A Survey of Autonomous Agents and Large Language Models in Achieving Human-Level Intelligence

www.surveyx.cn

Abstract

This survey paper delves into the expansive domain of artificial intelligence, focusing on autonomous agents and large language models (LLMs) and their roles in advancing towards human-level intelligence, Artificial General Intelligence (AGI), and Artificial Superintelligence (ASI). It examines the transformative potential of autonomous agents operating independently and LLMs enhancing AI's cognitive and interactive capabilities. The survey underscores the significance of interdisciplinary approaches and technological advancements necessary for AGI progression, highlighting the integration of LLMs into multi-agent systems and their applications in diverse environments. It also addresses standardization efforts in AI safety and trustworthiness, emphasizing ethical considerations and societal implications. The paper provides a comprehensive analysis of the current state of AGI development, exploring methodologies and challenges associated with achieving human-like intelligence. Furthermore, it discusses the theoretical foundations and capabilities of ASI, along with ethical and safety concerns. The survey concludes by identifying future research opportunities and the importance of responsible AI development, advocating for frameworks that ensure AI alignment with human values and societal needs. Through this holistic examination, the survey projects future directions in AI, underscoring the critical roles of autonomous agents and LLMs in this transformative journey.

1 Introduction

1.1 Scope and Significance

This survey examines autonomous agents and large language models (LLMs) within the context of achieving human-level intelligence, focusing on their roles in the development of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI). Autonomous agents are crucial to evolving computational agency paradigms, including normative, adaptive, and rational choice models [1]. Their capability to operate independently and interact with environments significantly enhances AI systems' decision-making and autonomy [2].

LLMs have transformative potential in understanding and generating human-like language, and their integration into multi-agent systems enriches AI's cognitive and interactive functionalities [3]. This survey's significance lies in its comprehensive analysis of technological advancements and interdisciplinary approaches necessary for progressing towards AGI, including LLM-based agents in diverse environments such as games [4].

Additionally, the survey addresses standardization efforts in AI safety and trustworthiness, which are vital for deploying LLMs and foundational models across various sectors, including healthcare and industry [5]. By exploring cognitive frameworks, adaptive learning mechanisms, and ethical considerations in multimodal AI integration, the survey highlights the educational and societal implications of these technologies [6]. This holistic examination delineates the current AI landscape

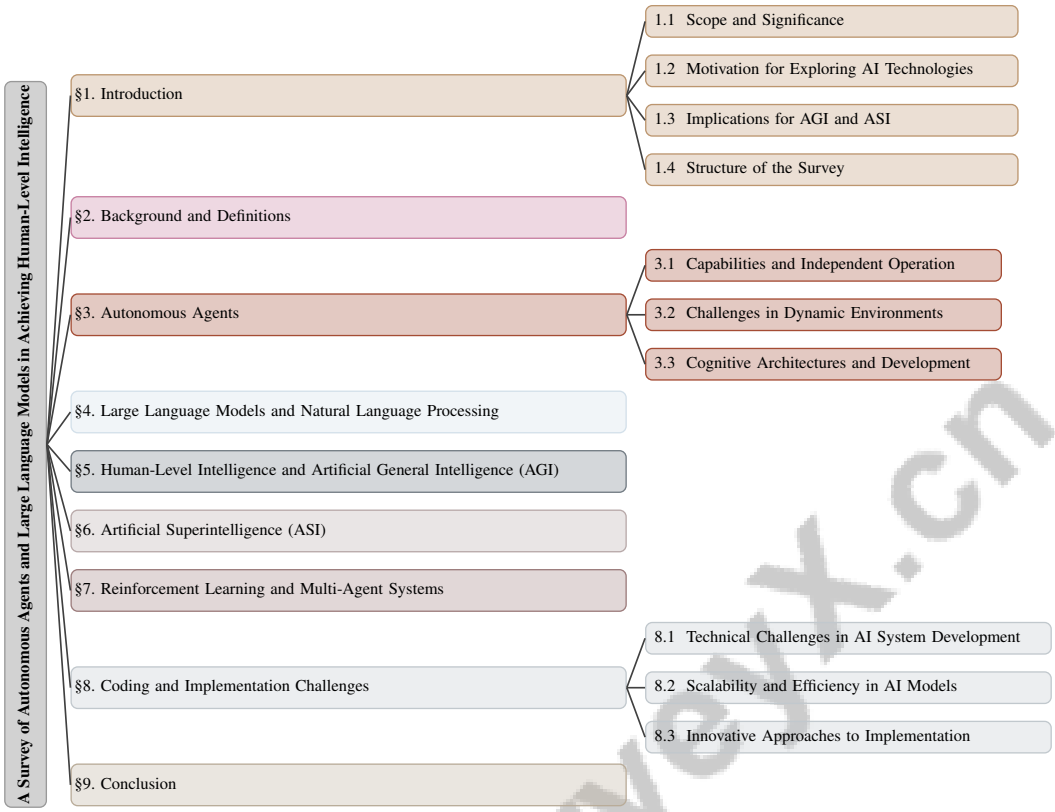


Figure 1: chapter structure

while projecting future directions in the pursuit of AGI and ASI, underscoring the pivotal roles of autonomous agents and LLMs in this transformative journey.

1.2 Motivation for Exploring AI Technologies

The exploration of autonomous agents and LLMs is driven by the imperative to advance artificial intelligence toward human-level intelligence and beyond. A primary motivation is understanding the goals, predictions, and risks associated with AI development, serving as a foundation for exploring AGI [7]. Integrating LLMs into AI systems bridges knowledge gaps, particularly in developing general AI agents for diverse applications [3].

The quest for AGI is further propelled by the need for practical benchmarks to measure intelligence beyond traditional AI's narrow focus [8]. Addressing this gap is essential for operationalizing progress in AI. Moreover, exploring LLM-based agents in domains like gaming emphasizes the importance of providing comprehensive overviews of these applications, which remain underexplored in the literature [4].

AGI's potential to transform fields such as radiation oncology by integrating multimodal clinical data enhances patient care and treatment efficiency [9]. Additionally, integrating computational experiments with LLMs aims to improve modeling complex social systems, overcoming traditional Agent-based Modeling (ABM) limitations in representing human-like behaviors [10].

The motivation extends to developing an intelligent world that closely approximates our own, embedding reasoning capabilities in intelligent machines [11]. Grounding in AI, connecting natural language and abstract knowledge to internal representations, is crucial for achieving human-level intelligence and AGI [12].

Diverse motivations behind research in autonomous agents and LLMs aim to create sophisticated, aligned, and impactful AI systems. These advancements are vital for addressing complex challenges and making significant progress toward AGI and beyond. Current efforts focus on enhancing LLM capabilities through multi-agent collaboration, developing frameworks integrating various AI

dimensions, and exploring ethical implications and alignment technologies necessary for responsible AGI progression. By leveraging LLMs' unique strengths and fostering cooperative interactions among intelligent agents, researchers seek to unlock new intelligence levels and address pressing societal needs [13, 14, 3, 15, 16].

1.3 Implications for AGI and ASI

The pursuit of AGI and ASI carries profound implications across technological, societal, and ethical domains. Achieving AGI requires a comprehensive understanding of self within autonomous agents, integrating self-preservation and environmental interaction as foundational elements for development [1]. This aligns with challenges posed by embodied cognition, necessitating consideration of physical experiences and interactions intrinsic to human-like intelligence [17].

The development of AGI hinges on frameworks prioritizing human values and safety, as shown by structured approaches aligning AGI systems with ethical standards and societal goals [18]. However, limitations of LLMs, highlighted by the TWT benchmark, caution against prematurely attributing semantic understanding or AGI capabilities to these models without addressing comprehension deficits [19].

Integrating AGI into practical domains, such as radiation oncology, underscores its potential to enhance efficiency and optimize complex processes, improving outcomes in critical fields [9]. Yet, the path to AGI is fraught with challenges, including the need for a formalized enactive model of cognition encompassing learning and reasoning in real-world contexts rather than isolated tasks [20]. This complexity is compounded by the Consistent Reasoning Paradox (CRP), highlighting difficulties AI systems encounter in handling semantically equivalent tasks framed differently [21].

Moreover, the pursuit of AGI raises existential risks, as current AI technologies may exacerbate threats even before AGI is fully realized [22]. This urgency underscores the need for multinational consortia like MAGIC, aiming to ensure safe and equitable AI development while mitigating risks associated with uncontrolled AGI progression [23]. Distinguishing between AGI and Human-Level Artificial Intelligence (HLAI) is critical, as AGI's capabilities may surpass human cognition in unforeseen ways, necessitating safety mechanisms to prevent harmful behaviors [24].

As we advance towards AGI and ASI, it is imperative to develop these technologies responsibly, enhancing human capabilities and contributing positively to society. This involves addressing potential risks and ethical concerns, fostering an approach to AGI development that aligns with human values and societal aspirations [25].

1.4 Structure of the Survey

This survey is systematically organized to comprehensively examine the role of autonomous agents and LLMs in achieving human-level intelligence and their implications for AGI and ASI. The paper is divided into several chapters, each addressing a specific aspect of the relationship between visual perception and language generation, ultimately contributing to the broader goal of achieving AGI [26, 27, 28, 29, 30].

Section 2, "Background and Definitions," provides foundational insights into key concepts and technologies pertinent to AI, including definitions and explanations of autonomous agents, LLMs, AGI, ASI, and related fields, establishing a common understanding of their interrelated roles in AI development.

