AI Scientists and Autonomous Research: A Survey

www.surveyx.cn

Abstract

This survey paper explores the transformative role of AI scientists, focusing on the integration of large language models (LLMs) and multi-agent systems in advancing autonomous research. AI scientists leverage sophisticated AI technologies to autonomously conduct research, enhancing data processing, hypothesis generation, and decision-making across domains such as education, healthcare, and robotics. The integration of LLMs with frameworks like Answer Set Programming highlights their potential to improve reasoning and generalization. Multi-agent systems further enhance collaboration and communication among AI agents, optimizing task execution and enabling efficient scientific inquiry. Despite challenges related to ethical considerations and the need for continuous learning, developing inclusive and transparent AI systems remains crucial. Future research directions emphasize refining AI models to enhance interpretability, adaptability, and efficiency, expanding their applicability across diverse scientific fields. By addressing these challenges and leveraging intelligent systems' capabilities, AI scientists are poised to significantly accelerate scientific breakthroughs and expand knowledge frontiers, ushering in a new era of AI-driven scientific discovery.

1 Introduction

1.1 Concept of AI Scientists

AI scientists are advanced systems engineered to autonomously conduct research by leveraging sophisticated artificial intelligence technologies to generate insights and automate complex tasks. These cognitive agents integrate large language models (LLMs) with cognitive architectures, enhancing performance across diverse environments [1]. By utilizing LLMs, AI scientists can perform autonomous research and data analysis, overcoming the limitations of traditional human-led methods.

Defined as systems employing advanced algorithms and models, AI scientists autonomously decompose complex tasks in fields such as robotics, thereby improving execution capabilities [2]. They can generate an extensive array of complex and novel behaviors without predetermined goals or constraints [3]. Additionally, AI scientists automate tasks traditionally performed by humans, including discourse analysis and the formalization of mathematical statements.

Beyond task automation, AI scientists are capable of executing complex activities autonomously, such as strategic negotiation and alliance-building in the Diplomacy game [4]. They combine computational experiments with LLM-based agents to model intricate social systems, showcasing their transformative potential in autonomous research [5]. Through these diverse applications, AI scientists represent a significant advancement in scientific discovery, broadening the scope of autonomous inquiry and enhancing research efficiency across multiple domains.

1.2 Significance of Integrating AI Technologies

The integration of AI technologies, particularly large language models (LLMs), is vital for advancing scientific discovery by improving the efficiency and scope of autonomous research. LLMs, such as Codex, enhance the efficiency and accessibility of mathematical formalization, a critical component

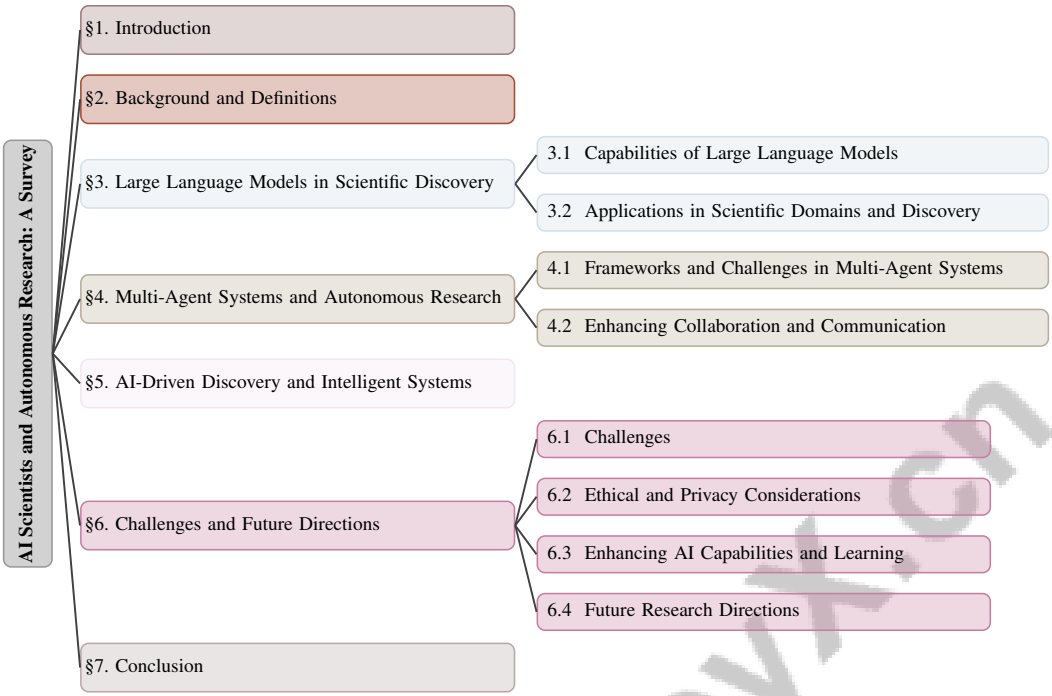


Figure 1: chapter structure

of scientific inquiry [6]. Furthermore, integrating LLMs with optimization algorithms marks a significant advancement in AI research, facilitating more effective problem-solving capabilities [7].

LLMs address challenges related to handling long inputs, essential for preserving contextual information in research [8]. Additionally, structured output constraints from a user-centered perspective are crucial for improving developer workflows and ensuring reliable outcomes [9].

The integration of AI technologies also addresses ethical considerations. The alignment of LLMs often reflects a colonial mindset, imposing Western values as universal and marginalizing other knowledge systems [10]. Hence, developing inclusive AI systems that represent diverse perspectives is imperative.

Moreover, self-evolving agents like Richelieu demonstrate the potential of leveraging experiences in simulated environments to enhance capabilities without relying on human-generated data [4]. This underscores the transformative impact of AI technologies in accelerating scientific breakthroughs and expanding the horizons of autonomous research.

In multi-modal applications, integrating diverse modalities in LLMs is crucial for expanding AI applications across various domains [11], including innovative methods for improving dental diagnosis and treatment through applications like ChatGPT [12].

