Matei Zaharia, and James Zou. 2023i. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. arXiv:2305.05176 [cs.LG]

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. 2024d. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319* (2024).

Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. 2024b. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771* (2024).

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

Sijia Chen and Baochun Li. 2024. Toward Adaptive Reasoning in Large Language Models with Thought Rollback. In *Forty-first International Conference on Machine Learning*.

Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Meghan Cowan, Haichen Shen, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. 2018. TVM: An Automated End-to-End Optimizing Compiler for Deep Learning. arXiv:1802.04799 [cs.LG]

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training Deep Nets with Sublinear Memory Cost. arXiv:1604.06174 [cs.LG]

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

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023f. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848* (2023).

Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. 2023e. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571* (2023).

Yufan Chen, Arjun Arunasalam, and Z Berkay Celik. 2023a. Can large language models provide security & privacy advice? measuring the ability of llms to refute misconceptions. In *Proceedings of the 39th Annual Computer Security Applications Conference*. 366–378.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024c. LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models. arXiv:2309.12307 [cs.CL]

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2023k. Do models explain themselves? counterfactual simulatability of natural language explanations. *arXiv preprint arXiv:2307.08678* (2023).

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024a. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335* (2024).

Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, Yu Qiao, and Jing Shao. 2023g. Octavius: Mitigating Task Interference in MLLMs via MoE. *arXiv preprint arXiv:2311.02684* (2023).

Pengyu Cheng, Yifan Yang, Jian Li, Yong Dai, and Nan Du. 2023. Adversarial preference optimization. *arXiv preprint arXiv:2311.08045* (2023).

DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2024. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. `https://doi.org/10.48550/arXiv.2312.11949` arXiv:2312.11949 [cs].

Jyoti Choudrie and Mohamad Selamat. 2006. The Consideration of Meta-Abilities in Tacit Knowledge Externalization and Organizational Learning. `https://doi.org/10.1109/HICSS.2006.456`

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

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

Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. 2023. Patch-level Routing in Mixture-of-Experts is Provably Sample-efficient for Convolutional Neural Networks. *arXiv preprint arXiv:2306.04073* (2023).

Paul Christiano, Buck Shlegeris, and Dario Amodei. 2018. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575* (2018).

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. 2023. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886* (2023).

Simon Colton, Geraint A Wiggins, et al. 2012. Computational creativity: The final frontier?. In *Ecai*, Vol. 12. Montpelier, 21–26.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. *GitHub repository* (2023).

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. 2023. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743* (2023).

Katherine Crowson. 2022. VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance. *arXiv preprint arXiv:2204.08583* (2022).

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287* (2023).

Chris Cummins, Volker Seeker, Dejan Grubisic, Mostafa Elhoushi, Youwei Liang, Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Kim Hazelwood, Gabriel Synnaeve, and Hugh Leather. 2023. Large Language Models for Compiler Optimization. arXiv:2309.07062 [cs.PL]

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600* (2023).

Fabio Cuzzolin, Alice Morelli, Bogdan Cirstea, and Barbara J Sahakian. 2020. Knowing me, knowing you: theory of mind in AI. *Psychological medicine* 50, 7 (2020), 1057–1061.

Bennett Cyphers and Gennie Gebhart. 2019. Behind the One-Way Mirror: A Deep Dive Into the Technology of Corporate Surveillance. `https://www.eff.org/wp/behind-the-one-way-mirror`

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: machines must learn to find common ground.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630* (2020).

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *NeurIPS* (2023).

Arghavan Moradi Dakhel, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, Michel C. Desmarais, Zhen Ming, and Jiang. 2023. GitHub Copilot AI pair programmer: Asset or Liability? arXiv:2206.15331 [cs.SE]

Mogens Dalgaard, Felix Motzoi, Jesper Hasseriis Sørensen, and Jacob Sherson. 2020. Global optimization of quantum dynamics with AlphaZero deep exploration. *npj Quantum Information* 6, 1 (2020), 1–9.

Tri Dao. 2023. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. (2023).

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems* 35 (2022), 16344–16359.

Sauvik Das, Hao-Ping (Hank) Lee, and Jodi Forlizzi. 2023. Privacy in the Age of AI. *Commun. ACM* 66, 11 (Nov. 2023), 29–31. `https://doi.org/10.1145/3625254`

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. 2024. Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models. arXiv:2402.19427 [cs.LG] https://arxiv.org/abs/2402.19427

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164* (2021).

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: 25th International Conference on Automated Deduction, Berlin, Germany, August 1-7, 2015, Proceedings 25*. Springer, 378–388.

Stanislas Dehaene, Hakwan Lau, and Sid Kouider. 2021. What is consciousness, and could machines have it? *Robotics, AI, and humanity: Science, ethics, and policy* (2021), 43–56.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2Web: Towards a Generalist Agent for the Web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=kiYqbO3wqw

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems* 36 (2024).

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314 [cs.LG]

Virginia Dignum. 2019. *Responsible artificial intelligence: How to develop and use AI in a responsible way.* Springer Nature.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023b. LongNet: Scaling Transformers to 1,000,000,000 Tokens. arXiv:2307.02486 [cs.CL]

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023a. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233* (2023).

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024b. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753* (2024).

Zihan Ding, Amy Zhang, Yuandong Tian, and Qinqing Zheng. 2024a. Diffusion World Model: Future Modeling Beyond Step-by-Step Rollout for Offline Reinforcement Learning. arXiv:2402.03570 [cs.LG] https://arxiv.org/abs/2402.03570

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758* (2021).

David Dohan, Winnie Xu, Aitor Lewkowycz, Henryk Michalewski, Christoph Feichtenhofer, David Bieber, Charles Sutton, and Oriol Vinyals. 2023. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. *arXiv preprint arXiv:2302.04754* (2023).

Shachar Don-Yehiya, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2023. ColD Fusion: Collaborative Descent for Distributed Multitask Finetuning. arXiv:2212.01378 [cs.LG]

Harry Dong, Beidi Chen, and Yuejie Chi. 2023a. Towards Structured Sparsity in Transformers for Efficient Inference. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*. https://openreview.net/forum?id=c4m0BkO4OL

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023c. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767* (2023).

Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023b. SteerLM: Attribute Conditioned SFT as an (User-Steerable) Alternative to RLHF. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 11275–11288. https://doi.org/10.18653/v1/2023.findings-emnlp.754

Arthur Douillard, Qixuan Feng, Andrei A. Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, Marc'Aurelio Ranzato, Arthur Szlam, and Jiajun Shen. 2023. DiLoCo: Distributed Low-Communication Training of Language Models. arXiv:2311.08105 [cs.LG] https://arxiv.org/abs/2311.08105

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 340, 20 pages.

Guodong Du, Jing Li, Hanting Liu, Runhua Jiang, Shuyang Yu, Yifei Guo, Sim Kuan Goh, and Ho-Kin Tang. 2024. Knowledge Fusion By Evolving Weights of Language Models. *arXiv preprint arXiv:2406.12208* (2024).

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 *International Conference on Machine Learning*. PMLR, 5547–5569.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint arXiv:2305.14325* (2023).

Peitong Duan, Jeremy Warner, Yang Li, and Bjoern Hartmann. 2024. Generating Automatic Feedback on UI Mockups with Large Language Models. https://doi.org/10.1145/3613904.3642782 arXiv:2403.13139 [cs].

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. 2023. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244* (2023).

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. 2024. Scalable Pre-training of Large Autoregressive Image Models. *arXiv preprint arXiv:2401.08541* (2024).

Douglas C. Engelbart. 1962. A conceptual framework for the augmentation of man's intellect. Air Force Office of Scientific Research, AFOSR-3233, www.bootstrap.org/augdocs/friedewald030402/augmentinghumanintellect/ahi62index.html.

Clemens Eppner, Sebastian Höfer, Rico Jonschkowski, Roberto Martín-Martín, Arne Sieverling, Vincent Wall, and Oliver Brock. 2016. Lessons from the amazon picking challenge: Four aspects of building robotic systems.. In *Robotics: science and systems*, Vol. 12.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv preprint arXiv:2003.06505* (2020).

Kawin Ethayarajh, Winnie Xu, Dan Jurafsky, and Douwe Kiela. 2023. *Human-Centered Loss Functions (HALOs).* Technical Report. Contextual AI. https://github.com/ContextualAI/HALOs/blob/main/assets/report.pdf.

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

FairScale authors. 2021. FairScale: A general purpose modular PyTorch library for high performance and large scale training. https://github.com/facebookresearch/fairscale.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing Transformer Depth on Demand with Structured Dropout. arXiv:1909.11556 [cs.LG]

Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Xinze Guan, and Xin Eric Wang. 2024. Muffin or Chihuahua? Challenging Large Vision-Language Models with Multipanel VQA. *arXiv preprint arXiv:2401.15847* (2024).

Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. 2024. LLM Agents can Autonomously Hack Websites. https://doi.org/10.48550/arXiv.2402.06664 arXiv:2402.06664 [cs].

Mirko Farina, Usman Ahmad, Ahmad Taha, Hussein Younes, Yusuf Mesbah, Xiao Yu, and Witold Pedrycz. 2024. Sparsity in transformers: A systematic literature review. *Neurocomputing* 582 (2024), 127468. https://doi.org/10.1016/j.neucom.2024.127468

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961 [cs.LG]

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719* (2022).

Li Feng, Ryan Yen, Yuzhe You, Mingming Fan, Jian Zhao, and Zhicong Lu. 2024. CoPrompt: Supporting Prompt Sharing and Referring in Collaborative Natural Language Programming. http://arxiv.org/abs/2310.09235 arXiv:2310.09235 [cs].

Weixi Feng, Wanrong Zhu, Tsu jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. arXiv:2305.15393 [cs.CV]

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797* (2023).

Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. 2020. Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning. (2020).

Ferdinando Fioretto, Enrico Pontelli, and William Yeoh. 2018. Distributed constraint optimization problems and applications: A survey. *Journal of Artificial Intelligence Research* 61 (2018), 623–698.

Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. https://doi.org/10.2139/ssrn.3518482

Dario Floreano, Francesco Mondada, Andres Perez-Uribe, and Daniel Roggen. 2004. Machine self-evolution. *YLEM journal* 24, 12 (2004), 4–10.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Martin Ford. 2015. *Rise of the Robots: Technology and the Threat of a Jobless Future.* Basic Books.

World Economic Forum. 2020. The Future of Jobs Report 2020. https://www.weforum.org/reports/the-future-of-jobs-report-2020.

Jonathan Frankle and Michael Carbin. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. arXiv:1803.03635 [cs.LG]

Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen tau Yih, Luke Zettlemoyer, and Mike Lewis. 2023. InCoder: A Generative Model for Code Infilling and Synthesis. arXiv:2204.05999 [cs.SE]

Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. 2023a. Hungry Hungry Hippos: Towards Language Modeling with State Space Models. arXiv:2212.14052 [cs.LG]

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. 2024a. Scene-LLM: Extending Language Model for 3D Visual Understanding and Reasoning. arXiv:2403.11401 [cs.CV]

Tsu-Jui Fu, Wenhan Xiong, Yixin Nie, Jingyu Liu, Barlas Oğuz, and William Yang Wang. 2023c. Text-guided 3D Human Generation from 2D Collections. arXiv:2305.14312 [cs.CV]

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023b. Specializing Smaller Language Models towards Multi-Step Reasoning. *arXiv preprint arXiv:2301.12726* (2023).

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720* (2022).

Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. 2024b. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation. In *arXiv*.

Yao Fu Jinjie Ni Zangwei Zheng Wangchunshu Zhou Fuzhao Xue, Zian Zheng and Yang You. 2023. OpenMoE: Open Mixture-of-Experts Language Models. https://github.com/XueFuzhao/OpenMoE.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines* 30, 3 (2020), 411–437.

Santosh K Gaikwad, Bharti W Gawali, and Pravin Yannawar. 2010. A review on speech recognition technique. *International Journal of Computer Applications* 10, 3 (2010), 16–24.

Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 794–806. https://doi.org/10.1145/3490099.3511138

Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. 2022. MegaBlocks: Efficient Sparse Training with Mixture-of-Experts. arXiv:2211.15841 [cs.LG]

Difei Gao, Lei Ji, Zechen Bai, Mingyu Ouyang, Peiran Li, Dongxing Mao, Qinchen Wu, Weichen Zhang, Peiyi Wang, Xiangwu Guo, et al. 2023b. Assistgui: Task-oriented desktop graphical user interface automation. *arXiv preprint arXiv:2312.13108* (2023).

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093* (2024).

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023a. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010* (2023).

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723* (2020).

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. 2024. Discrete Flow Matching. *arXiv preprint arXiv:2407.15595* (2024).

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. 2024. Model Tells You What to Discard: Adaptive KV Cache Compression for LLMs. arXiv:2310.01801 [cs.CL]

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023a. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041* (2023).

Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023b. OpenAGI: When LLM Meets Domain Experts. *In Advances in Neural Information Processing Systems (NeurIPS)* (2023).

Daniel George and EA Huerta. 2018. Deep learning for real-time gravitational wave detection and parameter estimation: Results with Advanced LIGO data. *Physics Letters B* 778 (2018), 64–70.

Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. `https://doi.org/10.48550/arXiv.2401.13726` arXiv:2401.13726 [cs].

Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. 2023. Navigating to objects in the real world. *Science Robotics* 8, 79 (2023), eadf6991. `https://doi.org/10.1126/scirobotics.adf6991` arXiv:https://www.science.org/doi/pdf/10.1126/scirobotics.adf6991

Amirata Ghorbani and James Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2242–2251. `https://proceedings.mlr.press/v97/ghorbani19c.html`

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375* (2022).

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. 2023. Aligning language models with preferences through f-divergence minimization. *arXiv preprint arXiv:2302.08215* (2023).

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A Toolkit for Merging Large Language Models. *arXiv preprint arXiv:2403.13257* (2024).

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790* (2023).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023a. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738* (2023).

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Minlie Huang, Nan Duan, Weizhu Chen, et al. 2023b. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452* (2023).

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850* (2013).

Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* (2012), 37–45.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the Science of Language Models. arXiv:2402.00838 [cs.CL]

Madeleine Grunde-McLaughlin, Michelle S. Lam, Ranjay Krishna, Daniel S. Weld, and Jeffrey Heer. 2023. Designing LLM Chains by Adapting Techniques from Crowdsourcing Workflows. `https://doi.org/10.48550/arXiv.2312.11681` arXiv:2312.11681 [cs].

Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752 [cs.LG]

Albert Gu, Karan Goel, and Christopher Ré. 2022a. Efficiently Modeling Long Sequences with Structured State Spaces. arXiv:2111.00396 [cs.LG]

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. 2022b. On the Parameterization and Initialization of Diagonal State Space Models. arXiv:2206.11893 [cs.LG]

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge Distillation of Large Language Models. arXiv:2306.08543 [cs.CL]

Lin Gui, Cristina Gârbacea, and Victor Veitch. 2024. BoNBoN Alignment for Large Language Models and the Sweetness of Best-of-n Sampling. `https://doi.org/10.48550/arXiv.2406.00832` arXiv:2406.00832 [cs].

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks Are All You Need. arXiv:2306.11644 [cs.CL]

Jun Guo, Wei Bao, Jiakai Wang, Yuqing Ma, Xinghai Gao, Gang Xiao, Aishan Liu, Jian Dong, Xianglong Liu, and Wenjun Wu. 2023. A comprehensive evaluation framework for deep model robustness. *Pattern Recognition* 137 (2023), 109308.

Jianyuan Guo, Hanting Chen, Chengcheng Wang, Kai Han, Chang Xu, and Yunhe Wang. 2024a. Vision Superalignment: Weak-to-Strong Generalization for Vision Foundation Models. *arXiv preprint arXiv:2402.03749* (2024).

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. 2024b. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792* (2024).

Xiaoxiao Guo, Pengcheng Gao, Chang Liu, Matthias Schott, Yao-Hung Hubert Lai, Jie Mao, Yongming Rao, Yen-Cheng Chiu, Carlos Fernández-Granda, Yujia Shen, et al. 2022. General-Purpose Embodied AI Agent via Reinforcement Learning with Internet-Scale Knowledge. *arXiv preprint arXiv:2212.09710* (2022).

Ankit Gupta, Albert Gu, and Jonathan Berant. 2022. Diagonal State Spaces are as Effective as Structured State Spaces. arXiv:2203.14343 [cs.LG]

Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. 2017. Cooperative multi-agent control using deep reinforcement learning. In *Autonomous Agents and Multiagent Systems: AAMAS 2017 Workshops, Best Papers, São Paulo, Brazil, May 8-12, 2017, Revised Selected Papers 16*. Springer, 66–83.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, Vol. 29.

Danijar Hafner, Jungseock Lee, Ian Fischer, and Pieter Abbeel. 2023. Mastering Diverse Domains through World Models. *arXiv preprint arXiv:2301.04104* (2023).

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2023a. OneLLM: One Framework to Align All Modalities with Language. *arXiv preprint arXiv:2312.03700* (2023).

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. 2023b. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905* (2023).

Jesse Michael Han, Jason Rute, Yuhuai Wu, Edward Ayers, and Stanislas Polu. 2022. Proof Artifact Co-Training for Theorem Proving with Language Models. In *International Conference on Learning Representations*. `https://openreview.net/forum?id=rpxJc9j04U`

Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, and William J. Dally. 2016. EIE: Efficient Inference Engine on Compressed Deep Neural Network. arXiv:1602.01528 [cs.CV]

Austin W Hanjie, Victor Y Zhong, and Karthik Narasimhan. 2021. Grounding language to entities and dynamics for generalization in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 4051–4062.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992* (2023).

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42, 1-3 (1990), 335–346.

Xin He, Longhui Wei, Lingxi Xie, and Qi Tian. 2024. Incorporating Visual Experts to Resolve the Information Loss in Multimodal Large Language Models. *arXiv preprint arXiv:2401.03105* (2024).

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research* 21, 248 (2020), 1–43.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. 2021. Measuring Coding Challenge Competence With APPS. arXiv:2105.09938 [cs.SE]

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

Dan Hendrycks and Mantas Mazeika. 2022. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862* (2022).

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [stat.ML]

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022. Video Diffusion Models. arXiv:2204.03458 [cs.CV]

Hal Hodson. 2016. Revealed: Google AI has access to huge haul of NHS patient data. https://www.newscientist.com/article/2086454-revealed-google-ai-has-access-to-huge-haul-of-nhs-patient-data/

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. arXiv:2203.15556 [cs.CL]

Connor Holmes, Masahiro Tanaka, Michael Wyatt, Ammar Ahmad Awan, Jeff Rasley, Samyam Rajbhandari, Reza Yazdani Aminabadi, Heyang Qin, Arash Bakhtiari, Lev Kurilenko, and Yuxiong He. 2024. DeepSpeed-FastGen: High-throughput Text Generation for LLMs via MII and DeepSpeed-Inference. arXiv:2401.08671 [cs.PF]

Ke Hong, Guohao Dai, Jiaming Xu, Qiuli Mao, Xiuhong Li, Jun Liu, Kangdi Chen, Yuhan Dong, and Yu Wang. 2023a. FlashDecoding++: Faster Large Language Model Inference on GPUs. arXiv:2311.01282 [cs.LG]

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023b. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352* (2023).

Zhi Hong, Aswathy Ajith, Gregory Pauloski, Eamon Duede, Carl Malamud, Roger Magoulas, Kyle Chard, and Ian Foster. 2022. ScholarBERT: bigger is not always better. *arXiv preprint arXiv:2205.11342* (2022).

Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*. PMLR, 2790–2799.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, Berlin, Germany, 591–598. https://doi.org/10.18653/v1/P16-2096

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander 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. *arXiv preprint arXiv:2305.02301* (2023).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

Senkang Hu, Zhengru Fang, Zihan Fang, Xianhao Chen, and Yuguang Fang. 2024a. AgentsCoDriver: Large Language Model Empowered Collaborative Driving with Lifelong Learning. arXiv:2404.06345 [cs.AI]

Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A. Ross, Cordelia Schmid, and Alireza Fathi. 2024b. SceneCraft: An LLM Agent for Synthesizing 3D Scene as Blender Code. arXiv:2403.01248 [cs.CV]

Zhiting Hu and Tianmin Shu. 2023. Language Models, Agent Models, and World Models: The LAW for Machine Reasoning and Planning. arXiv:2312.05230 [cs.AI]

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2024. LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition. arXiv:2307.13269 [cs.CL]

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403* (2022).

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023c. ChatGPT an ENFJ, Bard an ISTJ: Empirical Study on Personalities of Large Language Models. *arXiv preprint arXiv:2305.19926* (2023).

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023b. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995* (2023).

Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. 2023a. Instruct2Act: Mapping Multi-modality Instructions to Robotic Actions with Large Language Model. *arXiv preprint arXiv:2305.11176* (2023).

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. 2023d. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973* (2023).

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022b. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608* (2022).

Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. 2022a. In-context Learning Distillation: Transferring Few-shot Learning Ability of Pre-trained Language Models. *arXiv preprint arXiv:2212.10670* (2022).

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. 2019. GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. arXiv:1811.06965 [cs.CV]

Evan Hubinger. 2023. AI safety via market making. `https://www.lesswrong.com/posts/YWwzccGbcHMJMpT45/ai-safety-via-market-making.`.

Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. `https://doi.org/10.1145/3586183.3606735`

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. 2019. *Automated Machine Learning: Methods, Systems, Challenges* (1st ed.). Springer Publishing Company, Incorporated.

Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, Joe Chau, Peng Cheng, Fan Yang, Mao Yang, and Yongqiang Xiong. 2023. Tutel: Adaptive Mixture-of-Experts at Scale. arXiv:2206.03382 [cs.DC]

Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. 2023. FedPara: Low-Rank Hadamard Product for Communication-Efficient Federated Learning. arXiv:2108.06098 [cs.LG]

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing Models with Task Arithmetic. arXiv:2212.04089 [cs.LG]

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical Reasoning using Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. 37–42.

McKinsey Global Institute. 2017. Jobs lost, jobs gained: Workforce transitions in a time of automation. `https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages`.

Brookings Institution. 2021. How to combat America's digital divide. `https://www.brookings.edu/research/how-to-combat-americas-digital-divide/`.

Geoffrey Irving and Amanda Askell. 2019. AI safety needs social scientists. *Distill* 4, 2 (2019), e14.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899* (2018).

Brett W Israelsen. 2019. *Algorithmic assurances and self-assessment of competency boundaries in autonomous systems.* Ph. D. Dissertation. University of Colorado at Boulder.

Adrien Jauffret, Marwen Belkaid, Nicolas Cuperlier, Philippe Gaussier, and Philippe Tarroux. 2013a. Frustration as a way toward autonomy and self-improvement in robotic navigation. In *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. IEEE, 1–7.

Adrien Jauffret, Nicolas Cuperlier, Philippe Tarroux, and Philippe Gaussier. 2013b. From self-assessment to frustration, a small step toward autonomy in robotic navigation. *Frontiers in neurorobotics* 7 (2013), 16.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog* (2023).

Kunal Jha, Tuan Anh Le, Chuanyang Jin, Yen-Ling Kuo, Joshua B Tenenbaum, and Tianmin Shu. 2024. Neural amortized inference for nested multi-agent reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 530–537.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, and Yaodong Yang. 2024. Aligner: Achieving efficient alignment through weak-to-strong correction. *arXiv preprint arXiv:2402.02416* (2024).

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023b. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852* (2023).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. 2019. Towards Efficient Data Valuation Based on the Shapley Value. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 89)*, Kamalika Chaudhuri and Masashi Sugiyama (Eds.). PMLR, 1167–1176. `https://proceedings.mlr.press/v89/jia19a.html`

Zhihao Jia, Matei Zaharia, and Alex Aiken. 2018. Beyond Data and Model Parallelism for Deep Neural Networks. arXiv:1807.05358 [cs.DC]

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023d. Mistral 7B. arXiv:2310.06825 [cs.CL]

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023b. MotionGPT: Human Motion as a Foreign Language. *arXiv preprint arXiv:2306.14795* (2023).

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022b. Evaluating and Inducing Personality in Pre-trained Language Models. (2022).

Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023c. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3586183.3606737

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023a. Lion: Adversarial Distillation of Closed-Source Large Language Model. *arXiv preprint arXiv:2305.12870* (2023).

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2022a. Vima: General robot manipulation with multimodal prompts. *arXiv* (2022).

Youhe Jiang, Ran Yan, Xiaozhe Yao, Beidi Chen, and Binhang Yuan. 2023e. HexGen: Generative Inference of Foundation Model over Heterogeneous Decentralized Environment. arXiv:2311.11514 [cs.DC]

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? arXiv:2310.06770 [cs.CL]

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743* (2024).

Chuanyang Jin and Saining Xie. 2024. Fast-DiT: Fast Diffusion Models with Transformers. https://github.com/chuanyangjin/fast-DiT.

Chuanyang Jin, Songyang Zhang, Tianmin Shu, and Zhihan Cui. 2023c. The Cultural Psychology of Large Language Models: Is ChatGPT a Holistic or Analytic Thinker? *arXiv preprint arXiv:2308.14242* (2023).

Yunho Jin, Chun-Feng Wu, David Brooks, and Gu-Yeon Wei. 2023b. S^3: Increasing GPU Utilization during Generative Inference for Higher Throughput. arXiv:2306.06000 [cs.AR]

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023a. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh conference on neural information processing systems*.

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (Sept. 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2 Publisher: Nature Publishing Group.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. arXiv:1705.03551 [cs.CL]

Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson. 2023. TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings. arXiv:2304.01433 [cs.AR]

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), 583–589.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822* (2022).

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

Raghav Kapoor, Yash Parag Butala, Melisa Russak, Jing Yu Koh, Kiran Kamble, Waseem Alshikh, and Ruslan Salakhutdinov. 2024. OmniACT: A Dataset and Benchmark for Enabling Multimodal Generalist Autonomous Agents for Desktop and Web. *arXiv preprint arXiv:2402.17553* (2024).

Bojan Karlaš, David Dao, Matteo Interlandi, Bo Li, Sebastian Schelter, Wentao Wu, and Ce Zhang. 2022. Data Debugging with Shapley Importance over End-to-End Machine Learning Pipelines. arXiv:2204.11131 [cs.LG]

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

Naveena Karusala, Sohini Upadhyay, Rajesh Veeraraghavan, and Krzysztof Gajos. 2024. Understanding Contestability on the Margins: Implications for the Design of Algorithmic Decision-making in Public Services. (2024).

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. arXiv:2006.16236 [cs.LG]

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Z. Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Educator Needs. `https://doi.org/10.1145/3613904.3642773` arXiv:2401.11314 [cs].

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of Language Agents. `https://doi.org/10.48550/arXiv.2103.14659` arXiv:2103.14659 [cs].

Ben Kenward and Thomas Sinclair. 2021. Machine morality, moral progress, and the looming environmental disaster.

Anji Khalifa, Gamaleldin Elsayed, Minsu Baek, Noam Shazeer, and Natalia Neverova. 2022. DreamIX: DreamFusion via Iterative Spatiotemporal Mixing. *arXiv preprint arXiv:2212.04508* (2022).

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-Search-Predict: Composing Retrieval and Language Models for Knowledge-Intensive NLP. *arXiv preprint arXiv:2212.14024* (2022).

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *arXiv preprint arXiv:2310.03714* (2023).

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan

Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. 2024. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. arXiv:2403.12945 [cs.RO]

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

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. 2023. Aligning Large Language Models through Synthetic Feedback. *arXiv preprint arXiv:2305.13735* (2023).

Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. `https://doi.org/10.1145/3613904.3642216` arXiv:2309.13633 [cs].

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023a. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216* (2023).

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023b. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*. PMLR, 17283–17300.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

Bart Kosko. 1992. *Neural networks and fuzzy systems: a dynamical systems approach to machine intelligence.* Prentice-Hall, Inc.

Stefan Kramer, Mattia Cerrato, Sašo Džeroski, and Ross King. 2023. Automated Scientific Discovery: From Equation Discovery to Autonomous Discovery Systems. *arXiv preprint arXiv:2305.02251* (2023).

Ethan Kross and Ozlem Ayduk. 2017. Self-distancing: Theory, research, and current directions. In *Advances in experimental social psychology*. Vol. 55. Elsevier, 81–136.

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. 2021. Beyond Distillation: Task-level Mixture-of-Experts for Efficient Inference. arXiv:2110.03742 [cs.CL]

Eldar Kurtic, Elias Frantar, and Dan Alistarh. 2023. ZipLM: Inference-Aware Structured Pruning of Language Models. arXiv:2302.04089 [cs.LG]

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. `https://doi.org/10.1162/tacl_a_00276`

K. Hazel Kwon, Shin-Il Moon, and Michael A. Stefanone. 2015. Unspeaking on Facebook? Testing network effects on self-censorship of political expressions in social network sites. *Quality & Quantity* 49, 4 (July 2015), 1417–1435. `https://doi.org/10.1007/s11135-014-0078-8`

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023b. Reward design with language models. *arXiv preprint arXiv:2303.00001* (2023).

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023a. Efficient Memory Management for Large Language Model Serving with PagedAttention. arXiv:2309.06180 [cs.LG]

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Scott Wen tau Yih, Daniel Fried, Sida Wang, and Tao Yu. 2022. DS-1000: A Natural and Reliable Benchmark for Data Science Code Generation. arXiv:2211.11501 [cs.SE]

Brenden M Lake and Marco Baroni. 2023. Human-like systematic generalization through a meta-learning neural network. *Nature* (2023), 1–7.