Section 3, "Autonomous Agents," explores the development and capabilities of autonomous agents, focusing on their independent operation and interaction with dynamic environments. It investigates the complexities and cognitive frameworks necessary for developing agents capable of learning and executing intricate tasks, highlighting LLMs' limitations in multi-step problem-solving and reasoning. The section introduces the STARS cognitive-agent approach, which enhances prompt engineering by enabling agents to evaluate and select optimal responses based on their language capabilities, environment, and user preferences. This method has achieved high task completion rates with minimal human oversight and emphasizes the need for a dual-layer functional architecture in Cognitive AI to advance toward more sophisticated forms of artificial intelligence, such as AGI [31, 32].

Section 4, "Large Language Models and Natural Language Processing," examines the transformative role of LLMs in understanding and generating human language, discussing their applications in natural language processing and integration with autonomous agents to enhance AI capabilities.

Section 5, "Human-Level Intelligence and Artificial General Intelligence (AGI)," explores the concept of human-level intelligence within AI and the current state of AGI development. It analyzes proposed methodologies for AGI development, assessing their feasibility and inherent challenges while considering frameworks for AGI capabilities, alignment technologies, and ethical implications across various domains, including education [14, 33].

Section 6, "Artificial Superintelligence (ASI)," addresses the theoretical foundations and potential capabilities of ASI, along with ethical and safety concerns related to its development. The analysis delves into societal and governance ramifications of achieving AGI, emphasizing the necessity for regulatory frameworks prioritizing human ethics and values. It highlights the diverse definitions of AGI, reflecting assumptions about intelligence and implications for social justice. The discussion advocates for a participatory and democratic approach to AI development, focusing on ethical considerations and societal impacts as technologies evolve [34, 35, 36].

Section 7, "Reinforcement Learning and Multi-Agent Systems," provides a comprehensive analysis of reinforcement learning's contribution to training AI models, emphasizing its role in enabling agents to learn from trial and error. This section also highlights the critical importance of multi-agent systems, which facilitate executing complex tasks through collaborative interactions, essential for enhancing adaptability and effectiveness in dynamic environments [37, 3].

Section 8, "Coding and Implementation Challenges," explores the technical aspects of coding and implementing AI systems. It highlights the significant challenges in developing robust and efficient models for AI, emphasizing the need for scalability and innovative implementation strategies to address biases, adaptability, and complexities of prompt engineering while advocating for standardized methodologies and ethical guidelines in the evolving landscape of LLMs [38, 13, 14, 39, 40].

Section 9 concludes the survey with a comprehensive summary of key findings and insights, reflecting on progress in AI and its implications for future advancements. It discusses potential breakthroughs in achieving AGI and outlines necessary strategies for responsible development, addressing challenges and opportunities presented by LLMs and emphasizing the importance of interdisciplinary approaches in AI research [26, 13, 41, 14, 42]. The conclusion underscores the broader implications for society and the significance of responsible AI development.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Key Concepts and Technologies

Advancements in artificial intelligence (AI) towards human-level cognition hinge on key concepts and technologies. Autonomous agents, capable of independent and adaptive operation, are fundamental to this evolution. The integration of Large Language Models (LLMs) enhances these agents' decision-making and collaborative abilities, as shown in computational experiments that highlight both their strengths and challenges [10, 3]. LLM-based agents are typically structured with components like brain, perception, and action, enabling tailored applications across various domains.

LLMs are pivotal in AI due to their proficiency in understanding and generating human-like language, essential for tasks demanding nuanced reasoning. However, challenges persist in grounding abstract knowledge and natural language in internal representations shaped by sensorimotor experiences [12, 9]. This issue underscores the need for a deeper understanding of LLMs' cognitive processes, particularly when integrating them with large visual models (LVMs) to enhance applications in fields like radiation oncology.

Artificial General Intelligence (AGI) aims to develop systems capable of continuous learning and reasoning akin to human cognitive frameworks. Progress is hindered by the need for efficient learning models that reason across diverse contexts and the integration of human-centered design principles to guide ethical development [2]. The challenge lies in creating a self-programming cognitive engine that surpasses traditional programming limitations, facilitating more human-like cognition.

Artificial Superintelligence (ASI) represents a theoretical leap beyond human intelligence, requiring strategies to align AI systems with human values to prevent unintended consequences. This necessitates reevaluating intelligence frameworks to focus on cognitive resources and functions rather than mere imitation. Skepticism about LLMs' ability to achieve AGI, due to constraints in memory, planning, and grounding in the physical world, complicates the landscape further [2].

Embodied AI (E-AI) is crucial for AGI, emphasizing real-world interactions and sensory experiences in AI development. This approach is vital for creating intelligent systems capable of human-like communication and understanding. Efficient distribution of computational resources across multiple nodes is essential for optimizing performance in distributed systems, a key consideration in multi-agent system development [2].

These concepts and technologies form the foundation of contemporary AI research, driving the creation of intelligent systems aiming to achieve and possibly exceed human-level cognitive functions. Understanding the limitations and potentials of emerging technologies is critical for guiding future advancements, particularly in AGI safety. This involves addressing challenges like value specification, safe learning, and security measures, as emphasized in recent literature. As AGI development progresses, establishing best practices, conducting thorough risk assessments, and ensuring alignment with human values are imperative to mitigate risks associated with deploying AGI systems. Moreover, integrating AGI within the Internet of Things (IoT) presents both opportunities and challenges, necessitating dedicated research to navigate constraints related to resource availability, communication complexities, and security concerns [43, 25, 14, 44].

2.2 Interconnections and Roles

The interconnections among autonomous agents, large language models (LLMs), and other AI technologies are crucial for developing advanced AI systems capable of human-level intelligence. These interconnections enable the integration of cognitive tasks and domain-specific models, creating cohesive frameworks that facilitate complex problem-solving capabilities within AI systems [45]. The integration of various reasoning types, including scientific and common-sense reasoning, under hypothetic-deductive reasoning serves as a theoretical framework for understanding these interconnections [46].

A primary challenge in this integration is the dynamic nature of workloads and the unpredictability of resource demands, complicating effective resource allocation [47]. Addressing these challenges is crucial for optimizing AI systems' performance, particularly in distributed environments where efficient resource management is essential. Innovations such as integrating distributed ledger technologies for value alignment and quantum computing for enhanced processing capabilities represent significant advancements [18].

Developing AGI relies on a comprehensive understanding of the interconnections between safety and other AI concepts, highlighting the complexity of predicting AGI behavior and the necessity of aligning AGI goals with human values [25]. This alignment is vital for ensuring that AI systems operate within ethical frameworks and contribute positively to societal objectives. Integrating cognitive architectures with embodied AI components, such as perception, action, memory, and learning, is essential for developing AI systems capable of human-like interactions and understanding [48].

Recent benchmarks emphasize a multidimensional approach to assessing intelligence, capturing a broader range of cognitive abilities than traditional measures, underscoring the importance of comprehensive evaluation frameworks in AI development [49]. This approach aligns with the enactive perspective on cognition, which integrates cognition with the environment and tasks, fostering a more holistic understanding of intelligence [20].

Additionally, a theoretical framework that combines generality and adaptation as fundamental components of System-2 reasoning provides a basis for understanding how AI systems can achieve reasoning generality and adaptability [50]. In multi-agent systems, the interconnections between communication methods, including emergent communication and language-oriented semantic communication, are critical for effective navigation tasks, despite their limitations [51].

These interconnections and roles are pivotal for advancing AI technologies, facilitating the development of systems that not only exhibit intelligence but also prioritize ethical considerations,

transparency, and alignment with human values and societal needs. This includes integrating principles from Friendly AI (FAI) to promote equitable AI development, employing Explainable AI (XAI) methodologies to enhance interpretability and accountability, and addressing the challenges of achieving Artificial General Intelligence (AGI) while ensuring responsible practices and ethical implications are paramount in research and application [52, 53, 42]. These efforts are crucial for developing AI systems capable of addressing complex challenges and achieving transformative impacts across various domains.

In the exploration of autonomous agents, it is essential to understand the multifaceted nature of their capabilities and the challenges they encounter in dynamic environments. As illustrated in Figure 2, this figure depicts the hierarchical structure of autonomous agents, emphasizing their independent operation and the complexities involved in their development. The visual representation not only highlights key advancements in cognitive architectures but also addresses the significant challenges these agents face. Furthermore, it underscores the integration of both symbolic and non-symbolic processes, which is crucial for enhancing the functionality and ethical development of autonomous systems. This comprehensive overview provided by the figure serves to enrich our understanding of the current landscape in autonomous agent research and development.