Collectively, these developments highlight the importance of integrating AI technologies in scientific discovery. The incorporation of fine-tuned LLMs into academic research enhances efficiency and accessibility by automating labor-intensive processes like Systematic Literature Reviews (SLRs) while addressing significant societal and ethical challenges. This advancement promotes methodological transparency and reliability in research, paving the way for more inclusive AI-driven outcomes. By mitigating issues such as LLM hallucination and enhancing the rigor of knowledge synthesis, these tools can prevent the formation of scientific monocultures and support a more innovative research landscape. Furthermore, the emergence of autonomous AI systems capable of conducting independent scientific inquiries signifies a transformative shift in research methodologies, positioning AI as a collaborative partner in addressing complex global challenges [13, 14, 15, 16].

AI-generated, for reference only.

1.3 Objectives and Structure of the Survey

This survey provides a comprehensive analysis of the innovative concept of AI scientists, exploring their potential to autonomously conduct research, generate new hypotheses, and contribute to scientific literature, thereby transforming the landscape of scientific inquiry in the age of artificial intelligence [13, 17, 14, 15, 18]. The primary objectives include elucidating the integration of LLMs and multi-agent systems in scientific discovery and assessing the transformative potential of AI-driven discovery and intelligent systems. The survey highlights the capabilities, applications, and challenges associated with these technologies while addressing ethical and privacy considerations.

The paper is structured to systematically explore each aspect of AI scientists and autonomous research. The introduction defines the concept of AI scientists and discusses the significance of integrating AI technologies in scientific discovery. The background and definitions section provides foundational understanding of key concepts such as AI scientists, LLMs, multi-agent systems, and autonomous research. Following this, a detailed exploration of the role of LLMs in scientific discovery, including their capabilities and applications across various scientific domains, is presented.

Subsequent sections delve into the integration of multi-agent systems in autonomous research, highlighting their frameworks, challenges, and advantages in enhancing collaboration and communication among AI agents. The survey further examines AI-driven discovery and the role of intelligent systems in enabling autonomous research, emphasizing their potential to operate without human intervention and accelerate scientific breakthroughs.

The survey comprehensively addresses the multifaceted challenges and future trajectories in integrating AI scientists and autonomous research systems. It discusses critical ethical considerations, including the implications of AI on scientific integrity and accountability, alongside concerns regarding data privacy and safeguarding sensitive information. Emphasizing the necessity for continuous learning and adaptation within these systems, the discussion aims to mitigate risks, enhance reliability, and foster innovation in research methodologies. This highlights the importance of establishing robust frameworks to ensure responsible AI deployment in the scientific community, ultimately balancing the transformative potential of AI with the need for ethical oversight and methodological rigor [13, 19, 16, 14, 15]. Proposed future research directions aim to enhance AI capabilities and foster innovations in AI-driven discovery, contributing to the advancement of scientific research.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Definitions

AI scientists are autonomous systems engineered to conduct research and make decisions independently, utilizing domain-specific large language models (LLMs) in areas such as medicine. These systems enhance functionality through structured optimization processes and frameworks like the GraphAgent-Reasoner (GAR), which facilitates complex reasoning tasks and advances autonomous research capabilities [20, 21]. LLMs are instrumental in translating informal mathematical content into formal languages, crucial for software correctness verification [6].

LLMs, characterized by their extensive parameter size often exceeding 10 billion, are sophisticated AI systems that generate human-like text and play a crucial role in task decomposition across various domains, including robotics [2]. Their emergent behaviors require a nuanced understanding to optimize their application in scientific discovery. These models enhance decision-making through optimization algorithms and effectively process multi-source data, improving clinical decisions and boosting qualitative research efficiency in educational contexts, such as discourse data analysis using models like GPT-4 [22].

Multi-agent systems facilitate collaborative efforts among AI agents, enabling them to tackle complex tasks and generate hypotheses collectively [23]. These systems are pivotal in refining decision-making processes and designing agents for specific tasks, though challenges like the lack of controllability in LLM outputs can impede development efficiency [9].

Autonomous research refers to AI systems independently conducting scientific investigations, deriving insights, and making discoveries without human intervention. These systems automate various stages of the research process, from ideation to experimentation, using LLMs to enhance efficiency and

decision-making. They are designed to autonomously select and apply algorithms to solve specific problems, integrating algorithmic features with problem instances [14, 24].

In recent years, the emergence of Large Language Models (LLMs) has significantly influenced the landscape of scientific discovery. These models not only enhance the efficiency of data processing but also contribute to innovative problem-solving approaches, thereby integrating seamlessly with existing machine learning frameworks. The applications of LLMs are diverse, spanning fields such as educational research, medicine, and robotics, where they focus on optimization and addressing complex challenges.

To better illustrate this hierarchical structure and the multifaceted capabilities of LLMs, Figure 2 presents a comprehensive overview of their role in enhancing scientific research. This figure categorizes the various applications of LLMs, emphasizing their transformative potential in scientific inquiry. It highlights how these models are not merely tools but pivotal elements that drive innovation in methodologies and facilitate the integration of insights across different scientific domains. By visualizing these relationships, the figure underscores the profound impact of LLMs on the advancement of scientific knowledge and practices.