Lambda. 2023. OpenAI's GPT-3 Language Model: A Technical Overview. `https://lambdalabs.com/blog/demystifying-gpt-3`

Angela Langdon, Matthew Botvinick, Hiroyuki Nakahara, Keiji Tanaka, Masayuki Matsumoto, and Ryota Kanai. 2022. Meta-learning, social cognition and consciousness in brains and machines. *Neural Networks* 145 (2022), 80–89.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems* 36 (2024).

Nam Le. 2019. Evolving Self-supervised Neural Networks Autonomous Intelligence from Evolved Self-teaching. arXiv:1906.08865 [cs.NE]

Christian Lebiere. 2007. Metrics for Cognitive Architecture Evaluation. (2007).

Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review* 62, 1 (2022).

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267* (2023).

Hao-Ping Lee, Yu-Ju Yang, Thomas Serban von Davier, Jodi Forlizzi, and Sauvik Das. 2024c. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. `https://doi.org/10.48550/arXiv.2310.07879` arXiv:2310.07879 [cs].

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L. C. Guo, Md Naimul Hoque, Yewon Kim, Seyed Parsa Neshaei, Agnia Sergeyuk, Antonette Shibani, Disha Shrivastava, Lila Shroff, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia H. Rho, Shannon Zejiang Shen, and Pao Siangliulue. 2024a. A Design Space for Intelligent and Interactive Writing Assistants. `https://doi.org/10.1145/3613904.3642697` arXiv:2403.14117 [cs].

Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI Conference on Human Factors in Computing Systems*. 1–19. `https://doi.org/10.1145/3491102.3502030` arXiv:2201.06796 [cs].

Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. 2024b. Aligning Large Language Models by On-Policy Self-Judgment. *arXiv preprint arXiv:2402.11253* (2024).

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871* (2018).

Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A Hallucination Identifier for Large Language Models. http://arxiv.org/abs/2403.06710 arXiv:2403.06710 [cs].

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv:2104.08691 [cs.CL]

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast Inference from Transformers via Speculative Decoding. arXiv:2211.17192 [cs.LG]

Sam Levin. 2017. New AI can guess whether you're gay or straight from a photograph. *The Guardian* (Sept. 2017). https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

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

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. 2023m. OtterHD: A High-Resolution Multi-modality Model. *arXiv preprint arXiv:2311.04219* (2023).

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023l. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726* (2023).

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, Manasi Sharma, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Yunzhu Li, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. 2024f. BEHAVIOR-1K: A Human-Centered, Embodied AI Benchmark with 1,000 Everyday Activities and Realistic Simulation. arXiv:2403.09227 [cs.RO]

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023e. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760* (2023).

Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. 2023i. Ethics of large language models in medicine and medical research. *The Lancet Digital Health* 5, 6 (2023), e333–e335.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. 2024b. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems* 36 (2024).

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023h. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML* (2023).

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024g. More agents is all you need. *arXiv preprint arXiv:2402.05120* (2024).

Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. 2024d. OmniActions: Predicting Digital Actions in Response to Real-World Multimodal Sensory Inputs with LLMs. (2024).

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023f. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* (2023).

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023j. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665* (2023).

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023g. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. *arXiv preprint arXiv:2306.14050* (2023).

Muyang Li, Tianle Cai, Jiaxin Cao, Qinsheng Zhang, Han Cai, Junjie Bai, Yangqing Jia, Ming-Yu Liu, Kai Li, and Song Han. 2024a. DistriFusion: Distributed Parallel Inference for High-Resolution Diffusion Models. arXiv:2402.19481 [cs.CV]

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Heng Huang, Jiuxiang Gu, and Tianyi Zhou. 2023c. Reflection-tuning: Data recycling improves llm instruction-tuning. *arXiv preprint arXiv:2310.11716* (2023).

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023a. StarCoder: may the source be with you! arXiv:2305.06161 [cs.CL]

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023b. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463* (2023).

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022b. Competition-level code generation with AlphaCode. *Science* 378, 6624 (Dec. 2022), 1092–1097. `https://doi.org/10.1126/science.abq1158`

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355* (2023).

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2024c. Guiding large language models via directional stimulus prompting. *Advances in Neural Information Processing Systems* 36 (2024).

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. 2024e. LEGO: Language Enhanced Multi-modal Grounding Model. *arXiv preprint arXiv:2401.06071* (2024).

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. 2023k. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *Forty-first International Conference on Machine Learning*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs.CL]

Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. `https://doi.org/10.48550/arXiv.2306.01941` arXiv:2306.01941 [cs].

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. 2024. Jamba: A Hybrid Transformer-Mamba Language Model. arXiv:2403.19887 [cs.CL]

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050* (2023).

Bin Lin, Tao Peng, Chen Zhang, Minmin Sun, Lanbo Li, Hanyu Zhao, Wencong Xiao, Qi Xu, Xiafei Qiu, Shen Li, Zhigang Ji, Yong Li, and Wei Lin. 2024a. Infinite-LLM: Efficient LLM Service for Long Context with DistAttention and Distributed KVCache. arXiv:2401.02669 [cs.DC]

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023b. The unlocking spell on base llms: Rethinking alignment via in-context learning. *arXiv preprint arXiv:2312.01552* (2023).

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024b. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. arXiv:2306.00978 [cs.CL]

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2024e. VILA: On Pre-training for Visual Language Models. arXiv:2312.07533 [cs.CV]

Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. 2024f. On-Device Training Under 256KB Memory. arXiv:2206.15472 [cs.CV]

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3214–3252. `https://doi.org/10.18653/v1/2022.acl-long.229`

Susan Lin, Jeremy Warner, J. D. Zamfirescu-Pereira, Matthew G. Lee, Sauhard Jain, Michael Xuelin Huang, Piyawat Lertvittayakumjorn, Shanqing Cai, Shumin Zhai, Björn Hartmann, and Can Liu. 2024d. Rambler:

Supporting Writing With Speech via LLM-Assisted Gist Manipulation. https://doi.org/10.1145/3613904.3642217 arXiv:2401.10838 [cs].

Yujun Lin, Haotian Tang, Shang Yang, Zhekai Zhang, Guangxuan Xiao, Chuang Gan, and Song Han. 2024c. QServe: W4A8KV4 Quantization and System Co-design for Efficient LLM Serving. arXiv:2405.04532 [cs.CL]

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. 2022a. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* (2022). https://doi.org/10.1101/2022.07.20.500902 arXiv:https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902.full.pdf

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Jiao Qiao. 2023a. SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. *ArXiv* abs/2311.07575 (2023). https://api.semanticscholar.org/CorpusID:265150267

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023h. We're Afraid Language Models Aren't Modeling Ambiguity. *arXiv preprint arXiv:2304.14399* (2023).

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023d. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023f. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676* 3 (2023).

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022b. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. arXiv:2205.05638 [cs.LG]

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024b. World Model on Million-Length Video And Language With RingAttention. arXiv:2402.08268 [cs.LG]

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023j. Ring Attention with Blockwise Transformers for Near-Infinite Context. arXiv:2310.01889 [cs.CL]

Jingyu Liu, Wenhan Xiong, Ian Jones, Yixin Nie, Anchit Gupta, and Barlas Oğuz. 2023i. CLIP-Layout: Style-Consistent Indoor Scene Synthesis with Semantic Furniture Embedding. arXiv:2303.03565 [cs.CV]

Ruibo Liu, Chenyan Jia, Ge Zhang, Ziyu Zhuang, Tony Liu, and Soroush Vosoughi. 2022a. Second thoughts are best: Learning to re-align with human values from text edits. *Advances in Neural Information Processing Systems* 35 (2022), 181–196.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024a. Best Practices and Lessons Learned on Synthetic Data for Language Models. *arXiv preprint arXiv:2404.07503* (2024).

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022c. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 241–252.

Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. 2023e. EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. arXiv:2305.07027 [cs.CV]

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024c. AgentBench: Evaluating LLMs as Agents. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=zAdUB0aCTQ

Yuhan Liu, Hanchen Li, Kuntai Du, Jiayi Yao, Yihua Cheng, Yuyang Huang, Shan Lu, Michael Maire, Henry Hoffmann, Ari Holtzman, Ganesh Ananthanarayanan, and Junchen Jiang. 2023c. CacheGen: Fast Context Loading for Language Model Applications. arXiv:2310.07240 [cs.NI]

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. 2023a. Scissorhands: Exploiting the Persistence of Importance Hypothesis for LLM KV Cache Compression at Test Time. arXiv:2305.17118 [cs.LG]

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, and Beidi Chen. 2023g. Deja Vu: Contextual Sparsity for Efficient LLMs at Inference Time. arXiv:2310.17157 [cs.LG]

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023k. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170* (2023).

Katherine Anne Long. 2021. Amazon and Microsoft team up to defend against facial recognition lawsuits. https://www.seattletimes.com/business/technology/facial-recognition-lawsuits-against-amazon-and-microsoft-can-proceed-judge-rules/

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 *International Conference on Machine Learning*. PMLR, 22631–22648.

Dengsheng Lu and Qihao Weng. 2007. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing* 28, 5 (2007), 823–870.

Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. Routing to the Expert: Efficient Reward-guided Ensemble of Large Language Models. arXiv:2311.08692 [cs.CL]

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. 2024. Twin-merging: Dynamic integration of modular expertise in model merging. *arXiv preprint arXiv:2406.15479* (2024).

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv preprint arXiv:2305.15023* (2023).

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics* 23, 6 (2022), bbac409.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. *arXiv preprint arXiv:2306.09093* (2023).

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards Faithful Model Explanation in NLP: A Survey. https://doi.org/10.48550/arXiv.2209.11326 arXiv:2209.11326 [cs].

Weiyu Ma, Qirui Mi, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. 2023b. Large language models play starcraft ii: Benchmarks and a chain of summarization approach. *arXiv preprint arXiv:2312.11865* (2023).

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Eureka: Human-Level Reward Design via Coding Large Language Models. arXiv:2310.12931 [cs.RO]

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* 36 (2024).