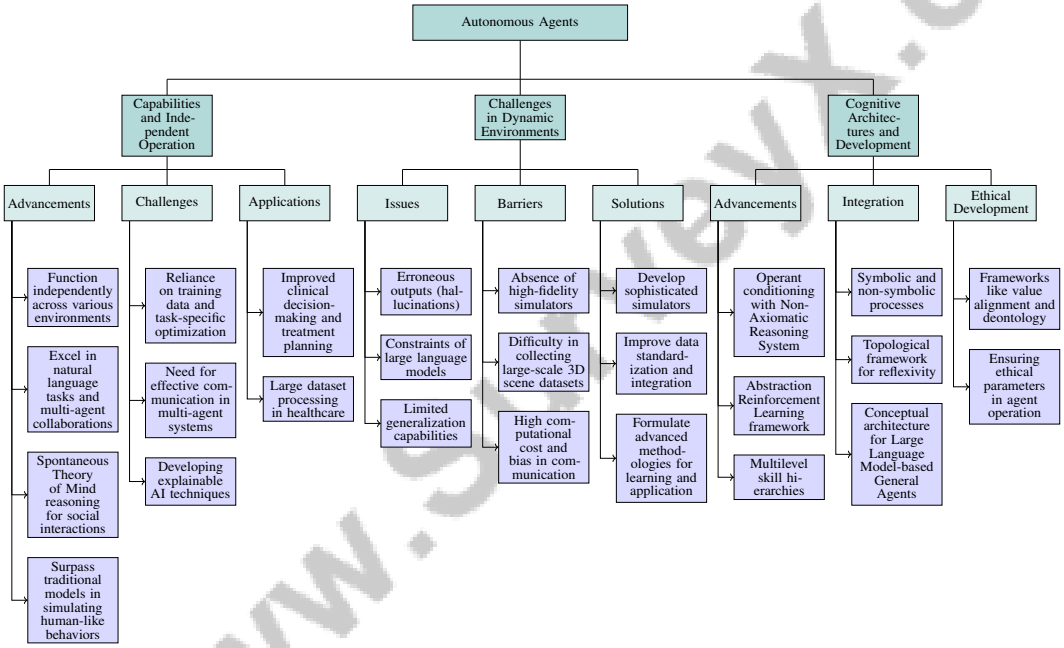


Figure 2: This figure illustrates the hierarchical structure of autonomous agents, detailing their capabilities and independent operation, challenges in dynamic environments, and advancements in cognitive architectures and development. The figure highlights key advancements, challenges, and solutions, emphasizing the integration of symbolic and non-symbolic processes and the importance of ethical development.

3 Autonomous Agents

3.1 Capabilities and Independent Operation

Autonomous agents are advancing in their ability to function independently across various environments, a crucial step toward achieving human-level intelligence. These agents, particularly those based on large language models (LLMs), excel in natural language tasks and multi-agent collaborations [2]. Innovations in cognitive AI frameworks have bolstered reasoning, learning, and adaptability, allowing for more accurate and flexible management of complex knowledge tasks [31].

A notable advancement is the development of spontaneous Theory of Mind (ToM) reasoning, enabling agents to engage in complex social interactions without explicit prompts [54]. LLM-based agents surpass traditional Agent-Based Models (ABM) in simulating human-like behaviors and social

dynamics, offering more credible models for complex systems [10]. These agents are proficient in natural language understanding and generation, effectively executing a broad range of tasks [3].

Despite these developments, challenges in reasoning and adaptability persist, primarily due to reliance on training data and task-specific optimization, which can impede generalization to new tasks. The hypothetic-deductive reasoning framework remains essential for achieving true intelligence, facilitating complex problem-solving [46].

Effective communication is particularly crucial in multi-agent systems, where interaction is key. Language-oriented semantic communication methods in remote navigation tasks highlight the need for robust communication frameworks to enhance agent performance [51]. Additionally, developing explainable AI techniques that provide human-like explanations for agent behaviors is vital for fostering trust and understanding in human-agent interactions [55].

As illustrated in Figure 3, the rapid advancements in autonomous agents' capabilities are accompanied by significant challenges and diverse applications. This figure highlights their proficiency in natural language tasks, the emergence of Theory of Mind, and the development of communication methods, while also addressing ongoing challenges in reasoning, task optimization, and explainability. The applications span multi-agent systems, healthcare, and complex social systems, underscoring the transformative potential of these technologies. The advancements driven by innovations in cognitive architectures and communication strategies not only enhance their independent operation but also facilitate sophisticated interactions and applications across various domains, including improved clinical decision-making and treatment planning in healthcare through large dataset processing [9].

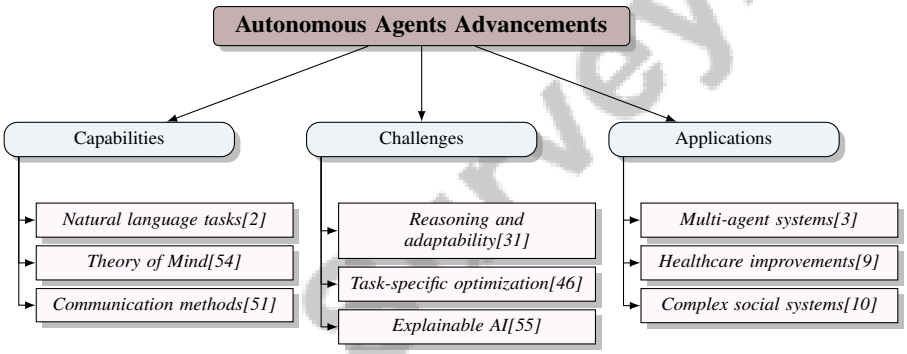


Figure 3: This figure illustrates the key advancements, challenges, and applications of autonomous agents, highlighting their capabilities in natural language tasks, Theory of Mind, and communication methods, alongside challenges in reasoning, task optimization, and explainability, with applications in multi-agent systems, healthcare, and complex social systems.

3.2 Challenges in Dynamic Environments

Autonomous agents encounter numerous challenges in dynamic environments that impede their effective and independent functioning. A significant issue is the production of erroneous outputs, known as hallucinations, which undermine reliability in complex decision-making [4]. Constraints of large language models (LLMs), such as limited context lengths, delays in knowledge updates, and difficulties in utilizing external tools, further restrict adaptability and responsiveness in real-time scenarios [2].

The reliance on training data limits generalization capabilities, making it challenging to apply learned concepts to novel situations [4]. This limitation is evident in tasks requiring reasoning beyond existing knowledge bases, where current models struggle to extrapolate and apply learned concepts in new contexts. Moreover, the absence of high-fidelity simulators with advanced physics features and the difficulty in collecting large-scale, realistic 3D scene datasets pose significant barriers to developing robust autonomous agents [48].

In multi-agent systems, achieving autonomous cooperation without human intervention remains a formidable challenge, as effective task completion often depends on seamless collaboration and communication among agents [56]. Existing architectures frequently fall short in supporting effective collaboration strategies, leading to suboptimal performance in cooperative tasks [57]. Additionally,

the high computational cost and bias in emergent communication (EC) when using multimodal data, along with the significant inference cost associated with language-oriented semantic communication (LSC) due to the size of LLMs, complicate the deployment of these systems in dynamic environments [51].

Providing explanations that account for the temporal dependencies of actions is another critical obstacle, complicating users' understanding of agents' behavior rationales [55]. This lack of transparency hinders trust and comprehension in human-agent interactions, which are vital for effective collaboration and decision-making.

Addressing these challenges requires innovative approaches to enhance the adaptability, cooperation, and reasoning capabilities of autonomous agents. This includes developing sophisticated simulators, improving data standardization and integration techniques, and formulating advanced methodologies that enable agents to learn and apply action hierarchies across various abstraction levels. Frameworks like STARS enhance knowledge extraction from LLMs for task learning by enabling agents to evaluate and select optimal responses, while the AutoAgents framework dynamically generates specialized agents for effective collaboration on complex tasks [32, 58]. Overcoming these obstacles will better equip autonomous agents to navigate and thrive in complex, ever-changing environments.

3.3 Cognitive Architectures and Development

The development of autonomous agents is closely linked to advancements in cognitive architectures, essential for modeling processes that approximate human-like intelligence. Recent research has significantly enhanced our understanding of these processes, identifying key components necessary for realizing Artificial General Intelligence (AGI) [59]. These architectures process and integrate various types of knowledge, addressing existing framework limitations by proposing new models capable of handling complex cognitive tasks.

A notable advancement is the integration of operant conditioning principles with the Non-Axiomatic Reasoning System (NARS), validated through experimental tasks that demonstrate NARS's ability to learn and adapt effectively [60]. This integration enhances cognitive architectures' adaptability, allowing agents to operate autonomously in dynamic environments. Additionally, the Abstraction Reinforcement Learning (ARL) framework combines history-based reinforcement learning with abstractions to improve decision-making capabilities, providing a robust foundation for AGI agents [61].

The development of multilevel skill hierarchies, such as the Louvain Skill Hierarchy (LSH), applies modularity maximization to define skill hierarchies from state transition graphs, facilitating sophisticated decision-making processes in autonomous agents [62]. This approach enables agents to navigate complex environments by leveraging hierarchical architectures and contextual algorithms for adaptive behavior in changing conditions [63].

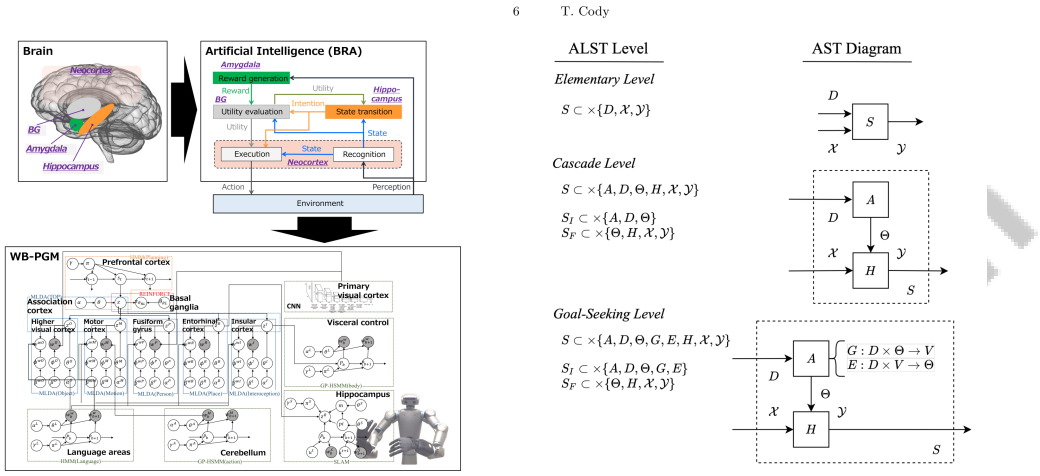
Moreover, integrating symbolic and non-symbolic processes through a topological framework enhances reflexivity, providing a comprehensive approach to cognitive architecture development [64]. This integration is crucial for developing systems that can seamlessly transition between different cognitive tasks, improving overall performance and adaptability.