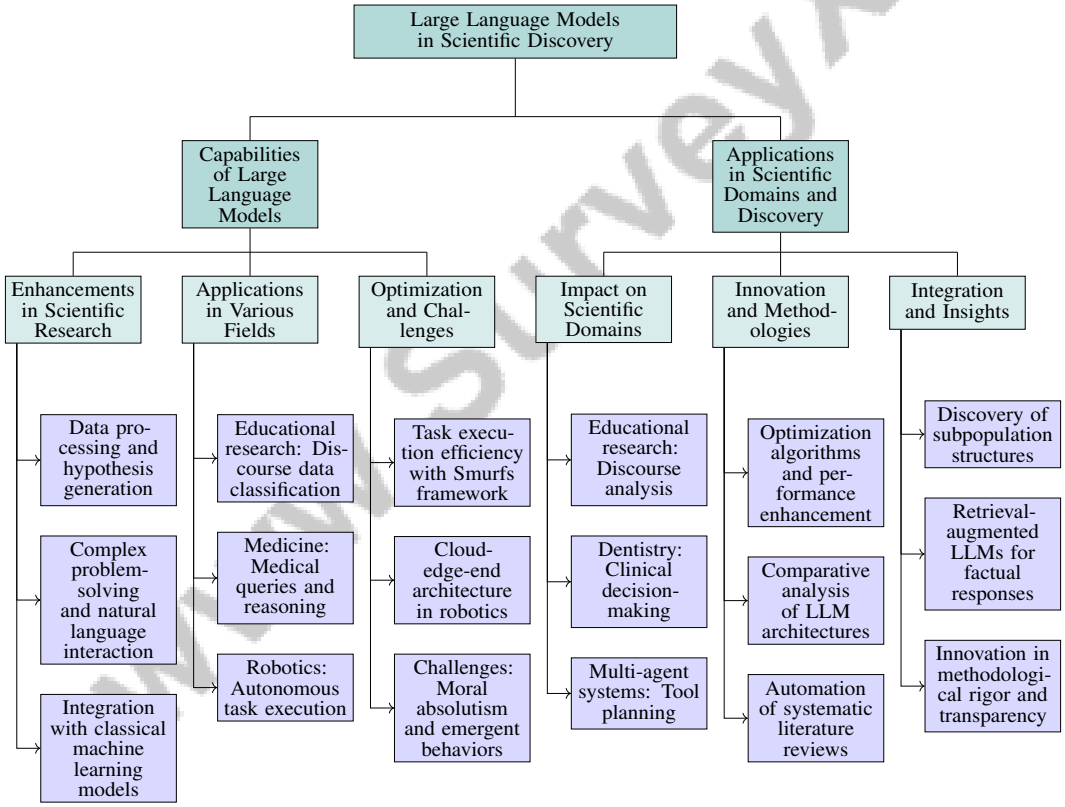


Figure 2: This figure illustrates the hierarchical structure of Large Language Models in Scientific Discovery, categorizing their capabilities and applications. It highlights their role in enhancing scientific research through data processing, problem-solving, and integration with machine learning models. The applications span various fields such as educational research, medicine, and robotics, with a focus on optimization and addressing challenges. Furthermore, the figure demonstrates the impact of LLMs on scientific domains, innovation in methodologies, and integration of insights, showcasing their transformative potential in scientific inquiry.

3 Large Language Models in Scientific Discovery

3.1 Capabilities of Large Language Models

Large Language Models (LLMs) have become pivotal in scientific research, enhancing data processing, hypothesis generation, and complex problem-solving. By enabling natural language interaction, LLMs facilitate intuitive engagement with complex datasets [25]. Their integration with classical machine learning models improves prediction performance across classification tasks, highlighting their scientific versatility [24]. In educational research, GPT-4 excels in classifying discourse data, streamlining qualitative research [22]. In medicine, AntGLM-Med-10B effectively answers medical queries, with fine-tuning enhancing reasoning and task performance [20].

The Smurfs framework optimizes LLMs' task execution efficiency by employing specialized agents for task decomposition and verification [26]. The cloud-edge-end architecture (CEEHA) demonstrates LLMs' potential in robotics by autonomously executing complex tasks [2]. Codex advances mathematical capabilities by transforming natural language statements into formal representations [6]. LLMs' ability to analyze complex graph structures is crucial for scientific research, as seen in scalable and accurate graph reasoning [21]. Optimizing LLMs with various algorithms further enhances their real-world applicability [7].

In healthcare, LLMs improve diagnostic accuracy and treatment planning by analyzing structured and unstructured patient data, particularly in dental applications [12]. Richelieu exemplifies enhanced diplomatic capabilities by adapting strategies based on past experiences, showcasing LLMs' potential in dynamic environments [4]. Despite their capabilities, LLMs face challenges such as moral absolutism, necessitating a deeper understanding of their emergent behaviors for effective scientific application [27].

As illustrated in Figure 3, the diverse capabilities of LLMs span across scientific, educational, medical, and technological domains. This figure categorizes key applications, emphasizing LLMs' roles in data processing, discourse classification, task decomposition, and more. These advancements highlight LLMs' role in expanding autonomous inquiry, fostering interdisciplinary collaboration, and enhancing scientific discovery [11].

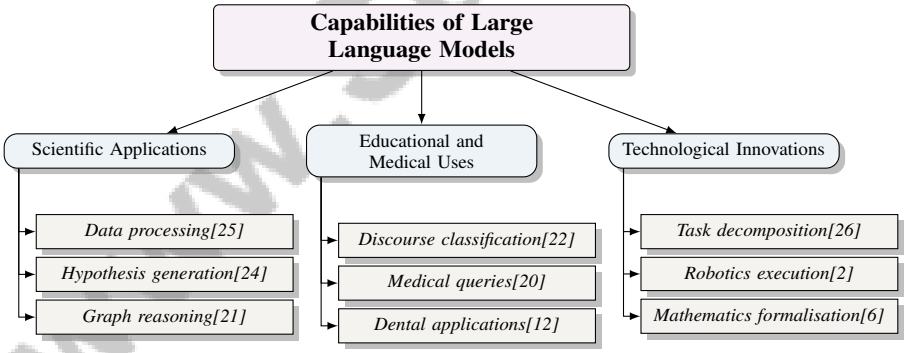


Figure 3: This figure illustrates the diverse capabilities of Large Language Models (LLMs) across scientific, educational, medical, and technological domains. It categorizes key applications, highlighting LLMs' role in data processing, discourse classification, task decomposition, and more.