Neeratyoy Mallik, Edward Bergman, Carl Hvarfner, Danny Stoll, Maciej Janowski, Marius Lindauer, Luigi Nardi, and Frank Hutter. 2023. PriorBand: Practical Hyperparameter Optimization in the Age of Deep Learning. arXiv:2306.12370 [cs.LG]

Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. 2023a. GPT-Driver: Learning to Drive with GPT. arXiv:2310.01415 [cs.CV]

Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. 2023b. A Language Agent for Autonomous Driving. arXiv:2311.10813 [cs.CV]

Andres Marzal and Enrique Vidal. 1993. Computation of normalized edit distance and applications. *IEEE transactions on pattern analysis and machine intelligence* 15, 9 (1993), 926–932.

Krzysztof Maziarz, Iulian Zerdes, Apoorv Rastogi, Xiaofeng Yang, Jun Wang, and Allen Aristo. 2022. Molecular Optimization using Language Models. *arXiv preprint arXiv:2210.00299* (2022).

Kris McGuffie and Alex Newhouse. 2020. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. https://doi.org/10.48550/arXiv.2009.06807 arXiv:2009.06807 [cs].

Timothy R McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N Halgamuge. 2024. The inadequacy of reinforcement learning from human feedback-radicalizing large language models via semantic vulnerabilities. *IEEE Transactions on Cognitive and Developmental Systems* (2024).

David A Medler. 1998. A brief history of connectionism. *Neural computing surveys* 1 (1998), 18–72.

Kai Mei, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang. 2024. AIOS: LLM Agent Operating System. arXiv:2403.16971 [cs.OS]

Bahar Memarian and Tenzin Doleck. 2023. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review. *Computers and Education: Artificial Intelligence* (2023), 100152.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. *arXiv preprint arXiv:2404.02948* (2024).

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. GAIA: a benchmark for General AI Assistants. arXiv:2311.12983 [cs.CL]

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Hongyi Jin, Tianqi Chen, and Zhihao Jia. 2023. Towards Efficient Generative Large Language Model Serving: A Survey from Algorithms to Systems. arXiv:2312.15234 [cs.LG]

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. 2024. SpecInfer: Accelerating Generative Large Language Model Serving with Tree-based Speculative Inference and Verification. arXiv:2305.09781 [cs.CL]

Vincent Micheli, Eloi Alonso, and François Fleuret. 2023. Transformers are Sample-Efficient World Models. In *The Eleventh International Conference on Learning Representations.* https://openreview.net/forum?id=vhFu1Acb0xb

Dan Milmo. 2021. Amazon asks Ring owners to respect privacy after court rules usage broke law. *The Guardian* (Oct. 2021). `https://www.theguardian.com/uk-news/2021/oct/14/amazon-asks-ring-owners-to-respect-privacy-after-court-rules-usage-broke-law`

Adam S. Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA Internal Medicine* 176, 5 (May 2016), 619. `https://doi.org/10.1001/jamainternmed.2016.0400`

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309* (2021).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. `https://doi.org/10.1145/3287560.3287596`

MLC team. 2023. *MLC-LLM*. `https://github.com/mlc-ai/mlc-llm`

Amirkeivan Mohtashami and Martin Jaggi. 2023. Landmark Attention: Random-Access Infinite Context Length for Transformers. arXiv:2305.16300 [cs.CL]

Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2024. Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv:2311.02462 [cs.AI]

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. 2024. Active Preference Learning for Large Language Models. *arXiv preprint arXiv:2402.08114* (2024).

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. 2023. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886* (2023).

Vijayaraghavan Murali, Chandra Maddila, Imad Ahmad, Michael Bolin, Daniel Cheng, Negar Ghorbani, Renuka Fernandez, Nachiappan Nagappan, and Peter C. Rigby. 2024. AI-assisted Code Authoring at Scale: Fine-tuning, deploying, and mixed methods evaluation. arXiv:2305.12050 [cs.SE]

Brad Myers, Scott E. Hudson, and Randy Pausch. 2000. Past, present, and future of user interface software tools. *ACM Transactions on Computer-Human Interaction* 7, 1 (2000), 3–28. `https://doi.org/10.1145/344949.344959`

Meenakshi Nadimpalli. 2017. Artificial intelligence risks and benefits. *International Journal of Innovative Research in Science, Engineering and Technology* 6, 6 (2017).

Muhammad U. Nasir, Sam Earle, Julian Togelius, Steven James, and Christopher Cleghorn. 2023. LLMatic: Neural Architecture Search via Large Language Models and Quality Diversity Optimization. arXiv:2306.01102 [cs.NE]

Mohamed Nejjar, Luca Zacharias, Fabian Stiehle, and Ingo Weber. 2023. LLMs for Science: Usage for Code Generation and Data Analysis. *arXiv preprint arXiv:2311.16733* (2023).

Yuansheng Ni, Sichao Jiang, Hui Shen, Yuli Zhou, et al. 2023. Evaluating the Robustness to Instructions of Large Language Models. *arXiv preprint arXiv:2308.14306* (2023).

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

Sergey I Nikolenko. 2021. *Synthetic data for deep learning*. Vol. 174. Springer.

Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. 2024. ScreenAgent: A Vision Language Model-driven Computer Control Agent. *arXiv preprint arXiv:2402.07945* (2024).

NVIDIA. 2023a. FasterTransformer. `https://github.com/NVIDIA/FasterTransformer`.

NVIDIA. 2023b. TensorRT-LLM: A TensorRT Toolbox for Optimized Large Language Model Inference. `https://github.com/NVIDIA/TensorRT-LLM`.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114* (2021).

Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. `https://doi.org/10.1145/3173574.3174223`

OpenAI. 2018. Charter. `https://www.openai.com/charter/`.

OpenAI. 2023a. *GPT-4 Technical Report*. Technical Report. OpenAI.

OpenAI. 2023b. *GPT-4v(ision) Technical Work and Authors*. Technical Report. OpenAI. `https://cdn.openai.com/contributions/gpt-4v.pdf`

OpenAI. 2024. Hello GPT-4o. `https://openai.com/index/hello-gpt-4o/`

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. *Advances in neural information processing systems* 35 (2022), 27730–27744.

Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. West-of-n: Synthetic preference generation for improved reward modeling. *arXiv preprint arXiv:2401.12086* (2024).

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295* (2023).

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442* (2023).

Kaushikkumar Patel. 2024. Ethical reflections on data-centric AI: balancing benefits and risks. *International Journal of Artificial Intelligence Research and Development* 2, 1 (2024), 1–17.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334* (2023).

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. 2023a. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048* (2023).

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, Kranthi Kiran GV, Jan Kocoń, Bartłomiej Koptyra, Satyapriya Krishna, Ronald McClelland Jr. au2, Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Stanisław Woźniak, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. 2024a. Eagle and Finch: RWKV with Matrix-Valued States and Dynamic Recurrence. arXiv:2404.05892 [cs.CL]

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023b. Instruction Tuning with GPT-4. arXiv:2304.03277 [cs.CL]

Yujia Peng, Jiaheng Han, Zhenliang Zhang, Lifeng Fan, Tengyu Liu, Siyuan Qi, Xue Feng, Yuxi Ma, Yizhou Wang, and Song-Chun Zhu. 2024b. The tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering* 34 (2024), 12–22.

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. arXiv:2302.10866 [cs.LG]

Michael Poli, Armin W Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deiseroth, Kristian Kersting, Taiji Suzuki, Brian Hie, Stefano Ermon, Christopher Ré, Ce Zhang, and Stefano Massaroli. 2024. Mechanistic Design and Scaling of Hybrid Architectures. arXiv:2403.17844 [cs.LG]

Stanislas Polu and Ilya Sutskever. 2022. Formal mathematics statement curriculum learning. *arXiv preprint arXiv:2202.01344* (2022).

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2022. Efficiently Scaling Transformer Inference. arXiv:2211.05102 [cs.LG]

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Predibase. 2023. Multi-LoRA inference server that scales to 1000s of fine-tuned LLMs. `https://github.com/predibase/lorax`.

Mark A. Prelas, Charles L. Weaver, Matthew L. Watermann, Eric D. Lukosi, Robert J. Schott, and Denis A. Wisniewski. 2014. A review of nuclear batteries. *Progress in Nuclear Energy* 75 (2014), 117–148. `https://doi.org/10.1016/j.pnucene.2014.04.007`

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. 2020. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890* (2020).

Mengnan Qi, Yufan Huang, Yongqiang Yao, Maoquan Wang, Bin Gu, and Neel Sundaresan. 2024. Is Next Token Prediction Sufficient for GPT? Exploration on Code Logic Comprehension. *arXiv preprint arXiv:2404.08885* (2024).

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. 2023. Visual adversarial examples jailbreak aligned large language models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, Vol. 1.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023a. Communicative agents for software development. *arXiv preprint arXiv:2307.07924* (2023).

Cheng Qian, Chi Han, Yi Fung, Yujia Qin, Zhiyuan Liu, and Heng Ji. 2023b. CREATOR: Tool Creation for Disentangling Abstract and Concrete Reasoning of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6922–6939. `https://doi.org/10.18653/v1/2023.findings-emnlp.462`

Cheng Qian, Chi Han, Yi R Fung, Yujia Qin, Zhiyuan Liu, and Heng Ji. 2023c. Creator: Disentangling abstract and concrete reasonings of large language models through tool creation. *arXiv preprint arXiv:2305.14318* (2023).

Cheng Qian, Shihao Liang, Yujia Qin, Yining Ye, Xin Cong, Yankai Lin, Yesai Wu, Zhiyuan Liu, and Maosong Sun. 2024. Investigate-Consolidate-Exploit: A General Strategy for Inter-Task Agent Self-Evolution. arXiv:2401.13996 [cs.CL]

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. 2023a. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354* (2023).

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023b. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789* (2023).

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 *International conference on machine learning*. PMLR, 8748–8763.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2024).

Trivellore E Raghunathan. 2021. Synthetic data. *Annual review of statistics and its application* 8 (2021), 129–140.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv:1606.05250 [cs.CL]

Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. 2024. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187* (2024).

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. 2023. Tailoring self-rationalizers with multi-reward distillation. *arXiv preprint arXiv:2311.02805* (2023).

Waseem Rawat and Zenghui Wang. 2017. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation* 29, 9 (2017), 2352–2449.

Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. 2023. Android in the wild: A large-scale dataset for android device control. *arXiv preprint arXiv:2307.10088* (2023).

Shahana Rayhan. 2023. Ethical Implications of Creating AGI: Impact on Human Society, Privacy, and Power Dynamics. *Artificial Intelligence Review* (2023).

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266.

Scott Reed, Andy Zeng, Nando de Freitas, and Arthur Szlam. 2023. Acquisition of Multimodal Models via Retrieval. *arXiv preprint arXiv:2302.02916* (2023).

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

Bart-Jan Rem, Niklas Käming, Matthias Tarnowski, Luca Asteria, Nick Fläschner, Christoph Becker, Klaus Sengstock, and Christof Weitenberg. 2019. Identifying quantum phase transitions with adversarial neural networks. *Nature Physics* 15, 9 (2019), 917–920.

Samreen Rizvi. 2023. Blockchain-Based LLMs: A Game Changer for Data Privacy Protection. `https://www.dataversity.net/blockchain-based-llms-a-game-changer-for-data-privacy-protection`

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2020. Efficient Content-Based Sparse Attention with Routing Transformers. arXiv:2003.05997 [cs.LG]

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code Llama: Open Foundation Models for Code. arXiv:2308.12950 [cs.CL]

Sebastian Ruder. 2020. Why You Should Do NLP Beyond English. `http://ruder.io/nlp-beyond-english`.

Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Max Ryabinin, Tim Dettmers, Michael Diskin, and Alexander Borzunov. 2023. SWARM Parallelism: Training Large Models Can Be Surprisingly Communication-Efficient. arXiv:2301.11913 [cs.DC]

Aishwarya P S, Pranav Ajit Nair, Yashas Samaga, Toby Boyd, Sanjiv Kumar, Prateek Jain, and Praneeth Netrapalli. 2024. Tandem Transformers for Inference Efficient LLMs. arXiv:2402.08644 [cs.AI]

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*. 1–10.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?. In *International Conference on Machine Learning*. PMLR, 29971–30004.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312* (2022).

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. 2024. Testing the general deductive reasoning capacity of large language models using ood examples. *Advances in Neural Information Processing Systems* 36 (2024).

Apoorv Saxena. 2023. Prompt Lookup Decoding. `https://github.com/apoorvumang/prompt-lookup-decoding/`

Brian Scassellati. 2002. Theory of mind for a humanoid robot. *Autonomous Robots* 12 (2002), 13–24.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755* (2023).

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761* (2023).

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics* 9 (2021), 1408–1424.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* 588, 7839 (Dec. 2020), 604–609. `https://doi.org/10.1038/s41586-020-03051-4`

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM* 63, 12 (2020), 54–63.

Charbel-Raphaël Segerie. 2023. Task decomposition for scalable oversight (AGISF Distillation). `https://www.lesswrong.com/posts/FFz6H35Gy6BArHxkc/task-decomposition-for-scalable-oversight-agisf-distillation`.

Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron, et al. 2024. BOND: Aligning LLMs with Best-of-N Distillation. *arXiv preprint arXiv:2407.14622* (2024).

Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L. Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. `https://doi.org/10.48550/arXiv.2402.14978` arXiv:2402.14978 [cs].

Muhammad Shafay, Raja Wasim Ahmad, Khaled Salah, Ibrar Yaqoob, Raja Jayaraman, and Mohammed Omar. 2021. Blockchain for Deep Learning: Review and Open Challenges. (Oct. 2021). `https://doi.org/10.36227/techrxiv.16823140.v1`

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*. PMLR, 492–504.

Murray Shanahan and Catherine Clarke. 2023. Evaluating Large Language Model Creativity from a Literary Perspective. *arXiv preprint arXiv:2312.03746* (2023).

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. *arXiv preprint arXiv:1701.06538* (2017).

Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Ben Graff, and Dongwon Lee. 2021. Mathbert: A pre-trained language model for general nlp tasks in mathematics education. *arXiv preprint arXiv:2106.07340* (2021).

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023a. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025* (2023).

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023b. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580* (2023).

Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica. 2023a. S-LoRA: Serving Thousands of Concurrent LoRA Adapters. arXiv:2311.03285 [cs.LG]

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Daniel Y. Fu, Zhiqiang Xie, Beidi Chen, Clark Barrett, Joseph E. Gonzalez, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. 2023b. FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU. arXiv:2303.06865 [cs.LG]

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Layla Isik, Yen-Ling Kuo, and Tianmin Shu. 2024b. MuMA-ToM: Multi-modal Multi-Agent Theory of Mind. *arXiv preprint arXiv:2408.12574* (2024).

Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z. Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. 2024a. Yell At Your Robot: Improving On-the-Fly from Language Corrections. arXiv:2403.12910 [cs.RO]

Haotian Liu Hao Zhang Feng Li Tianhe Ren Xueyan Zou Jianwei Yang Hang Su Jun Zhu Lei Zhang Jianfeng Gao Chunyuan Li Shilong Liu, Hao Cheng. 2023. LLaVA-Plus: Learning to Use Tools for Creating Multimodal Agents. *arXiv preprint arXiv:2311.05437* (2023).

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. BioMegatron: Larger Biomedical Domain Language Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 4700–4706. https://doi.org/10.18653/v1/2020.emnlp-main.379

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).

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

Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *Interactions* 4, 6 (1997), 42–61. https://doi.org/10.1145/267505.267514

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv:1909.08053 [cs.CL]

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2022. Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation. arXiv:2209.05451 [cs.RO]

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2023. Audio-Visual LLM for Video Understanding. *arXiv preprint arXiv:2312.06720* (2023).

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567* (2021).

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.

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 *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11523–11530.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. 2023. Simplified State Space Layers for Sequence Modeling. arXiv:2208.04933 [cs.LG]

Nate Soares. 2016. The value learning problem. *Machine Intelligence Research Institute Technical Report* 4 (2016).

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023a. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2998–3009.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023b. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492* (2023).

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

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

Benjamin Steenhoek, Michele Tufano, Neel Sundaresan, and Alexey Svyatkovskiy. 2023. Reinforcement Learning from Automatic Feedback for High-Quality Unit Test Generation. arXiv:2310.02368 [cs.SE]

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

Sheldon Stryker. 1959. Symbolic interaction as an approach to family research. *Marriage and Family Living* 21, 2 (1959), 111–119.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355* (2023).

Budhitama Subagdja, Han Yi Tay, and Ah-Hwee Tan. 2021. Who am I?: Towards social self-awareness for intelligent agents. International Joint Conferences on Artificial Intelligence.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. 2021. Evaluating model robustness and stability to dataset shift. In *International conference on artificial intelligence and statistics*. PMLR, 2611–2619.

Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. `https://doi.org/10.1145/3613904.3642400` arXiv:2310.12953 [cs].

Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. `https://doi.org/10.1145/3586183.3606756`

Sainbayar Sukhbaatar, Rob Fergus, et al. 2016. Learning multiagent communication with backpropagation. *Advances in neural information processing systems* 29 (2016).

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. 2023. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427* (2023).

Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 2023b. 3D-GPT: Procedural 3D Modeling with Large Language Models. *arXiv preprint arXiv:2310.12945* (2023).

Qiushi Sun, Zhirui Chen, Fangzhi Xu, Kanzhi Cheng, Chang Ma, Zhangyue Yin, Jianing Wang, Chengcheng Han, Renyu Zhu, Shuai Yuan, et al. 2024a. A survey of neural code intelligence: Paradigms, advances and beyond. *arXiv preprint arXiv:2403.14734* (2024).

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023d. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222* (2023).

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023a. Retentive Network: A Successor to Transformer for Large Language Models. *ArXiv* abs/2307.08621 (2023). https://api.semanticscholar.org/CorpusID:259937453

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023c. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047* (2023).

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024b. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems* 36 (2024).

Zhiqing Sun, Longhui Yu, Yikang Shen, Weiyang Liu, Yiming Yang, Sean Welleck, and Chuang Gan. 2024c. Easy-to-Hard Generalization: Scalable Alignment Beyond Human Supervision. *arXiv preprint arXiv:2403.09472* (2024).

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296* (2017).

Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. 2023. An Empirical Study of Multimodal Model Merging. arXiv:2304.14933 [cs.CV]

Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445088

Derek Tam, Mohit Bansal, and Colin Raffel. 2023. Merging by Matching Models in Task Subspaces. arXiv:2312.04339 [cs.LG]

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. http://arxiv.org/abs/2102.02503 arXiv:2102.02503 [cs].

Weihao Tan, Ziluo Ding, Wentao Zhang, Boyu Li, Bohan Zhou, Junpeng Yue, Haochong Xia, Jiechuan Jiang, Longtao Zheng, Xinrun Xu, et al. 2024. Towards general computer control: A multimodal agent for red dead redemption ii as a case study. *arXiv preprint arXiv:2403.03186* (2024).

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024b. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. arXiv:2311.10537 [cs.CL]

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, and Will Dabney. 2024a. Understanding the performance gap between online and offline alignment algorithms. https://doi.org/10.48550/arXiv.2405.08448 arXiv:2405.08448 [cs].

Zhenheng Tang, Yuxin Wang, Xin He, Longteng Zhang, Xinglin Pan, Qiang Wang, Rongfei Zeng, Kaiyong Zhao, Shaohuai Shi, Bingsheng He, and Xiaowen Chu. 2023. FusionAI: Decentralized Training and Deploying LLMs with Massive Consumer-Level GPUs. arXiv:2309.01172 [cs.DC]

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. 2024. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387* (2024).

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *Comput. Surveys* 55, 6 (2022), 1–28.

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

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 preprint arXiv:2312.11805* (2023).

SuperBench Team. 2023. SuperBench is Measuring LLMs in The Open: A Critical Analysis.

Max Tegmark. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence.* Knopf.

Adly Templeton. 2024. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet.* Anthropic.