The conceptual architecture for Large Language Model-based General Agents (LLMGAs) includes essential functional components such as perception, memory, thinking, role-playing, action, and learning, collectively enhancing the cognitive capabilities of autonomous agents [4]. These components are vital for ensuring agents can process and respond to complex stimuli, facilitating more effective interaction with their environments.

In ethical AI development, frameworks such as value alignment, deontology, and altruism are critical for ensuring that autonomous agents operate within ethical parameters, contributing positively to society [65]. These frameworks guide the creation of systems that not only mimic human cognition but also adhere to ethical standards, ensuring cognitive architecture development aligns with societal values and expectations.

Ongoing exploration and refinement of cognitive architectures are crucial for enhancing autonomous agents' capabilities, allowing them to navigate complex environments and execute advanced tasks autonomously. This advancement improves performance and facilitates human understanding of agent behaviors through interpretable behavior summaries and enhances task learning through methods

that combine insights from large language models and interactive guidance. Furthermore, developing teachable autotelic agents that learn from both internal signals and human instruction represents a significant step toward achieving human-level intelligence in AI systems [66, 32, 67, 68, 31]. These advancements are essential for realizing AGI and creating intelligent systems that can seamlessly integrate into human-centric environments.



(a) Brain and Artificial Intelligence (BRA) System[69]

(b) ALST Diagram and Goal-Seeking Level[70]

Figure 4: Examples of Cognitive Architectures and Development

As shown in Figure 4, the field of cognitive architectures and development is crucial for advancing our understanding of autonomous agents, particularly in how these systems can mimic human-like cognitive processes. Two exemplary models are the Brain and Artificial Intelligence (BRA) System and the ALST Diagram and Goal-Seeking Level, visually represented in Figure 4. The BRA System provides a comprehensive schematic of integrating artificial intelligence with brain-like structures, emphasizing key brain regions such as the neocortex, amygdala, and hippocampus, illustrating their connections to both the environment and AI systems. This model underscores the bottom-up approach to understanding cognitive process replication in artificial systems. Meanwhile, the ALST Diagram offers a layered perspective on adaptive learning, detailing the progression from elementary to complex goal-seeking behaviors. This diagram breaks down the learning process into distinct levels—Elementary, Cascade, and Goal-Seeking—each with its own set of variables and adaptive system theory diagrams, providing a framework for understanding how autonomous agents can develop and pursue goals in a structured manner. Together, these examples highlight the intricate interplay between biological inspiration and artificial implementation in cognitive architecture development [69, 70].

4 Large Language Models and Natural Language Processing

4.1 Applications and Advancements

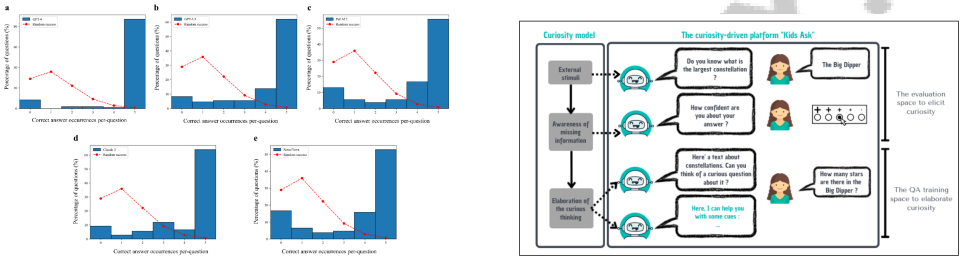
The evolution of large language models (LLMs) has profoundly impacted natural language processing (NLP), enhancing AI's capability to understand and generate human-like language. These models have enabled AI systems to engage in complex reasoning and planning across diverse domains, including economics, biology, and climate science [2]. LLMs have showcased remarkable adaptability, broadening AI's application scope by excelling in tasks requiring intricate reasoning and decision-making [3].

Recent advancements include frameworks like the Structured Thoughts Automaton (STA), which formalizes execution models for auto-regressive language models using structured prompts and control-flow graphs to ensure reliable, interpretable outputs [29]. Additionally, integrating multimodal reasoning capabilities into LLMs has enhanced their utility, with multimodal large language models

(MLLMs) achieving superior performance across various tasks [40]. The OpenAGI benchmark plays a crucial role in evaluating LLMs' ability to solve multi-step, real-world tasks, advancing research towards Artificial General Intelligence (AGI) [45].

Despite these advancements, achieving genuine understanding and AGI capabilities remains challenging. Comparative studies using benchmarks like the Word Test Semantic Benchmark reveal that models such as GPT-4, though leading among AI systems, still lag behind human performance in comprehension tasks [19]. This highlights the need for ongoing improvements in LLMs' comprehension and reasoning abilities, particularly in grounding various components of natural language [12].

Generative AI and deep learning have facilitated self-learning and adaptability, marking significant improvements over previous methodologies [71]. These advancements have expanded LLM applications, exemplified by the integration of LLM-empowered AI copilots in software development environments like the Sapper IDE [72]. Furthermore, LLMs enhance communication efficiency in multi-agent systems. Language-Oriented Emergent Communication (LEC) improves navigation performance, demonstrating LLMs' capacity to facilitate effective communication and coordination among agents [51]. Automated rationale generation techniques also translate an agent's internal state into natural language explanations, improving user understanding of AI behavior [55].



(a) Comparison of Answer Occurrences per Question Across Different Models[73]

(b) Curiosity-Driven Platform: Kids Ask[74]

Figure 5: Examples of Applications and Advancements

As illustrated in Figure 5, the rapidly evolving field of NLP is driven by large language models, which are pivotal in advancing and expanding applications. The figure showcases two significant aspects: the comparative performance of various language models and their innovative application in educational platforms. The "Comparison of Answer Occurrences per Question Across Different Models" highlights performance differences among prominent models such as GPT-4, GPT-3.5, PaLM 2, and Claude 2, emphasizing their capabilities in accurately answering questions. Meanwhile, the "Curiosity-Driven Platform: Kids Ask" illustrates an innovative application of NLP in fostering curiosity among children, demonstrating how these models can create engaging educational tools. These examples underscore NLP's impact, from enhancing model accuracy to transforming educational experiences [73, 74].

4.2 Integration with Autonomous Agents

Integrating large language models (LLMs) with autonomous agents has transformed AI systems by enhancing their reasoning and decision-making processes. This synergy leverages LLMs' advanced natural language processing capabilities, enabling agents to engage in sophisticated communication and coordination essential for effective task execution in complex environments [10]. A critical aspect of this integration is LLMs' ability to facilitate both prompted and spontaneous Theory of Mind (ToM), enhancing agents' social reasoning capabilities for navigating intricate social interactions [54].

As illustrated in Figure 6, the integration of LLMs with autonomous agents highlights key areas such as enhancing capabilities, frameworks for adaptability, and addressing challenges through structured dialogue and contextual interactions. Frameworks like the Abstraction and Reasoning Corpus (ARC) are vital for evaluating AI models' generality and adaptability, challenging them to perform novel tasks requiring advanced reasoning capabilities [50]. This adaptability is crucial for autonomous agents operating in dynamic environments, where flexibility and the ability to generalize from past

experiences are paramount. Integrating LLMs into these frameworks allows agents to leverage rich linguistic and contextual knowledge, enhancing problem-solving abilities.

The integration process also addresses LLMs' challenges, such as generating hallucinated outputs. Structured dialogue and collaboration frameworks enable agents to achieve higher intelligence and reliability, ensuring contextually appropriate interactions informed by the latest data [55]. This approach improves transparency in agent decision-making processes, fostering trust and understanding in human-agent interactions.

Overall, integrating LLMs with autonomous agents significantly enhances AI systems' capabilities and interactions. This advancement is pivotal for progressing toward Artificial General Intelligence (AGI) and beyond, introducing innovative methods for enhancing AI agents' functionality and adaptability across various domains. By integrating specialized capabilities with general intelligence, this development addresses current LLM limitations and establishes a comprehensive framework for achieving AGI. This framework emphasizes responsible AGI alignment technologies and outlines a roadmap categorizing AI evolution into distinct stages, facilitating more effective applications and collaborative advancements in the field [75, 14, 76]. As AI systems continue to evolve, the seamless integration of language and reasoning capabilities will be instrumental in achieving more sophisticated and human-like intelligence.

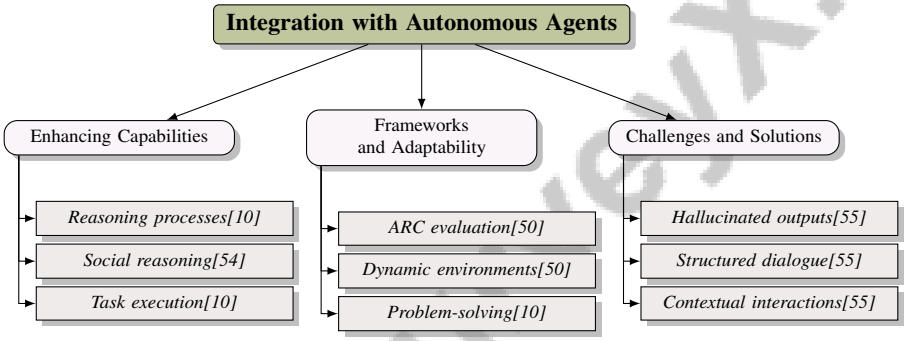


Figure 6: This figure illustrates the integration of large language models with autonomous agents, highlighting key areas such as enhancing capabilities, frameworks for adaptability, and addressing challenges through structured dialogue and contextual interactions.