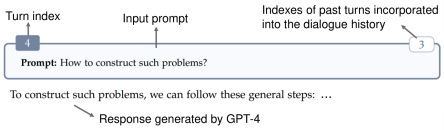
3.2 Applications in Scientific Domains and Discovery

LLMs significantly impact various scientific domains by enhancing data processing and decision-making. In educational research, they improve discourse analysis and automatic deductive coding, offering deeper insights into educational methodologies [22]. In dentistry, the Multi-Modal LLM AI System (MMLLM) enhances clinical decision-making and patient care through treatment plans and diagnostic reports [12]. The Smurfs framework advances LLM applications in multi-agent systems for tool planning, optimizing research processes through specialized agents [26].

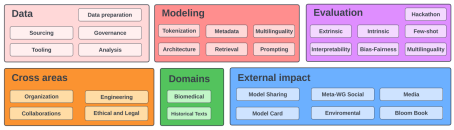
Understanding LLM emergent behaviors is crucial for successful applications across scientific domains, as demonstrated in complex systems analysis [27]. LLMs drive innovation by developing

optimization algorithms that enhance performance and enable diverse scientific applications [7]. Comparative analyses of LLM architectures, such as GPT-3 and T5, provide insights into their effectiveness and applications across scientific domains, emphasizing the importance of selecting appropriate architectures for specific research contexts [11].

The studies illustrate LLMs' transformative influence on scientific methodologies, automating systematic literature reviews and increasing efficiency and accuracy while addressing challenges like LLM hallucination. LLMs facilitate discovering subpopulation structures within datasets, yielding insights crucial for analytical tasks. The development of retrieval-augmented LLMs, such as the RETA-LLM toolkit, showcases their ability to generate factual responses by integrating external information, enriching research across disciplines. This integration fosters innovation and sets new standards for methodological rigor and transparency in scientific inquiry [13, 28, 29, 30].



(a) Dialogue History with Turn Index, Input Prompt, and Response Generated by GPT-4[31]



(b) The image represents a structured framework for understanding and addressing various aspects of machine learning and related fields.[32]

Figure 4: Examples of Applications in Scientific Domains and Discovery

As illustrated in Figure 4, LLMs like GPT-4 play a crucial role in scientific discovery, offering innovative applications across domains. The examples demonstrate their transformative potential, showcasing their ability to engage in complex dialogues and provide structured frameworks for machine learning. The first example highlights GPT-4's capacity to generate insightful responses to scientific queries, maintaining context and continuity in discussions. The second example presents a comprehensive framework for understanding machine learning, segmented into key areas such as data sourcing, modeling, and multilinguality. This structured approach aids in organizing and addressing various facets of machine learning, exemplifying how LLMs facilitate deeper understanding and innovation in scientific research, driving advancements in the field [31, 32].

4 Multi-Agent Systems and Autonomous Research

4.1 Frameworks and Challenges in Multi-Agent Systems

Multi-agent systems (MAS) are pivotal in advancing autonomous research, enabling AI agents to collaborate on complex tasks through sophisticated frameworks that enhance interactions and task execution. The Smurfs framework exemplifies this by utilizing specialized agents for task decomposition, optimizing research processes and efficiency [26]. Similarly, the CRAT framework employs agents like the Unknown Terms Detector and Knowledge Graph Constructor to improve translation quality, demonstrating MAS's potential in complex workflows [33].

Integrating large language models (LLMs) with cognitive architectures in MAS poses challenges, particularly in improving algorithm selection and execution. The LLM-ML method enhances classification accuracy by combining LLMs with classical machine learning techniques, addressing data distribution shifts [24]. The LLM-POET method further adapts LLMs into the POET algorithm, tackling complexities in environment generation within MAS [3].

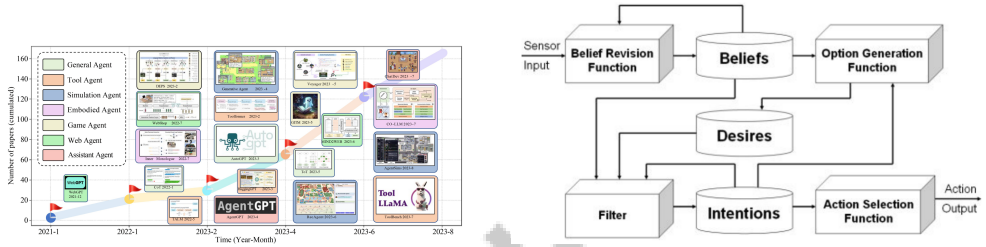
Challenges such as human-like polarization in LLM agents, which can lead to polarized opinions and complicate collaboration, are notable [34]. Frameworks like the role-playing inception prompting method guide agents towards autonomous task completion, facilitating efficient data generation and analysis [35].

Frameworks such as the multi-robot collaboration benchmark assess LLM effectiveness in coordinating tasks among multiple robots, highlighting the need for scalable and accurate task planning [36]. The GAR strategy improves MAS by enabling explicit reasoning and scalability, addressing challenges in scientific discovery [21].

In specialized domains, the PatExpert framework orchestrates tasks among expert agents using a meta-agent, showcasing MAS's potential in patent workflows [37]. The CEEHA framework facilitates AI agent collaboration by integrating cloud-based LLMs for policy generation, illustrating MAS's role in enhancing task execution [2].

Critiques of coloniality and moral absolutism advocate for a decolonial framework in AI alignment, emphasizing diverse perspectives in MAS [10]. The autoformalisation method using Codex highlights challenges in maintaining context across multiple statements in proofs, a significant MAS obstacle [6]. A structured framework categorizing LLM research into pre-training, fine-tuning, efficiency, inference, and challenges offers a comprehensive overview of MAS's current landscape and future directions [11].