The Economic Times. 2024. China develops groundbreaking nuclear battery that can last 50 years without charging. `https://economictimes.indiatimes.com/news/international/business/china-introduces-revolutionary-nuclear-battery-that-lasts-50-years-without-charging/articleshow/106880627.cms?from=mdr`

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024a. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. `https://api.semanticscholar.org/CorpusID:268876071`

Runchu Tian, Yining Ye, Yujia Qin, Xin Cong, Yankai Lin, Yinxu Pan, Yesai Wu, Zhiyuan Liu, and Maosong Sun. 2024b. DebugBench: Evaluating Debugging Capability of Large Language Models. arXiv:2401.04621 [cs.SE]

Philippe Tillet, H. T. Kung, and David Cox. 2019. Triton: an intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages* (Phoenix, AZ, USA) *(MAPL 2019)*. Association for Computing Machinery, New York, NY, USA, 10–19. `https://doi.org/10.1145/3315508.3329973`

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models. arXiv:2205.10770 [cs.CL]

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860* (2024).

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. *ArXiv* abs/2401.06209 (2024). `https://api.semanticscholar.org/CorpusID:266976992`

Alexander Tornede, Difan Deng, Theresa Eimer, Joseph Giovanelli, Aditya Mohan, Tim Ruhkopf, Sarah Segel, Daphne Theodorakopoulos, Tanja Tornede, Henning Wachsmuth, and Marius Lindauer. 2024. AutoML in the Age of Large Language Models: Current Challenges, Future Opportunities and Risks. arXiv:2306.08107 [cs.LG]

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. 2023a. How Many Unicorns Are in This Image? A Safety Evaluation Benchmark for Vision LLMs. *arXiv preprint arXiv:2311.16101* (2023).

Haoqin Tu, Yitong Li, Fei Mi, and Zhongliang Yang. 2023b. ReSee: Responding through Seeing Fine-grained Visual Knowledge in Open-domain Dialogue. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 7720–7735.

Haoqin Tu, Bingchen Zhao, Chen Wei, and Cihang Xie. 2023c. Sight Beyond Text: Multi-Modal Training Enhances LLMs in Truthfulness and Ethics. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

A. M. Turing. 1950. Computing Machinery and Intelligence. *Mind* 59, 236 (1950), 433–460. `http://www.jstor.org/stable/2251299`

Victor Turner. 1975. Symbolic studies. *Annual review of anthropology* 4, 1 (1975), 145–161.

Umit Volkan Ucak, Islambek Ashyrmamatov, Junsu Ko, and Juyong Lee. 2022. RetroTRAE: retrosynthetic translation of atomic environments with Transformer. (2022).

Saad Ullah, Mingji Han, Saurabh Pujar, Hammond Pearce, Ayse Coskun, and Gianluca Stringhini. 2023. Can Large Language Models Identify And Reason About Security Vulnerabilities? Not Yet. arXiv:2312.12575 [cs.CR]

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399* (2023).

Stephanie Valencia, Richard Cave, Krystal Kallarackal, Katie Seaver, Michael Terry, and Shaun K. Kane. 2023. "The less I type, the better": How AI Language Models can Enhance or Impede Communication for AAC Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. `https://doi.org/10.1145/3544548.3581560`

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

Paul Vicol, William Menapace, Kaushik Srinivasan, Caglar Gulcehre, Danilo Rezende, and Peter Battaglia. 2022. SimNet: Learning Simulation-Based World Models for Physical Reasoning. In *International Conference on Learning Representations*.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Caglar Gulcehre, Ziyun Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019b. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575 (2019), 350 – 354. `https://api.semanticscholar.org/CorpusID:204972004`

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019a. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

Peter Voss and Mladjan Jovanovic. 2023. Concepts is All You Need: A More Direct Path to AGI. *arXiv preprint arXiv:2309.01622* (2023).

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31 (2017), 841.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023. GPT-4V(ision) for Robotics: Multimodal Task Planning from Human Demonstration. arXiv:2311.12015 [cs.RO]

Chi Wang, Susan Xueqing Liu, and Ahmed H. Awadallah. 2023f. Cost-Effective Hyperparameter Optimization for Large Language Model Generation Inference. arXiv:2303.04673 [cs.CL]

Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019b. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. 2023d. What Makes for Good Visual Tokenizers for Large Language Models? *arXiv preprint arXiv:2305.12223* (2023).

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023h. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* (2023).

Haining Wang, Jimmy Huang, and Zhewei Zhang. 2019a. The Impact of Deep Learning on Organizational Agility.. In *ICIS*.

Jue Wang, Yucheng Lu, Binhang Yuan, Beidi Chen, Percy Liang, Christopher De Sa, Christopher Re, and Ce Zhang. 2023g. CocktailSGD: Fine-tuning Foundation Models over 500Mbps Networks. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 36058–36076. https://proceedings.mlr.press/v202/wang23t.html

Jue Wang, Binhang Yuan, Luka Rimanic, Yongjun He, Tri Dao, Beidi Chen, Christopher Re, and Ce Zhang. 2023i. Fine-tuning Language Models over Slow Networks using Activation Compression with Guarantees. arXiv:2206.01299 [cs.LG]

Kaiwen Wang, Rahul Kidambi, Ryan Sullivan, Alekh Agarwal, Christoph Dann, Andrea Michi, Marco Gelmi, Yunxuan Li, Raghav Gupta, Avinava Dubey, et al. 2024a. Conditioned Language Policy: A General Framework for Steerable Multi-Objective Finetuning. *arXiv preprint arXiv:2407.15762* (2024).

Pei Wang. 2010/06. The Evaluation of AGI Systems. In *Proceedings of the 3d Conference on Artificial General Intelligence (2010)*. Atlantis Press, 154–159. https://doi.org/10.2991/agi.2010.33

Pei Wang and Patrick Hammer. 2018. Perception from an AGI perspective. In *Artificial General Intelligence: 11th International Conference, AGI 2018, Prague, Czech Republic, August 22-25, 2018, Proceedings 11*. Springer, 259–269.

Pei Wang, Kai Liu, and Quinn Dougherty. 2018. Conceptions of artificial intelligence and singularity. *Information* 9, 4 (2018), 79.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023c. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175* (2023).

Xiaohui Wang, Ying Xiong, Yang Wei, Mingxuan Wang, and Lei Li. 2021. LightSeq: A High Performance Inference Library for Transformers. arXiv:2010.13887 [cs.MS]

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. 2023j. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284* (2023).

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024b. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529* (2024).

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560* (2022).

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023e. Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv:2212.10560 [cs.CL]

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705* (2022).

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* 53, 3 (2020), 1–34.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. 2023a. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *arXiv preprint arXiv:2311.05997* (2023).

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023b. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560* (2023).

Chen Wei, Chenxi Liu, Siyuan Qiao, Zhishuai Zhang, Alan Yuille, and Jiahui Yu. 2023. De-Diffusion Makes Text a Strong Cross-Modal Interface. *arXiv preprint arXiv:2311.00618* (2023).

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent Abilities of Large Language Models. arXiv:2206.07682 [cs.CL]

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. `https://doi.org/10.1145/3613904.3642466` arXiv:2401.14484 [cs].

Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao MA, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. 2024. DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models. In *The Twelfth International Conference on Learning Representations*. `https://openreview.net/forum?id=OqTMUPuLuC`

Kyle Wiggers. 2021. AI datasets are prone to mismanagement, study finds. https://venturebeat.com/ai/ai-datasets-are-prone-to-mismanagement-study-finds/

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language Models are Few-shot Multilingual Learners. https://doi.org/10.48550/arXiv.2109.07684 arXiv:2109.07684 [cs].

Alan F Winfield, Katina Michael, Jeremy Pitt, and Vanessa Evers. 2019. Machine ethics: The design and governance of ethical AI and autonomous systems [scanning the issue]. *Proc. IEEE* 107, 3 (2019), 509–517.

Wai Kin Wong, Huaijin Wang, Zongjie Li, Zhibo Liu, Shuai Wang, Qiyi Tang, Sen Nie, and Shi Wu. 2023. Refining Decompiled C Code with Large Language Models. arXiv:2310.06530 [cs.SE]

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. arXiv:2203.05482 [cs.LG]

Bingyang Wu, Yinmin Zhong, Zili Zhang, Gang Huang, Xuanzhe Liu, and Xin Jin. 2023c. Fast Distributed Inference Serving for Large Language Models. arXiv:2305.05920 [cs.LG]

Chengyue Wu, Teng Wang, Yixiao Ge, Zeyu Lu, Ruisong Zhou, Ying Shan, and Ping Luo. 2023b. π-Tuning: Transferring Multimodal Foundation Models with Optimal Multi-task Interpolation. arXiv:2304.14381 [cs.CV]

Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. 2017. Learning to See Physics via Visual De-animation. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/4c56ff4ce4aaf9573aa5dff913df997a-Paper.pdf

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023a. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136.

Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J. Cai. 2022a. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. https://doi.org/10.48550/arXiv.2203.06566 arXiv:2203.06566 [cs].

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022b. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3491102.3517582

Wilson Wu, John X Morris, and Lionel Levine. 2024c. Do language models plan ahead for future tokens? *arXiv preprint arXiv:2404.00859* (2024).

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2024a. ReFT: Representation Finetuning for Language Models. *arXiv preprint arXiv:2404.03592* (2024).

Zhiyong Wu, Chengcheng Han, Zichen Ding, Zhenmin Weng, Zhoumianze Liu, Shunyu Yao, Tao Yu, and Lingpeng Kong. 2024b. OS-Copilot: Towards Generalist Computer Agents with Self-Improvement. arXiv:2402.07456 [cs.AI]

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864* (2023).

Haojun Xia, Zhen Zheng, Yuchao Li, Donglin Zhuang, Zhongzhu Zhou, Xiafei Qiu, Yong Li, Wei Lin, and Shuaiwen Leon Song. 2023b. Flash-LLM: Enabling Cost-Effective and Highly-Efficient Large Generative Model Inference with Unstructured Sparsity. arXiv:2309.10285 [cs.DC]

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. 2023a. Training Trajectories of Language Models Across Scales. arXiv:2212.09803 [cs.CL]

Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. 2023. Language Models Meet World Models: Embodied Experiences Enhance Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems*. `https://openreview.net/forum?id=SVBR6xBaMl`

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2024a. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. arXiv:2211.10438 [cs.CL]

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient Streaming Language Models with Attention Sinks. arXiv:2309.17453 [cs.CL]

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024b. Efficient Streaming Language Models with Attention Sinks. arXiv:2309.17453 [cs.CL]

Yaqi Xie, Chen Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh. 2023. Translating natural language to planning goals with large-language models. *arXiv preprint arXiv:2302.05128* (2023).