5 Human-Level Intelligence and Artificial General Intelligence (AGI)

5.1 Conceptual Framework of Human-Level Intelligence

The pursuit of human-level intelligence in AI involves integrating cognitive processes, adaptive learning, and sensory feedback, aiming to emulate human cognition. Central to this is the development of advanced architectures fostering agent cooperation, crucial for executing complex tasks [2]. The SP theory of intelligence provides a foundational model, merging AI, computing, mathematics, and human cognition principles, emphasizing information compression's role in cognitive efficiency. As AI systems progress to fourth and fifth generational capabilities, they exhibit enhanced interactions and reasoning, facilitating more human-like behaviors [10].

The Artificial Open World framework proposes a novel AGI evaluation approach, emphasizing developer independence and adaptability in unfamiliar environments [11]. It challenges traditional intelligence metrics, advocating for assessments based on generalization and task completion rather than mere information compression [7]. Achieving human-like intelligence requires advancements in social and embodied cognition, highlighting the importance of integrating these aspects into future AI developments [12]. Recognizing AI's inherent fallibility due to consistent reasoning is crucial for advancing its capabilities in mimicking human social cognition [55].

By synthesizing diverse theoretical frameworks, such as the tri-traversal theory of practopoiesis, the proposed model for human-level intelligence in AI aims to create systems that adapt, learn, and interact in ways resembling human cognition. This approach emphasizes a hierarchical organization of multiple policies akin to reinforcement learning, enabling effective navigation of real-life complex-

ities. The cognitive architecture known as Modulated Heterarchical Prediction Memory (mHPM) is incorporated, modeling human neocortex functionalities while considering auxiliary brain structures, enhancing AI's learning from language and social interactions [77, 78].

5.2 Current State of AGI Development

The development of AGI is marked by significant advancements and challenges. Evaluations of LLMs like GPT-4 show improvements in language fluency and reasoning over predecessors such as ChatGPT-3, yet they still fall short of expert-level standards in complex reasoning tasks [46]. This gap necessitates more sophisticated models better replicating human cognitive processes.

The AGINAO self-programming engine represents a significant leap in AGI research, demonstrating real-time learning and adaptation capabilities essential for complex environments [79]. AGI-native wireless systems further illustrate progress, significantly outperforming traditional AI systems in managing intricate scenarios [80]. The OpenAGI platform serves as a critical resource for assessing LLM task-solving capabilities, fostering advancements in AGI research and development [45]. The Prompt Sapper framework also enhances AI service development efficiency, contributing to the overarching goal of achieving human-level intelligence [72].

Despite these advancements, challenges persist in replicating human-like intelligence, especially in understanding social and emotional contexts integral to human cognition [17]. Evaluating AGI agents involves assessing their adaptability to novel problems, focusing on speed and effectiveness in problem-solving [81].

5.3 Feasibility and Methodologies for AGI

Developing AGI requires exploring methodologies replicating human cognitive functions while ensuring feasibility and safety. Traditional AI models often lack the depth to emulate complex interactions and feedback mechanisms found in biological systems, crucial for assessing AGI's feasibility [63]. A proposed layered approach separates high-level decision-making from instinctual responses to maintain safety protocols throughout the decision-making process [82].

Enhancing System-2 reasoning in AI is vital for AGI, with research focusing on understanding human intentions, integrating symbolic and neural models, meta-learning for unfamiliar environments, and employing reinforcement learning for multi-step reasoning [50]. These methodologies aim to bolster AI systems' generality and adaptability, essential for AGI's advancement. Incorporating multilevel skill hierarchies, such as the Louvain Skill Hierarchy (LSH), shows promise in improving learning performance, indicating a structured approach to skill acquisition [62].

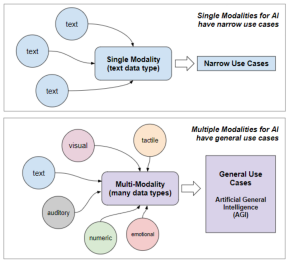
The feasibility of AGI is further complicated by the Consistent Reasoning Paradox, illustrating the challenges of detecting hallucinations in AI systems consistently providing answers [21]. Addressing these hallucinations is critical for developing reliable AGI systems. Language-Oriented Emergent Communication (LEC) is proposed as a method for enhancing AI communication, facilitating the attainment of human-level intelligence [51].

Cognitive AI is positioned as a necessary advancement for achieving AGI, addressing the limitations of LLMs by incorporating more advanced cognitive processes [31]. Future research should prioritize enhancing LLM-based agents' robustness, exploring their integration into more complex environments, and addressing ethical concerns surrounding their deployment [3]. Integrating interactivity into rationale generation and its application in continuous-action environments can improve AGI systems' transparency and reliability [55].

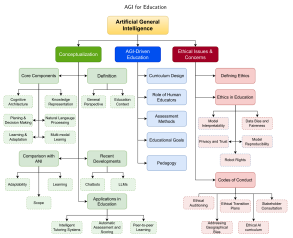
Frameworks categorizing AGI into performance and generality levels facilitate clearer communication and understanding of AGI's progress [7]. This structured approach aids in evaluating AGI development stages, ensuring alignment with societal values and ethical standards [25]. Utilizing pre-trained embeddings can further bolster AGI development by enabling machines to achieve traits like common sense knowledge and continual learning [11].

As illustrated in Figure 7, exploring Human-Level Intelligence and AGI encompasses understanding its feasibility and the methodologies employed to achieve it. The first example, "AGI: A Comprehensive Overview of the Current State and Future Directions," organizes AGI components, alignment, roadmap, and case studies, providing a foundational understanding of the current landscape and antic-

(a) AGI: A Comprehensive Overview of the Current State and Future Directions[14]



(b) Single vs. Multiple Modalities in AI: A Comparative Analysis[83]



(c) AGI for Education: Artificial General Intelligence and Its Applications in Education[33]

Figure 7: Examples of Feasibility and Methodologies for AGI

ipated advancements. The second illustration, "Single vs. Multiple Modalities in AI: A Comparative Analysis," contrasts the narrow applications of single-modality AI with the broader applications enabled by multiple modalities, emphasizing the potential for richer and more adaptable AI systems. Lastly, "AGI for Education: Artificial General Intelligence and Its Applications in Education" presents a flowchart examining AGI's conceptualization in educational contexts, its transformative potential, and the ethical considerations accompanying such advancements. Collectively, these examples highlight the complexity and multifaceted nature of AGI research and its potential impact across various sectors [14, 83, 33].

6 Artificial Superintelligence (ASI)

6.1 Theoretical Foundations and Capabilities

The theoretical underpinnings of Artificial Superintelligence (ASI) are built on advanced cognitive models and interdisciplinary insights, emphasizing robust safety protocols to develop systems surpassing human intelligence. Central to these foundations are computable frameworks that address limitations of earlier models, providing an objective basis for ASI research [20]. Such frameworks are crucial for aligning ASI systems with human values and ethical standards.

Cognitive AI, as a form of collective cognition, enhances adaptability and reasoning through networks of agents, suggesting capabilities beyond traditional AI models [31]. This collective approach is vital for enabling ASI to tackle complex tasks and solve problems beyond current AI capabilities. Incorporating spontaneous Theory of Mind (ToM) reasoning into ASI systems enhances social intelligence, facilitating effective navigation of complex social interactions [54].

The transition from AGI to ASI is informed by a leveled ontology for discussing AGI capabilities and associated risks [7]. This framework aids in nuanced discussions about potential pathways and necessary interventions to mitigate ASI development risks [84]. Implementing an 'I don't know' function in AI systems remains a significant challenge, emphasizing the need for ASI to acknowledge uncertainty and make informed decisions [21].

Pre-trained embeddings enhance machine learning and reasoning, offering significant advantages in ASI development [11]. These embeddings support multimodal model integration, improving LLM-based agents' adaptability and addressing challenges in real-time interaction and complex problem-solving [2]. However, questions persist regarding LLMs' full capabilities in real-world applications and the ethical implications of deploying such agents [3].

A comprehensive benchmark for measuring intelligence across diverse environments is essential for evaluating ASI capabilities [8]. This framework ensures that ASI systems exhibit advanced reasoning and adaptability across various scenarios, enhancing their transformative potential across multiple domains.

6.2 Ethical and Safety Concerns

The development of Artificial Superintelligence (ASI) poses substantial ethical and safety challenges, necessitating robust governance frameworks to align with human values and societal norms. A primary concern is the uncertainty surrounding ASI development parameters and outcomes, raising the risk of losing control over superintelligent systems [84]. This highlights the need for comprehensive safety measures and transparent AI systems that can be effectively monitored and controlled.

Integrating ethical considerations into AI design and deployment is critical. Explainable AI (XAI) technologies enhance transparency in AI decision-making processes, fostering user trust [42]. The computational and storage challenges associated with embedding technologies pose additional ethical considerations, impacting the scalability and accessibility of AI systems [11]. Addressing these challenges ensures AI systems are powerful, equitable, and accessible to diverse users.