Collectively, these frameworks underscore MAS's transformative potential in enhancing scientific discovery by automating critical processes such as systematic literature reviews, patent analysis, and independent research generation. Recent advancements in fine-tuned LLMs illustrate their ability to streamline literature synthesis while maintaining accuracy and transparency, advocating for updates to reporting guidelines. The PatExpert framework exemplifies AI's optimization of patent workflows through a coordinated multi-agent approach, enhancing efficiency and compliance. Moreover, The AI Scientist's introduction signifies a leap toward fully automated scientific research, capable of autonomously generating and validating new ideas. These innovations herald a new era in research, where AI not only supports but actively drives scientific inquiry across diverse fields [13, 14, 37].



(a) Timeline of Agent Research and Development[18]

(b) A Simplified Model of a Decision-Making System[38]

Figure 5: Examples of Frameworks and Challenges in Multi-Agent Systems

As shown in Figure 5, the exploration of multi-agent systems and autonomous research is a dynamic field. The timeline charts the trajectory of agent research and development from 2021 to 2023, highlighting the diversification of agents into categories such as General Agent, Tool Agent, Simulation Agent, Embodied Agent, Game Agent, and Web Agent, emphasizing the broadening scope and specialization within the field. The second example illustrates the decision-making process within these agents, outlining how sensor input informs a belief revision function that updates the agent's beliefs. Together, these examples encapsulate the frameworks and challenges faced in MAS, providing insights into the mechanisms driving autonomous decision-making and the progression of research in this area [18, 38].

4.2 Enhancing Collaboration and Communication

Method Name	Communication Strategies	Framework Integration	Role of Language
CTG[39]	Task Grammar	Task Grammar	Task Grammar
PA[40]	Prediction With Reasoning	Centralized And Decentralized	Underlying Language Models
RPF[35]	Conversation Protocols	Structured Role Assignments	Task-specifier Agent

Table 1: Comparison of communication strategies, framework integration, and language roles in various multi-agent system methodologies. This table highlights the distinct approaches of CTG, PA, and RPF methods in enhancing agent collaboration and task execution through advanced frameworks and linguistic capabilities.

Multi-agent systems (MAS) significantly enhance collaboration and communication among AI agents through advanced frameworks and methodologies that promote decentralized control and sophisticated communication strategies. The Collaborative Task Grammar (CTG) allows agents

to autonomously determine their roles and collaborate without predefined assignments, enhancing flexibility and scalability [39]. This adaptability enables agents to respond effectively to changing environments and tasks, promoting efficient teamwork and resource allocation.

As illustrated in Figure 6, the integration of various frameworks, communication strategies, and benchmarks plays a crucial role in enhancing collaboration within MAS. This figure highlights key components such as the PreAct framework, which enhances agents' planning abilities by reflecting on predicted outcomes and adjusting strategies accordingly. By doing so, agents improve their decision-making processes in unforeseen circumstances [40]. This predictive capability is crucial for optimizing collaborative efforts in complex environments.

Structured communication approaches, such as the CAMEL framework, reduce confusion and enhance cooperation among agents by clearly defining roles and expectations [35]. This methodology ensures effective coordination and information sharing, improving overall system performance and reliability.

Hybrid frameworks that combine centralized and decentralized planning, as seen in multi-robot collaboration benchmarks, improve task success rates and scalability [36]. These frameworks allow agents to leverage the strengths of both planning paradigms, optimizing task execution and resource utilization.

Language plays a pivotal role in facilitating generalization and enhancing communication among agents. Empirical studies reveal how language supports agents in generalizing learned concepts across various scenarios, highlighting the trade-offs between teaching modalities [41]. This linguistic capability is essential for effective knowledge transfer and collaborative problem-solving.

Benchmarks like ARENA, which support diverse multi-agent scenarios, promote innovation and collaboration in MAS research [23]. These platforms provide robust environments for testing and refining communication strategies, enabling agents to develop more effective collaborative behaviors.

Table 1 provides a detailed comparison of the communication strategies, framework integration, and the role of language in various methodologies employed within multi-agent systems, illustrating their impact on enhancing collaboration and communication. The methodologies and frameworks discussed underscore the transformative potential of MAS in enhancing collaboration and communication among AI agents, particularly in automating systematic literature reviews, improving teacher professional development through effective content knowledge identification, and leveraging advanced language models for versatile applications across various industries. These advancements streamline labor-intensive processes while advocating for methodological transparency and reliability in research practices, ultimately setting new standards for AI integration in academic and professional contexts [13, 42, 16]. By leveraging advanced planning, prediction, and communication strategies, MAS facilitate more efficient autonomous research systems, advancing the capabilities of intelligent systems in scientific discovery.

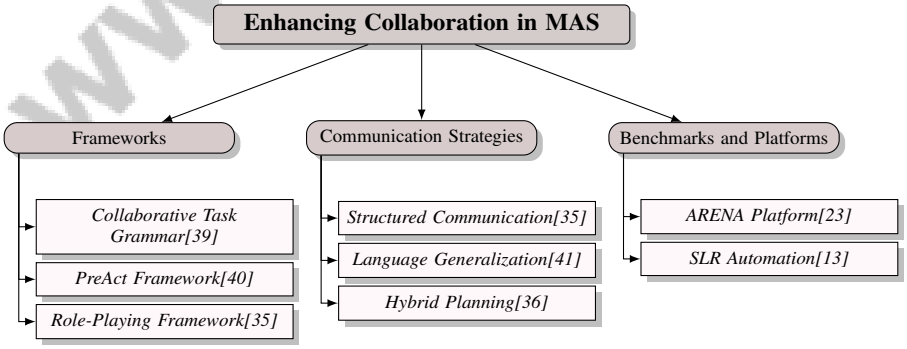


Figure 6: This figure illustrates the key frameworks, communication strategies, and benchmarks that enhance collaboration and communication in multi-agent systems (MAS). It highlights the integration of advanced frameworks such as the Collaborative Task Grammar, PreAct Framework, and Role-Playing Framework, alongside structured communication strategies and language generalization techniques. Additionally, it showcases benchmarks like the ARENA Platform and SLR Automation, which support the development and evaluation of MAS.