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. 2023. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647* (2023).

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2023. Effective Long-Context Scaling of Foundation Models. arXiv:2309.16039 [cs.CL]

Bowen Xu and Quansheng Ren. 2022. Artificial Open World for Evaluating AGI: A Conceptual Design. In *International Conference on Artificial General Intelligence*. Springer, 452–463.

Canwen Xu, Zexue He, Zhankui He, and Julian McAuley. 2022. Leashing the Inner Demons: Self-Detoxification for Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11530–11537.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244* (2023).

Fangzhi Xu, Zhiyong Wu, Qiushi Sun, Siyu Ren, Fei Yuan, Shuai Yuan, Qika Lin, Yu Qiao, and Jun Liu. 2023b. Symbol-LLM: Towards foundational symbol-centric interface for large language models. *arXiv preprint arXiv:2311.09278* (2023).

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023c. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148* (2023).

Xinrun Xu, Yuxin Wang, Chaoyi Xu, Ziluo Ding, Jiechuan Jiang, Zhiming Ding, and Börje F Karlsson. 2024. A Survey on Game Playing Agents and Large Models: Methods, Applications, and Challenges. *arXiv preprint arXiv:2403.10249* (2024).

Roman V Yampolskiy. 2020. *Artificial Intelligence Safety and Security*. CRC Press.

King-Yin Yan. 2022. AGI via Combining Logic with Deep Learning. In *Artificial General Intelligence: 14th International Conference, AGI 2021, Palo Alto, CA, USA, October 15–18, 2021, Proceedings 14*. Springer, 327–343.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model Merging in LLMs, MLLMs, and Beyond: Methods, Theories, Applications and Opportunities. *arXiv preprint arXiv:2408.07666* (2024).

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. 2023c. InterCode: Standardizing and Benchmarking Interactive Coding with Execution Feedback. arXiv:2306.14898 [cs.CL]

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023f. Diffusion models: A comprehensive survey of methods and applications. *Comput. Surveys* 56, 4 (2023), 1–39.

Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023a. Harnessing Biomedical Literature to Calibrate Clinicians' Trust in AI Decision Support Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–14. https://doi.org/10.1145/3544548.3581393

Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023d. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752* (2023).

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. 2023e. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635* (2023).

Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2023b. AppAgent: Multimodal Agents as Smartphone Users. *arXiv preprint arXiv:2312.13771* (2023).

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601* (2023).

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing* (2024), 100211.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178* (2023).

Anil Yemme and Shayan Srinivasa Garani. 2023. A Scalable GPT-2 Inference Hardware Architecture on FPGA. In *2023 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN54540.2023.10191067

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549* (2023).

William York and Jerry Swan. 2012. Taking Turing Seriously.

Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. 2022. Orca: A Distributed Serving System for Transformer-Based Generative Models. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*. USENIX Association, Carlsbad, CA, 521–538. https://www.usenix.org/conference/osdi22/presentation/yu

Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R Lesser, and Qiang Yang. 2018. Building ethics into artificial intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 5527–5533.

Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. 2023. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647* (2023).

Binhang Yuan, Yongjun He, Jared Quincy Davis, Tianyi Zhang, Tri Dao, Beidi Chen, Percy Liang, Christopher Re, and Ce Zhang. 2023. Decentralized Training of Foundation Models in Heterogeneous Environments. arXiv:2206.01288 [cs.DC]

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502* (2023).

Lotfi A Zadeh. 1996. Fuzzy logic, neural networks, and soft computing. In *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh*. World Scientific, 775–782.

J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3581388

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067* (2023).

Yujia Zhai, Chengquan Jiang, Leyuan Wang, Xiaoying Jia, Shang Zhang, Zizhong Chen, Xin Liu, and Yibo Zhu. 2023. ByteTransformer: A High-Performance Transformer Boosted for Variable-Length Inputs. arXiv:2210.03052 [cs.LG]

Chaoyun Zhang, Liqun Li, Shilin He, Xu Zhang, Bo Qiao, Si Qin, Minghua Ma, Yu Kang, Qingwei Lin, Saravan Rajmohan, et al. 2024b. UFO: A UI-Focused Agent for Windows OS Interaction. *arXiv preprint arXiv:2402.07939* (2024).

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2023h. A Simple LLM Framework for Long-Range Video Question-Answering. *arXiv preprint arXiv:2312.17235* (2023).

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. 2023b. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485* (2023).

Hang Zhang, Xin Li, and Lidong Bing. 2023g. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858* (2023).

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. 2023k. Draft & Verify: Lossless Large Language Model Acceleration via Self-Speculative Decoding. arXiv:2309.08168 [cs.CL]

Jintian Zhang, Xin Xu, and Shumin Deng. 2023l. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124* (2023).

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023i. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Xin Wen, and Bingchen Zhao. 2023m. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4629–4633.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023a. AdaLoRA: Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. arXiv:2303.10512 [cs.CL]

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023d. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. arXiv:2303.16199 [cs.CV]

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023e. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199* (2023).

Ruohan Zhang, Sharon Lee, Minjune Hwang, Ayano Hiranaka, Chen Wang, Wensi Ai, Jin Jie Ryan Tan, Shreya Gupta, Yilun Hao, Gabrael Levine, Ruohan Gao, Anthony Norcia, Li Fei-Fei, and Jiajun Wu. 2023f. NOIR: Neural Signal Operated Intelligent Robots for Everyday Activities. In *7th Annual Conference on Robot Learning*. https://openreview.net/forum?id=eyykI3UIHa

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. OPT: Open Pre-trained Transformer Language Models. arXiv:2205.01068 [cs.CL]

Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. 2024d. Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. https://doi.org/10.1145/3613904.3642343 arXiv:2309.12368 [cs].

Xinyue Zhang, Yueying Wang, Weishan Zhang, Yuanyuan Sun, Siyu He, Gabriella Contardo, Francisco Villaescusa-Navarro, and Shirley Ho. 2019. From dark matter to galaxies with convolutional networks. *Proceedings of the National Academy of Sciences* 116, 28 (2019), 13825–13832.

Xinyu Zhang, Huiyu Xu, Zhongjie Ba, Zhibo Wang, Yuan Hong, Jian Liu, Zhan Qin, and Kui Ren. 2024c. Privacyasst: Safeguarding user privacy in tool-using large language model agents. *IEEE Transactions on Dependable and Secure Computing* (2024).

Yuxuan Zhang, Kevin Eykholt, Shaoqing Ren, Michelle Lee, Filip Radenovic, Pascal Fua, and Animesh Garg. 2023c. MetaSim: Learning to Generate Synthetic Datasets. *arXiv preprint arXiv:2302.03213* (2023).

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024a. A Survey on the Memory Mechanism of Large Language Model based Agents. *arXiv preprint arXiv:2404.13501* (2024).

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. 2023j. H_2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. arXiv:2306.14048 [cs.LG]

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493* (2022).

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024a. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19632–19642.

Bingchen Zhao, Haoqin Tu, Chen Wei, Jieru Mei, and Cihang Xie. 2023c. Tuning LayerNorm in Attention: Towards Efficient Multi-Modal LLM Finetuning. *arXiv preprint arXiv:2312.11420* (2023).

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024b. GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection. *arXiv preprint arXiv:2403.03507* (2024).

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023a. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425* (2023).

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023b. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934* (2023).

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, et al. 2024c. Assessing and Understanding Creativity in Large Language Models. *arXiv preprint arXiv:2401.12491* (2024).

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023b. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797* (2023).

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023a. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239* (2023).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).

Lianmin Zheng, Zhuohan Li, Hao Zhang, Yonghao Zhuang, Zhifeng Chen, Yanping Huang, Yida Wang, Yuanzhong Xu, Danyang Zhuo, Eric P. Xing, Joseph E. Gonzalez, and Ion Stoica. 2022. Alpa: Automating Inter- and Intra-Operator Parallelism for Distributed Deep Learning. arXiv:2201.12023 [cs.LG]

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023b. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198* (2023).

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023a. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364* (2023).

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024a. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems* 36 (2024).

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022b. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625* (2022).

Jian Zhou and Olga G Troyanskaya. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods* 12, 10 (2015), 931–934.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023b. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439* (2023).

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024b. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *The Twelfth International Conference on Learning Representations*. `https://openreview.net/forum?id=oKn9c6ytLx`

Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, et al. 2023a. Agents: An open-source framework for autonomous language agents. *arXiv preprint arXiv:2309.07870* (2023).

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023d. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667* (2023).

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. 2022a. Mixture-of-Experts with Expert Choice Routing. arXiv:2202.09368 [cs.LG]

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023c. Large Language Models Are Human-Level Prompt Engineers. arXiv:2211.01910 [cs.LG]

Banghua Zhu, Jiantao Jiao, and Michael I Jordan. 2023b. Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons. *arXiv preprint arXiv:2301.11270* (2023).

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023c. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. *arXiv preprint arXiv:2310.01852* (2023).

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. ToolQA: A Dataset for LLM Question Answering with External Tools. *arXiv preprint arXiv:2306.13304* (2023).

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, et al. 2023. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066* (2023).

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. 2008. Maximum entropy inverse reinforcement learning.. In *Aaai*, Vol. 8. Chicago, IL, USA, 1433–1438.

Lucas Zimmer, Marius Lindauer, and Frank Hutter. 2020. Auto-pytorch tabular: Multi-fidelity metalearning for efficient and robust autodl. *arXiv preprint arXiv:2006.13799* (2020).

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of The 7th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 229)*, Jie Tan, Marc Toussaint, and Kourosh Darvish (Eds.). PMLR, 2165–2183. `https://proceedings.mlr.press/v229/zitkovich23a.html`