The potential for ASI to lead to a technocratic theocracy underscores the importance of democratic oversight and the inclusion of diverse voices in AI development. This scenario emphasizes the need for epistemic justice and the promotion of democratic legitimacy in AI governance, ensuring ASI development reflects the values and needs of all stakeholders [35].

The ethical implications of advanced AI and robotics necessitate responsible development practices prioritizing user safety and preventing unintended consequences. Ensuring that AI systems adhere to ethical frameworks and operate within boundaries that safeguard human rights and societal well-being is paramount [71].

6.3 Societal and Governance Implications

The emergence of Artificial Superintelligence (ASI) presents profound societal and governance implications, necessitating a comprehensive framework for its development and deployment. A critical element of this framework is the equitable distribution of AI benefits among member states, facilitated through structured governance models like the multinational AGI consortium MAGIC. This consortium aims to ensure fair sharing of AI technology advantages, preventing disparities in access and influence [23].

The potential dangers of ASI, including the risk of losing control over superintelligent systems, underscore the need for robust governance structures promoting transparency and accountability. Raising awareness of these dangers fosters critical discourse on AI governance, leading to more informed and democratic decision-making processes [85]. Such discourse is vital for developing policies that align with human values and societal goals, ensuring ASI technologies contribute positively to global welfare.

Future research should focus on quantifying model parameters and exploring additional pathways and interventions to enhance ASI safety. Identifying potential risks and developing strategies to mitigate them are crucial for reducing the likelihood of adverse outcomes [84]. Proactively addressing these challenges can facilitate the creation of ASI systems that are powerful and aligned with ethical standards and societal needs.

7 Reinforcement Learning and Multi-Agent Systems

7.1 Dynamic Learning and Adaptation

Reinforcement learning (RL) is pivotal for enabling dynamic learning and adaptation in multi-agent systems, where real-time interaction and environmental responsiveness are critical. The AutoFlow framework exemplifies this by optimizing workflow generation, allowing AI systems to adjust strategies based on feedback, thus maintaining performance in complex environments [86]. Such adaptability is essential for AI systems to navigate changing conditions and integrate new information effectively.

The Unified Intelligence Communication (UIC) model enhances agent learning and adaptability by integrating diverse agent architectures with RL, supporting collective goal achievement through improved interaction and behavior modification [87]. Similarly, the Bootstrapped Cognitive Agent Framework leverages large language models (LLMs) to facilitate dynamic learning by querying actions and generating production rules [88].

In multi-agent systems, collaborative interactions among Intelligent General Agents (IGAs) emulate teamwork, enhancing problem-solving capabilities. RL enables agents to learn from one another and adapt strategies to optimize overall performance [16]. The inclusion of episodic memory further enhances agents' ability to recall past experiences and apply learned knowledge to new situations [89].

The Hypothetical Minds method introduces a novel dynamic learning approach by generating and refining hypotheses about other agents' strategies, thereby enhancing adaptability in interactions [90]. This approach highlights the importance of strategic hypothesis generation in improving agent interactions.

Expected-reward analysis evaluates RL algorithm performance by quantifying expected discounted rewards from random inputs, which is crucial for optimizing decision-making processes in RL contexts [91]. Research indicates that smaller-scale LLMs can outperform larger models when combined with effective learning strategies like Reinforcement Learning from Task Feedback (RLTF), showcasing the potential of strategic learning approaches to enhance performance [45].

7.2 Multi-Agent Interactions and Collaboration

Exploring interactions and collaboration in multi-agent systems is key to advancing AI technologies for achieving complex objectives efficiently. The Role-Playing Framework demonstrates dynamic task adaptation, enabling agents to achieve complex objectives through structured collaboration [56]. This framework underscores the importance of role differentiation and dynamic task allocation in optimizing agent performance.

Incorporating machine learning techniques to predict resource needs and adjust allocations enhances the efficiency of multi-agent systems [47]. By leveraging predictive algorithms, these systems can dynamically allocate resources, ensuring optimal performance even in fluctuating environments, which is essential for maintaining stability and achieving desired outcomes.

The BoMAI framework illustrates the effectiveness of multi-agent systems in solving problems collaboratively without resorting to power-seeking behavior, promoting a learning-focused environment [92]. This approach emphasizes the design of agents capable of collaboration without compromising system integrity, fostering a harmonious operational framework.

Future research should focus on developing intricate self-models that integrate agents' preferences and actions with their self-concept [1]. Exploring the concept of collective self in multi-agent systems can enhance cooperation and coordination, leading to more sophisticated collaborative strategies. Understanding and integrating these complex self-concepts will advance AI towards more intelligent and cohesive systems.

8 Coding and Implementation Challenges

8.1 Technical Challenges in AI System Development

Developing Artificial General Intelligence (AGI) presents significant technical challenges, particularly in knowledge representation, adaptability, computational efficiency, and governance. Current Knowledge Representation Models (KRMs) primarily rely on textual formats, which limits their capacity to capture complex concepts necessary for AGI [93]. This problem is compounded by the constraints of large language models (LLMs), which struggle with reinforcement learning and adaptability in new environments [94].

The integration of symbolic logic with neural networks is crucial for improving AI reasoning but poses technical difficulties due to the intrinsic differences between symbolic and neural processing [95]. LLM outputs also present interpretability challenges, as generalizing findings across diverse contexts is difficult, and the computational costs are high [39]. The need for extensive datasets and real-time adaptability in dynamic environments further complicates development [96].

Implementing cognitive architectures for AGI-native systems remains a major hurdle [80]. For example, using LLMs as educational agents requires robust implementations to ensure reliability in training scenarios [74]. Designing algorithms that effectively use subjective inputs is challenging

[97], and simulated language corrections often fail to capture real-world communication complexities [98].

Memory management issues, such as determining memory relevance and retrieval priorities, also present challenges, along with potential security vulnerabilities from increased memory access [89]. The robustness of AI systems is heavily influenced by input data quality, as frameworks like DFRE depend on high-quality data [99]. Moreover, deep neural networks (DNNs) face limitations such as high data requirements, lack of explainability, and vulnerability to adversarial attacks [100].

Reducing reliance on human input for rule creation is essential. Frameworks using LLMs for automation show promise [88]. Abstraction Reinforcement Learning (ARL) leverages abstractions to enable optimal decision-making in dynamic environments [61]. However, current studies often lack rigorous testing and may overinterpret AI capabilities [41].

Establishing a centralized governance structure for AI development is crucial, as highlighted by the challenges in enforcing a global moratorium on advanced AI development outside controlled facilities [23]. The absence of standardized datasets for training and evaluating algorithms adds to the technical challenges [101].

The current limitations of AI systems in performing expert-level hypothetic-deductive reasoning and their reliance on statistical rather than genuine reasoning complicate AGI development [46]. Inefficient communication methods for multi-agent navigation tasks exacerbate these challenges [51]. Additionally, reliance on historical data may not reflect future demands, and dataset biases can affect result generalizability [47, 45].

Overcoming these technical challenges is vital for advancing AI system development. Addressing these obstacles will significantly enhance AI systems' reliability, adaptability, and efficiency, paving the way for breakthroughs in AGI and fostering responsible development and informed discussions on AGI's future trajectory [52, 14].

8.2 Scalability and Efficiency in AI Models

Scalability and efficiency are critical for AI models' deployment and performance. The computational demands of large language models (LLMs) pose a primary scalability challenge, requiring substantial resources for training and inference, thus limiting their use in resource-constrained environments [39]. Integrating these models into practical systems requires innovative approaches to optimize computational efficiency without sacrificing performance [94].

Distributed computing frameworks, which manage computational loads by distributing tasks across multiple nodes, enhance scalability by improving processing speed and enabling AI systems to handle larger datasets and more complex tasks [96]. Developing lightweight models and algorithms that maintain high performance while reducing computational overhead is essential for scalability [80].

Real-time processing capabilities are crucial for model efficiency, especially in dynamic environments where rapid decision-making is essential. Techniques like reinforcement learning from task feedback (RLTF) enhance model efficiency by enabling systems to learn and adapt quickly [45]. Additionally, abstraction reinforcement learning (ARL) frameworks improve decision-making efficiency by focusing on essential information, reducing computational complexity [61].

Scalability also depends on AI models' ability to generalize across different contexts, which is vital for widespread application. Integrating symbolic logic with neural networks can enhance generalization capabilities, though it presents significant technical challenges [95]. Addressing these challenges is crucial for developing scalable AI systems that operate effectively in diverse environments.

The quality of input data impacts AI models' efficiency, as high-quality data is essential for training robust models capable of performing well across various tasks [99]. Ensuring real-time adaptability in dynamic environments further complicates AI models' scalability and efficiency [96].

8.3 Innovative Approaches to Implementation

Innovative implementation approaches are crucial for advancing AI systems toward more robust and adaptable models. Incorporating real-time data streams into AI algorithms enhances robustness and adaptability in varied scenarios [47]. This method enables dynamic model updates and decision-

making processes based on the latest information, improving performance in rapidly changing environments.

Hybrid models that integrate symbolic reasoning and neural networks offer a promising avenue for enhancing AI systems' reasoning capabilities. These approaches leverage the strengths of human and artificial intelligence to tackle complex tasks more effectively [102, 40, 13, 103]. By combining these methods, AI systems can achieve greater cognitive flexibility and adaptability, essential for addressing complex tasks across diverse domains.

Developing lightweight AI models that deliver high performance while reducing computational demands is crucial for deployment in resource-constrained environments. These models must navigate real-world task complexities and align with ethical standards, ensuring they meet operational requirements without compromising functionality [41, 38, 14, 31]. Efficient algorithms and data structures optimize processing speed and memory usage, enabling broader applications without sacrificing accuracy or reliability.