5 AI-Driven Discovery and Intelligent Systems

5.1 Intelligent Systems and AI-Driven Discovery

Intelligent systems are pivotal in automating research processes and enhancing decision-making across diverse fields by integrating large language models (LLMs) and multi-agent frameworks. This integration allows operations with minimal human intervention, significantly improving task execution efficiency, especially in robotics [2]. The Richelieu model exemplifies this potential by autonomously refining diplomatic strategies through self-play, highlighting the role of intelligent systems in enabling autonomous research and strategic decision-making [4].

In healthcare, intelligent systems like the AntGLM-Med-10B model enhance medical reasoning and address complex queries with precision [20]. Similarly, LLMs in dentistry improve diagnostic and treatment planning, leading to better clinical outcomes [12]. These systems are architected to understand model behaviors, crucial for refining functionality and enabling autonomous scientific inquiries [27]. By employing hierarchical structures and advanced algorithms, these systems operate with minimal oversight, expanding the scope of autonomous research.

These advancements underscore the transformative potential of intelligent systems in AI-driven scientific discovery. Fully automated research processes, exemplified by the AI Scientist's ability to generate novel ideas, conduct experiments, and produce publishable papers efficiently, enhance scientific exploration accuracy and efficacy. The integration of fine-tuned LLMs in automating Systematic Literature Reviews streamlines labor-intensive methodologies while ensuring high standards of factual accuracy and reliability. Collectively, these innovations foster more sustainable and effective research practices, paving the way for creativity and innovation in addressing complex global challenges [13, 14]. Intelligent systems are poised to revolutionize scientific research, driving innovation and expanding knowledge frontiers.

5.2 Enhancing Reasoning, Generalization, and Decision-Making

AI systems, particularly those utilizing large language models (LLMs), have significantly advanced reasoning, generalization, and decision-making in scientific research. Integrating LLMs with Answer Set Programming (ASP) exemplifies this progress, enabling natural language queries to be converted into logical representations processed by ASP, thereby enhancing reasoning capabilities [43].

The LLM+ASP method also enhances generalization capabilities, allowing effective application of learned knowledge across diverse contexts. This approach supports robust reasoning and task adaptation without extensive retraining, essential for tackling complex tasks like question-answering and automated planning while maintaining high accuracy and reliability [13, 44, 43, 28, 20]. By converting complex queries into structured logical formats, this method supports robust decision-making processes, resulting in accurate and reliable outputs.

Furthermore, integrating LLMs with advanced algorithms enhances decision-making in dynamic and uncertain environments. This is evident in autonomous robotic systems, where AI agents navigate complex settings and make real-time decisions despite incomplete or rapidly changing data. These systems utilize sophisticated models to analyze vast information, optimize decision-making, and adapt to new challenges while maintaining high accuracy and efficiency. Frameworks like The AI Scientist demonstrate AI's potential in dynamic contexts by autonomously generating ideas, conducting experiments, and evaluating findings [13, 14, 17, 45]. Leveraging LLMs in natural language processing alongside logical reasoning frameworks, AI systems achieve higher cognitive functions, enabling them to address complex scientific inquiries with greater precision and efficacy.

The integration of LLMs with ASP and other reasoning frameworks highlights AI systems' transformative potential in advancing scientific discovery by automating complex processes like systematic literature reviews, enhancing factual accuracy through retrieval-augmented techniques, and improving reasoning capabilities across research domains. This synergy streamlines labor-intensive methodologies and establishes new standards for methodological transparency and reliability, democratizing access to advanced reasoning abilities in academic research [13, 28, 43, 46]. This integration empowers AI systems to reason, generalize, and make informed decisions, driving innovation and advancing knowledge across various scientific domains.

6 Challenges and Future Directions

6.1 Challenges, Limitations, and Considerations

The implementation of AI scientists and autonomous research systems faces significant challenges that impede their full potential in scientific discovery. A major issue is the lack of transparency in large language models (LLMs), which complicates interpretability and optimization, highlighting the need for more transparent systems [27]. Models like Richelieu reveal limitations when applied to complex real-world scenarios, indicating the necessity for improvements in environments with incomplete information [4]. The dependency on extensive labeled datasets further limits practical applications, particularly in educational research [22].

In medicine, the discrepancy between LLM capabilities and clinical workflows stems from insufficient training on specialized data, necessitating comprehensive, domain-specific datasets to enhance applicability [20]. Ethical issues, such as data privacy and biases in model training, are critical limitations in clinical applications [12]. Developing inclusive and context-sensitive AI alignment frameworks is crucial to address these ethical challenges [10].

Scalability, computational costs, and the interpretability of optimization processes also pose significant hurdles for broader LLM applications [7]. Continuous innovation in AI research is essential to effectively integrate these technologies into scientific discovery, addressing both technical and ethical challenges.

6.2 Ethical and Privacy Considerations

Ethical and privacy considerations are critical in AI-driven research, particularly as LLMs and multi-agent systems become more capable. Integrating these technologies demands careful attention to privacy and security, especially concerning user data in multi-agent systems, where transparency in agent behavior is crucial for maintaining trust [47].

Addressing biases in LLM-based agents is vital due to their significant impact on diverse demographics. The development of safe and responsible large language models (SRLLM) aims to mitigate such biases [48]. Additionally, the growing autonomy of LLMs raises concerns about their influence on public opinion, necessitating ethical guidelines for their deployment [34].

In AI-driven mental health care, privacy concerns are pressing, as AI systems' reliance for diagnosis introduces risks of misdiagnosis, emphasizing the need for rigorous ethical standards and privacy protections, especially for platforms handling Protected Health Information (PHI) [49, 50].

Ethical approval and informed consent from participants are crucial in AI research, as shown in studies automating psychological hypothesis generation [51]. Additionally, datasets must be curated to eliminate ethical risks and sensitive content, ensuring AI systems are trained on data that respects privacy considerations [8].

These considerations highlight the importance of robust ethical frameworks and privacy safeguards in AI research, promoting responsible AI technology development while minimizing potential harms [52].

6.3 Enhancing AI Capabilities and Learning

Enhancing AI capabilities and continuous learning is crucial for advancing autonomous research and scientific discovery. Integrating LLMs with reinforcement learning techniques refines decision-making processes and adaptability, enabling AI systems to learn from dynamic environments [4].

Hybrid models combining LLMs with classical machine learning algorithms can enhance performance by addressing specific task limitations, such as classification and regression [24]. This synergy improves accuracy and broadens applicability across diverse domains.

Meta-learning frameworks facilitate continuous learning by allowing models to adapt quickly to new tasks with minimal data, particularly in data-scarce environments [26]. Feedback mechanisms enable AI systems to learn from user interactions, enhancing responsiveness and alignment with user expectations [47].

Transfer learning techniques are vital for enhancing AI capabilities, enabling systems to leverage knowledge from pre-trained models for new tasks, improving performance with limited training data [20]. Collaboration among AI agents in multi-agent systems enhances collective capabilities, enabling more effective problem-solving [33].

By adopting these advanced strategies, the scientific community can significantly improve AI functionality and learning capabilities, streamlining processes like literature reviews and data mining in chemical literature. These innovations pave the way for AI to become a collaborative partner in research, addressing complex global challenges with increased creativity and effectiveness [13, 17, 19, 14, 15].

6.4 Future Research Directions

Future research in AI-driven discovery aims to enhance the adaptability and efficacy of AI scientists across various scientific domains. Refining behavioral taxonomies and developing new evaluation methods will improve model interpretability and predictability [27]. Exploring efficient training methods and aligning LLMs with human values are crucial for generalizing across tasks with minimal fine-tuning [11].

Improving the adaptability of cloud-edge-end architecture to dynamic environments and optimizing task decomposition algorithms are vital for enhancing hierarchical LLM performance [2]. Enhancing the robustness and efficiency of optimization algorithms for LLMs and exploring interdisciplinary applications are essential to address foundational interaction theories [7].

In dental applications, future research should focus on data privacy, model accuracy, and expanding LLM applications across dental contexts [12]. Extending frameworks like Richelieu to other multi-agent scenarios will enhance AI-driven discovery, facilitating complex interactions among agents [4].

Improving prompt engineering and integrating contextual information are critical for enhancing AI-driven discovery in educational technology [22]. In the medical domain, optimizing training processes and applying techniques like the Verification-of-Choice approach across diverse tasks will contribute to more reliable AI systems [20].

These research directions aim to develop fully automated systems capable of conducting independent scientific research, generating novel ideas, and synthesizing literature. Such advancements seek to create robust, efficient, and ethical AI systems that streamline research processes, automate experiment execution, and significantly expand scientific knowledge across fields. Frameworks like The AI Scientist enable LLMs to autonomously produce high-quality research papers, while fine-tuned models enhance systematic literature review efficiency. Together, these innovations promise to revolutionize the scientific landscape, making research more accessible and fostering continuous innovation in addressing global challenges [13, 14].

7 Conclusion

AI scientists are reshaping the landscape of scientific research by leveraging the power of large language models (LLMs) and multi-agent systems, which are essential for enhancing autonomous research capabilities. This survey has highlighted the significant impact of LLMs in improving data processing, generating hypotheses, and facilitating decision-making across various fields, such as education, healthcare, and robotics. The integration of LLMs with frameworks like Answer Set Programming underscores their ability to enhance reasoning and generalization.

The use of multi-agent systems promotes better collaboration and communication among AI entities, optimizing task execution and streamlining scientific inquiry. Despite the challenges posed by ethical considerations and the need for continuous adaptation, the focus remains on developing AI systems that are inclusive and transparent.

Future research should concentrate on refining AI models to enhance their interpretability, adaptability, and efficiency, thereby increasing their utility in diverse scientific disciplines. By addressing these challenges and capitalizing on the capabilities of intelligent systems, AI scientists are poised to drive scientific advancements and extend the boundaries of knowledge, heralding a new era of discovery powered by sophisticated AI technologies.

References

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

[2] Zhirong Luan, Yujun Lai, Rundong Huang, Yan Yan, Jingwei Wang, Jizhou Lu, and Badong Chen. Hierarchical large language models in cloud edge end architecture for heterogeneous robot cluster control, 2024.

[3] Fuma Aki, Riku Ikeda, Takumi Saito, Ciaran Regan, and Mizuki Oka. Llm-poet: Evolving complex environments using large language models, 2024.

[4] Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. Richelieu: Self-evolving llm-based agents for ai diplomacy, 2024.

[5] 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.

[6] Ayush Agrawal, Siddhartha Gadgil, Navin Goyal, Ashvni Narayanan, and Anand Tadipatri. Towards a mathematics formalisation assistant using large language models, 2022.

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

[8] Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. Enhancing large language model with self-controlled memory framework, 2024.

[9] Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. "we need structured output": Towards user-centered constraints on large language model output, 2024.

[10] Kush R. Varshney. Decolonial ai alignment: Openness, viśeṣa-dharma, and including excluded knowledges, 2024.

[11] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[12] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, and Bing Shi. Chatgpt for shaping the future of dentistry: The potential of multi-modal large language model, 2023.

[13] Teo Susnjak, Peter Hwang, Napoleon H. Reyes, Andre L. C. Barczak, Timothy R. McIntosh, and Surangika Ranathunga. Automating research synthesis with domain-specific large language model fine-tuning, 2024.

[14] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024.

[15] Lisa Messeri and MJ Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024.

[16] Hitesh Mohapatra and Soumya Ranjan Mishra. Exploring ai tool's versatile responses: An in-depth analysis across different industries and its performance evaluation, 2023.

[17] Kexin Chen, Hanqun Cao, Junyou Li, Yuyang Du, Menghao Guo, Xin Zeng, Lanqing Li, Jiezhong Qiu, Pheng Ann Heng, and Guangyong Chen. An autonomous large language model agent for chemical literature data mining, 2024.