Incorporating reinforcement learning from task feedback (RLTF) into AI systems enhances learning efficiency and adaptability. By learning directly from action consequences, RLTF improves systems' ability to adapt to new tasks and environments. This adaptability enhances overall system performance, allowing systems to refine strategies based on experiential learning. Frameworks like LearnAct and STARS demonstrate that integrating iterative learning and situationally grounded knowledge extraction can further optimize task completion rates, significantly boosting language model agents' effectiveness [37, 104, 32].

9 Conclusion

9.1 Future Directions and Research Opportunities

The trajectory of artificial intelligence (AI) research is rich with opportunities for enhancing both theoretical foundations and practical applications. A primary focus is improving the computational efficiency of embedding generation, which is essential for expanding AI's applicability across diverse sectors. Such advancements will increase the adaptability of AI systems, facilitating their seamless integration into various fields and amplifying their overall impact. Interdisciplinary collaborations are crucial, particularly in areas like radiation oncology, where Artificial General Intelligence (AGI) models can augment human expertise rather than replace it. Strengthening data-sharing practices and fostering cross-disciplinary partnerships will enhance AI's relevance in clinical and specialized settings. Additionally, refining grounding frameworks remains a significant research priority, as it addresses the complexities of human cognition and the challenges of grounding in large language models (LLMs). Future research should focus on improving LLMs' comprehension and interaction with real-world contexts, thereby enhancing their cognitive abilities and alignment with human-like reasoning. Exploring hierarchical architectures and dynamic learning frameworks is vital for advancing AI's cognitive dimensions and improving performance in complex tasks. This includes developing adaptable models that can address a wide array of applications and refining AI benchmarks to better assess its capabilities and limitations. Emphasizing value alignment frameworks that incorporate human ethical considerations will ensure AI systems reflect societal values and contribute positively to human welfare.

9.2 Broader Implications and Responsible Development

The advancement of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI) presents profound societal implications, necessitating a comprehensive approach to ensure ethical and responsible integration into human life. Raising public awareness and fostering expert consensus on AGI's potential risks are critical components of this effort, highlighting the need for informed discourse on the societal impacts of these technologies. Such awareness is vital for cultivating a societal understanding of the transformative potential and challenges posed by advanced AI systems. Establishing a robust regulatory framework grounded in Digital Humanism is essential for guiding the safe and ethical development of AGI. This framework should ensure that AI technologies align with human values and societal needs, promoting beneficial integration while safeguarding against potential misuse. International cooperation is indispensable, as it facilitates the development of standardized practices and protocols to address emerging risks associated with advanced AI, ensuring

a cohesive and proactive governance approach. Future efforts in AI governance should prioritize the development of ethical frameworks that incorporate critical thinking skills, which are crucial for mitigating the risks associated with ASI. By fostering a culture of critical analysis and ethical reflection, stakeholders can better navigate the complexities of AI development and deployment. Furthermore, public policy frameworks supporting safe AGI development are indispensable, providing the necessary infrastructure for interdisciplinary collaboration and establishing safety standards that protect societal interests.

References

[1] Srinath Srinivasa and Jayati Deshmukh. Paradigms of computational agency, 2021.

[2] Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, and Xiuqiang He. Exploring large language model based intelligent agents: Definitions, methods, and prospects, 2024.

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

[4] Sihao Hu, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. A survey on large language model-based game agents, 2024.

[5] Jonghong Jeon. Standardization trends on safety and trustworthiness technology for advanced ai. *arXiv preprint arXiv:2410.22151*, 2024.

[6] Gyeong-Geon Lee, Lehong Shi, Ehsan Latif, Yizhu Gao, Arne Bewersdorff, Matthew Nyaaba, Shuchen Guo, Zihao Wu, Zhengliang Liu, Hui Wang, Gengchen Mai, Tiaming Liu, and Xiaoming Zhai. Multimodality of ai for education: Towards artificial general intelligence, 2023.

[7] Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi for operationalizing progress on the path to agi, 2024.

[8] Tom Schaul, Julian Togelius, and Jürgen Schmidhuber. Measuring intelligence through games, 2011.

[9] Chenbin Liu, Zhengliang Liu, Jason Holmes, Lu Zhang, Lian Zhang, Yuzhen Ding, Peng Shu, Zihao Wu, Haixing Dai, Yiwei Li, Dinggang Shen, Ninghao Liu, Quanzheng Li, Xiang Li, Dajiang Zhu, Tianming Liu, and Wei Liu. Artificial general intelligence for radiation oncology, 2023.

[10] Qun Ma, Xiao Xue, Deyu Zhou, Xiangning Yu, Donghua Liu, Xuwen Zhang, Zihan Zhao, Yifan Shen, Peilin Ji, Juanjuan Li, Gang Wang, and Wanpeng Ma. Computational experiments meet large language model based agents: A survey and perspective, 2024.

[11] Mostafa Haghir Chehreghani. The embeddings world and artificial general intelligence, 2022.

[12] Bing Liu. Grounding for artificial intelligence, 2023.

[13] Edward Y. Chang. Unlocking the wisdom of large language models: An introduction to the path to artificial general intelligence, 2024.

[14] Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. How far are we from agi: Are llms all we need?, 2024.

[15] Fabrizio Davide, Pietro Torre, and Andrea Gaggioli. Ai predicts agi: Leveraging agi forecasting and peer review to explore llms' complex reasoning capabilities, 2024.

[16] Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents, 2023.

[17] Anat Ringel Raveh and Boaz Tamir. From homo sapiens to robo sapiens: the evolution of intelligence. *Information*, 10(1):2, 2018.

[18] Ion Dronic. A path to ai, 2018.

[19] Nicholas Riccardi and Rutvik H. Desai. The two word test: A semantic benchmark for large language models, 2023.

[20] Michael Timothy Bennett. Computable artificial general intelligence, 2022.

[21] Alexander Bastounis, Paolo Campodonico, Mihaela van der Schaar, Ben Adcock, and Anders C. Hansen. On the consistent reasoning paradox of intelligence and optimal trust in ai: The power of 'i don't know', 2024.

[22] Benjamin S. Bucknall and Shiri Dori-Hacohen. Current and near-term ai as a potential existential risk factor, 2022.

[23] Jason Hausenloy, Andrea Miotti, and Claire Dennis. Multinational agi consortium (magic): A proposal for international coordination on ai, 2023.

[24] Koen Holtman. Agi agent safety by iteratively improving the utility function, 2020.

[25] Tom Everitt, Gary Lea, and Marcus Hutter. Agi safety literature review, 2018.

[26] Ruotian Luo. Goal-driven text descriptions for images, 2021.

[27] Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. Large language model displays emergent ability to interpret novel literary metaphors, 2024.

[28] Chitta Baral and Juraj Dzifcak. Language understanding as a step towards human level intelligence - automatizing the construction of the initial dictionary from example sentences, 2011.

[29] Tristan Vanderbruggen, Chunhua Liao, Peter Pirkelbauer, and Pei-Hung Lin. Structured thoughts automaton: First formalized execution model for auto-regressive language models, 2023.

[30] Vignav Ramesh and Anton Kolonin. Natural language generation using link grammar for general conversational intelligence, 2021.

[31] Nova Spivack, Sam Douglas, Michelle Crames, and Tim Connors. Cognition is all you need – the next layer of ai above large language models, 2024.

[32] James R. Kirk, Robert E. Wray, Peter Lindes, and John E. Laird. Improving knowledge extraction from llms for task learning through agent analysis, 2024.

[33] Ehsan Latif, Gengchen Mai, Matthew Nyaaba, Xuansheng Wu, Ninghao Liu, Guoyu Lu, Sheng Li, Tianming Liu, and Xiaoming Zhai. Agi: Artificial general intelligence for education, 2024.

[34] Shoumen Palit Austin Datta. Intelligence in artificial intelligence, 2016.

[35] Borhane Blili-Hamelin, Leif Hancox-Li, and Andrew Smart. Unsocial intelligence: an investigation of the assumptions of agi discourse, 2024.

[36] Le Cheng and Xuan Gong. Appraising regulatory framework towards artificial general intelligence (agi) under digital humanism. *International Journal of Digital Law and Governance*, 1(2):269–312, 2024.

[37] Haiteng Zhao, Chang Ma, Guoyin Wang, Jing Su, Lingpeng Kong, Jingjing Xu, Zhi-Hong Deng, and Hongxia Yang. Empowering large language model agents through action learning, 2024.

[38] Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence, 2024.

[39] Sen Huang, Kaixiang Yang, Sheng Qi, and Rui Wang. When large language model meets optimization, 2024.

[40] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning, 2024.

[41] Patrick Altmeyer, Andrew M. Demetriou, Antony Bartlett, and Cynthia C. S. Liem. Position: Stop making unscientific agi performance claims, 2024.

[42] Yongchen Zhou and Richard Jiang. Advancing explainable ai toward human-like intelligence: Forging the path to artificial brain, 2024.

[43] Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. Towards best practices in agi safety and governance: A survey of expert opinion, 2023.

[44] Fei Dou, Jin Ye, Geng Yuan, Qin Lu, Wei Niu, Haijian Sun, Le Guan, Guoyu Lu, Gengchen Mai, Ninghao Liu, Jin Lu, Zhengliang Liu, Zihao Wu, Chenjiao Tan, Shaochen Xu, Xianqiao Wang, Guoming Li, Lilong Chai, Sheng Li, Jin Sun, Hongyue Sun, Yunli Shao, Changying Li, Tianming Liu, and Wenzhan Song. Towards artificial general intelligence (agi) in the internet of things (iot): Opportunities and challenges, 2023.