[18] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[19] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[20] Qiang Li, Xiaoyan Yang, Haowen Wang, Qin Wang, Lei Liu, Junjie Wang, Yang Zhang, Mingyuan Chu, Sen Hu, Yicheng Chen, Yue Shen, Cong Fan, Wangshu Zhang, Teng Xu, Jinjie Gu, Jing Zheng, and Guannan Zhang Ant Group. From beginner to expert: Modeling medical knowledge into general llms, 2024.

[21] Yuwei Hu, Runlin Lei, Xinyi Huang, Zhewei Wei, and Yongchao Liu. Scalable and accurate graph reasoning with llm-based multi-agents, 2024.

[22] Lishan Zhang, Han Wu, Xiaoshan Huang, Tengfei Duan, and Hanxiang Du. Automatic deductive coding in discourse analysis: an application of large language models in learning analytics, 2024.

[23] Yuhang Song, Andrzej Wojcicki, Thomas Lukasiewicz, Jianyi Wang, Abi Aryan, Zhenghua Xu, Mai Xu, Zihan Ding, and Lianlong Wu. Arena: A general evaluation platform and building toolkit for multi-agent intelligence, 2019.

[24] Yuhang Wu, Yingfei Wang, Chu Wang, and Zeyu Zheng. Large language model enhanced machine learning estimators for classification, 2024.

[25] Maojun Sun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang. A survey on large language model-based agents for statistics and data science, 2024.

[26] Junzhi Chen, Juhao Liang, and Benyou Wang. Smurfs: Leveraging multiple proficiency agents with context-efficiency for tool planning, 2024.

[27] Ari Holtzman, Peter West, and Luke Zettlemoyer. Generative models as a complex systems science: How can we make sense of large language model behavior?, 2023.

[28] Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. Reta-llm: A retrieval-augmented large language model toolkit, 2023.

[29] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model, 2024.

[30] Adrian de Wynter. Awes, laws, and flaws from today's llm research, 2024.

[31] Qingxiu Dong, Li Dong, Ke Xu, Guangyan Zhou, Yaru Hao, Zhifang Sui, and Furu Wei. Large language model for science: A study on p vs. np, 2023.

[32] Christopher Akiki, Giada Pistilli, Margot Mieskes, Matthias Gallé, Thomas Wolf, Suzana Ilić, and Yacine Jernite. Bigscience: A case study in the social construction of a multilingual large language model, 2022.

[33] Meiqi Chen, Fandong Meng, Yingxue Zhang, Yan Zhang, and Jie Zhou. Crat: A multi-agent framework for causality-enhanced reflective and retrieval-augmented translation with large language models, 2024.

[34] Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Fernando P. Santos, Yong Li, and James Evans. Emergence of human-like polarization among large language model agents, 2025.

[35] 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.

[36] Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems?, 2024.

[37] Sakhinana Sagar Srinivas, Vijay Sri Vaikunth, and Venkataramana Runkana. Towards automated patent workflows: Ai-orchestrated multi-agent framework for intellectual property management and analysis, 2024.

[38] Marx Viana, Paulo Alencar, and Carlos Lucena. Towards an adaptive and normative multi-agent system metamodel and language: Existing approaches and research opportunities, 2021.

[39] Amy Fang and Hadas Kress-Gazit. High-level, collaborative task planning grammar and execution for heterogeneous agents, 2024.

[40] Dayuan Fu, Jianzhao Huang, Siyuan Lu, Guanting Dong, Yejie Wang, Keqing He, and Weiran Xu. Preact: Prediction enhances agent's planning ability, 2024.

[41] Dhara Yu, Noah D. Goodman, and Jesse Mu. Characterizing tradeoffs between teaching via language and demonstrations in multi-agent systems, 2023.

[42] Kaiqi Yang, Yucheng Chu, Taylor Darwin, Ahreum Han, Hang Li, Hongzhi Wen, Yasemin Copur-Gencturk, Jiliang Tang, and Hui Liu. Content knowledge identification with multi-agent large language models (llms), 2024.

[43] Zhun Yang, Adam Ishay, and Joohyung Lee. Coupling large language models with logic programming for robust and general reasoning from text, 2023.

[44] Haoming Li, Zhaoliang Chen, Jonathan Zhang, and Fei Liu. Lasp: Surveying the state-of-the-art in large language model-assisted ai planning, 2024.

[45] Zach Johnson and Jeremy Straub. Development of regai: Rubric enabled generative artificial intelligence, 2024.

[46] Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Democratizing reasoning ability: Tailored learning from large language model, 2023.

[47] Burak Aksar, Yara Rizk, and Tathagata Chakraborti. Tess: A multi-intent parser for conversational multi-agent systems with decentralized natural language understanding models, 2023.

[48] Shaina Raza, Oluwanifemi Bamgbose, Shardul Ghuge, Fatemeh Tavakol, Deepak John Reji, and Syed Raza Bashir. Developing safe and responsible large language model : Can we balance bias reduction and language understanding in large language models?, 2025.

[49] Tianyu He, Guanghui Fu, Yijing Yu, Fan Wang, Jianqiang Li, Qing Zhao, Changwei Song, Hongzhi Qi, Dan Luo, Huijing Zou, and Bing Xiang Yang. Towards a psychological generalist ai: A survey of current applications of large language models and future prospects, 2023.

[50] V. K. Cody Bumgardner, Mitchell A. Klusty, W. Vaiden Logan, Samuel E. Armstrong, Caroline N. Leach, Kenneth L. Calvert, Caylin Hickey, and Jeff Talbert. Institutional platform for secure self-service large language model exploration, 2025.

[51] Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. Automating psychological hypothesis generation with ai: when large language models meet causal graph, 2024.

[52] Jitao Bai, Simiao Zhang, and Zhonghao Chen. Is there any social principle for llm-based agents?, 2023.

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.