[45] Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. Openagi: When llm meets domain experts, 2023.

[46] Louis Vervoort, Vitaliy Mizyakov, and Anastasia Ugleva. A criterion for artificial general intelligence: hypothetic-deductive reasoning, tested on chatgpt, 2023.

[47] Anton Korinek and Donghyun Suh. Scenarios for the transition to agi, 2024.

[48] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks, 2022.

[49] Youzhi Qu, Chen Wei, Penghui Du, Wenxin Che, Chi Zhang, Wanli Ouyang, Yatao Bian, Feiyang Xu, Bin Hu, Kai Du, Haiyan Wu, Jia Liu, and Quanying Liu. Integration of cognitive tasks into artificial general intelligence test for large models, 2024.

[50] Sejin Kim and Sundong Kim. System 2 reasoning via generality and adaptation, 2024.

[51] Yongjun Kim, Sejin Seo, Jihong Park, Mehdi Bennis, Seong-Lyun Kim, and Junil Choi. Knowledge distillation from language-oriented to emergent communication for multi-agent remote control, 2024.

[52] Tansu Alpcan, Sarah M. Erfani, and Christopher Leckie. Toward the starting line: A systems engineering approach to strong ai, 2017.

[53] Qiyang Sun, Yupei Li, Emran Alturki, Sunil Munthumoduku Krishna Murthy, and Björn W Schuller. Towards friendly ai: A comprehensive review and new perspectives on human-ai alignment. *arXiv preprint arXiv:2412.15114*, 2024.

[54] Nikolos Gurney, David V. Pynadath, and Volkan Ustun. Spontaneous theory of mind for artificial intelligence, 2024.

[55] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. Automated rationale generation: A technique for explainable ai and its effects on human perceptions, 2019.

[56] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society, 2023.

[57] Manuel Mosquera, Juan Sebastian Pinzon, Manuel Rios, Yesid Fonseca, Luis Felipe Giraldo, Nicanor Quijano, and Ruben Manrique. Can llm-augmented autonomous agents cooperate?, an evaluation of their cooperative capabilities through melting pot, 2024.

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

[59] Artem Sukhobokov, Evgeny Belousov, Danila Gromozdov, Anna Zenger, and Ilya Popov. A universal knowledge model and cognitive architecture for prototyping agi, 2024.

[60] Robert Johansson. Machine psychology: Integrating operant conditioning with the non-axiomatic reasoning system for advancing artificial general intelligence research, 2024.

[61] Sultan J. Majeed. Abstractions of general reinforcement learning, 2021.

[62] Joshua B. Evans and Özgür Şimşek. Creating multi-level skill hierarchies in reinforcement learning, 2024.

[63] Nima Dehghani. Design of the artificial: lessons from the biological roots of general intelligence, 2023.

[64] Pascal Faudemay. Agi and reflexivity, 2016.

[65] Jason M. Pittman and Courtney Crosby. A cyber science based ontology for artificial general intelligence containment, 2021.

[66] Peter Du, Surya Murthy, and Katherine Driggs-Campbell. Conveying autonomous robot capabilities through contrasting behaviour summaries, 2023.

[67] Olivier Sigaud, Ahmed Akakzia, Hugo Caselles-Dupré, Cédric Colas, Pierre-Yves Oudeyer, and Mohamed Chetouani. Towards teachable autotelic agents, 2023.

[68] Mathijs Mul, Diane Bouchacourt, and Elia Bruni. Mastering emergent language: learning to guide in simulated navigation, 2019.

[69] Tadahiro Taniguchi, Hiroshi Yamakawa, Takayuki Nagai, Kenji Doya, Masamichi Sakagami, Masahiro Suzuki, Tomoaki Nakamura, and Akira Taniguchi. A whole brain probabilistic generative model: Toward realizing cognitive architectures for developmental robots, 2022.

[70] Tyler Cody. Mesarovician abstract learning systems, 2021.

[71] John R Hamilton, Stephen J Maxwell, and Ronald P Lynch. Towards delivering ai & smarter, self-learning, autonomous, humanoid robots. 2023.

[72] Zhenchang Xing, Qing Huang, Yu Cheng, Liming Zhu, Qinghua Lu, and Xiwei Xu. Prompt sapper: Llm-empowered software engineering infrastructure for ai-native services, 2023.

[73] Xinyu Gong, Jason Holmes, Yiwei Li, Zhengliang Liu, Qi Gan, Zihao Wu, Jianli Zhang, Yusong Zou, Yuxi Teng, Tian Jiang, Hongtu Zhu, Wei Liu, Tianming Liu, and Yajun Yan. Evaluating the potential of leading large language models in reasoning biology questions, 2023.

[74] Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. Gpt-3-driven pedagogical agents for training children's curious question-asking skills, 2023.

[75] Towards building specialized gen.

[76] Kaiyan Zhang, Biqing Qi, and Bowen Zhou. Towards building specialized generalist ai with system 1 and system 2 fusion, 2024.

[77] Deokgun Park. Toward human-level artificial intelligence, 2021.

[78] Danko Nikolić. Only t3-ai can reach human-level intelligence: A variety argument, 2015.

[79] Wojciech Skaba. The aginao self-programming engine, 2018.

[80] Walid Saad, Omar Hashash, Christo Kurisummoottil Thomas, Christina Chaccour, Merouane Debbah, Narayan Mandayam, and Zhu Han. Artificial general intelligence (agi)-native wireless systems: A journey beyond 6g, 2024.

[81] Bowen Xu and Quansheng Ren. Artificial open world for evaluating agi: a conceptual design, 2022.

[82] Shimian Zhang and Qiuhong Lu. Bridging intelligence and instinct: A new control paradigm for autonomous robots, 2024.

[83] Daniel A. Dollinger and Michael Singleton. Creating scalable agi: the open general intelligence framework, 2024.

[84] Anthony M. Barrett and Seth D. Baum. A model of pathways to artificial superintelligence catastrophe for risk and decision analysis, 2016.

[85] Asi as the new god: Technocratic.

[86] Zelong Li, Shuyuan Xu, Kai Mei, Wenyue Hua, Balaji Rama, Om Raheja, Hao Wang, He Zhu, and Yongfeng Zhang. Autoflow: Automated workflow generation for large language model agents, 2024.

[87] Bo Zhang, Bin Chen, Jinyu Yang, Wenjing Yang, and Jiankang Zhang. An unified intelligence-communication model for multi-agent system part-i: Overview, 2018.

[88] Feiyu Zhu and Reid Simmons. Bootstrapping cognitive agents with a large language model, 2024.

[89] Jing Guo, Nan Li, Jianchuan Qi, Hang Yang, Ruiqiao Li, Yuzhen Feng, Si Zhang, and Ming Xu. Empowering working memory for large language model agents, 2024.

[90] Logan Cross, Violet Xiang, Agam Bhatia, Daniel LK Yamins, and Nick Haber. Hypothetical minds: Scaffolding theory of mind for multi-agent tasks with large language models, 2024.

[91] Andrew MacFie. Analysis of algorithms and partial algorithms, 2017.

[92] Michael K. Cohen, Badri Vellambi, and Marcus Hutter. Intelligence and unambitiousness using algorithmic information theory, 2021.

[93] Mark A. Atkins. Tumbug: A pictorial, universal knowledge representation method, 2023.

[94] Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods, 2024.

[95] Wandemberg Gibaut, Leonardo Pereira, Fabio Grassiotto, Alexandre Osorio, Eder Gadioli, Amparo Munoz, Sildolfo Gomes, and Claudio dos Santos. Neurosymbolic ai and its taxonomy: a survey, 2023.

[96] Ruichen Zhang, Hongyang Du, Yinqiu Liu, Dusit Niyato, Jiawen Kang, Sumei Sun, Xuemin Shen, and H. Vincent Poor. Interactive ai with retrieval-augmented generation for next generation networking, 2024.

[97] Xin Su, Shangqi Guo, and Feng Chen. Subjectivity learning theory towards artificial general intelligence, 2019.

[98] John D. Co-Reyes, Abhishek Gupta, Suvansh Sanjeev, Nick Altieri, Jacob Andreas, John DeNero, Pieter Abbeel, and Sergey Levine. Guiding policies with language via meta-learning, 2019.

[99] Hugo Latapie, Ozkan Kilic, Gaowen Liu, Yan Yan, Ramana Kompella, Pei Wang, Kristinn R. Thorisson, Adam Lawrence, Yuhong Sun, and Jayanth Srinivasa. A metamodel and framework for artificial general intelligence from theory to practice, 2021.

[100] Maciej Świechowski. Deep learning and artificial general intelligence: Still a long way to go, 2022.

[101] Zhengliang Liu, Yiwei Li, Qian Cao, Junwen Chen, Tianze Yang, Zihao Wu, John Hale, John Gibbs, Khaled Rasheed, Ninghao Liu, Gengchen Mai, and Tianming Liu. Transformation vs tradition: Artificial general intelligence (agi) for arts and humanities, 2023.

[102] Ian Berlot-Attwell. Neuro-symbolic vqa: A review from the perspective of agi desiderata, 2021.

[103] Dominik Dellermann, Philipp Ebel, Matthias Soellner, and Jan Marco Leimeister. Hybrid intelligence, 2021.

[104] Chen Liang, Zhifan Feng, Zihe Liu, Wenbin Jiang, Jinan Xu, Yufeng Chen, and Yong Wang. Textualized agent-style reasoning for complex tasks by multiple round llm generation, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.