Autonomous Generalist Scientist:
Towards and Beyond Human-level Automatic Research Using Foundation Model-Based AI Agents and Robots (A Position)

Starkson Zhang [* 1] **Alfredo Pearson** [* 2] **Zhenting Wang** [3]

Abstract

The accelerating pace of scientific research highlights the need for more efficient, accurate, and comprehensive methodologies. Traditional research methods are often hindered by the limitations of manual experimentation and data collection in isolated environments, resulting in slow and resource-intensive processes. Besides, multidisciplinary research presents significant challenges due to the complexities of integrating knowledge from various fields, often surpassing the expertise of individual researchers. The limited knowledge base of single researchers constrains the scope and depth of inquiry, complicating efforts to fully explore complex interdisciplinary problems. To address these challenges, it is crucial to develop automatic systems that can dynamically interact with both physical and virtual environments while facilitating the integration of knowledge across multiple disciplines. Foundation AI models, such as large language models, are trained on vast amount of data from diverse sources, enabling them to acquire knowledge across various scientific disciplines. Therefore, it is promising to build the automatic research systems based on these foundation models. This paper envisions a groundbreaking framework for automating the entire research lifecycle, leveraging advanced software agents and robotics. By incorporating AI technologies across various stages of research - from literature review and hypothesis generation to experimentation conducting and manuscript writing - this project aims to significantly reduce the time and resources required for scientific discovery. This paper explores the current challenges in academic research and the transformative potential of automating the research lifecycle. The envisioned framework, which integrates advanced AI and robotics, aims to accelerate cross-disciplinary inquiry, democratize access to high-level research, and address complex global challenges. In addition, the paper discusses the societal and ethical implications of automated research, including shifts in scientific training and the innovative advancements required to overcome current technological limitations. Github: https://github.com/universea/Auto-Research.

1. Introduction

Scientific research plays a crucial and profound role in advancing knowledge, driving technological innovation, solving problems, providing education and training, enhancing societal welfare, fostering global collaboration, impacting the economy, and enriching cultural and intellectual life. It not only expands our understanding of the natural world, technological possibilities, and social phenomena but also stimulates economic growth by creating and improving technologies, thereby enhancing quality of life and productivity (Gibbons & Johnston, 1974; Noy & Zhang, 2023).

Despite its critical importance, the current landscape of academic research is characterized by inherent complexity and a slow pace, which often impede rapid scientific progress. Traditional research methodologies involve labor-intensive processes, extensive literature reviews, and meticulous experimental design and execution, all of which consume valuable time and resources (Rossoni et al., 2023; Vamathevan et al., 2019). Additionally, the reliance on specialized expertise limits the progress and innovative capacity of research due to the dependence on a limited pool of experts (Fabrykowska, 2020). Cross-domain knowledge plays a critical role in advancing research, particularly in addressing complex global challenges

---

[*]Equal contribution [1]Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, Canada [2]Istituto Italiano di Tecnologia, Genoa, Italy. Ph.D. program of national interest in Robotics and Intelligent Machines (DRIM) [3]Computer Science Department, Rutgers University, New Jersey, United States.

such as sustainable development and health sciences (Daniel et al., 2022; Freeth & Caniglia, 2020). Multidisciplinary collaboration has demonstrated significant advantages by bringing together diverse skills and perspectives, leading to more comprehensive and innovative research outcomes. However, these collaborations often face substantial challenges, including differences in disciplinary cultures(Daniel et al., 2022), specific methodologies(MacLeod, 2018), and the considerable time and resources required to coordinate across fields. These barriers can impede effective communication, integration of concepts, and the development of unified research frameworks.

Recent advancements in Artificial Intelligence (AI), notably in Large Language Models (LLMs), have ushered in transformative capabilities in generating and understanding human-like text across diverse disciplines. Trained on extensive datasets that integrate various fields, these models excel in applying multidisciplinary knowledge, thereby significantly enhancing scientific research. The inherent capacity of generative AI to traverse and link disparate domains of knowledge makes it exceptionally conducive to interdisciplinary research (Li et al., 2024a). These AI systems have demonstrated remarkable efficiency in tasks such as information aggregation (Kang & Xiong, 2024), idea generation (Baek et al., 2024), coding (Jiang et al., 2024), and academic writing (Irons et al., 2023). Moreover, they have been employed in self-driven hypothesis generation and the exploration of new scientific questions autonomously (Zenil et al., 2023), as well as in specialized areas like biomedical question answering, image analysis (Tu et al., 2024), and data-driven models in medical research (Gao et al., 2024), while also fostering creativity in both scientific and artistic domains (Wingström et al., 2024). These tools not only accelerate data processing and analysis but also uncover patterns and correlations that may elude human researchers, enhancing both the depth and breadth of discoveries (Shir, 2024). However, despite these significant advancements, the application of AI and LLMs remains largely confined to specific, narrow tasks or purely data-centric studies that do not involve interactions with the physical world. This limitation underscores the need for continued development in AI technologies to fully realize their potential in broader scientific contexts, including more tangible, real-world applications. As the field evolves, integrating these sophisticated tools with autonomous agent and robotic systems could potentially unlock unprecedented opportunities in research and beyond.

To date, the AI community has yet to demonstrate the ability to interact with both the physical and virtual worlds in a manner that allows for fully autonomous scientific research, akin to that conducted by human scientists. One of the primary challenges lies in the current limitations of AI systems' ability to seamlessly integrate and interact with both virtual and physical environments. These systems often struggle with independently searching for, collecting data, and performing manipulating tasks across different domains, which is essential for conducting comprehensive research. This limitation is particularly significant in fields such as biology, medicine, and engineering, where interaction with the physical world is crucial. For example, in biomedical research, AI systems must be able to handle complex physical tasks such as manipulating biological samples or operating lab equipment, in addition to analyzing vast amounts of virtual data (da Silva, 2024). The inability to perform these tasks autonomously across both worlds presents a significant barrier to developing AI scientists capable of independently conducting research. Overcoming this challenge is critical to advancing the field and enabling AI systems to perform scientific research with the same level of autonomy and adaptability as human scientists.

The motivation behind the autonomous generalist scientist auto research project is multifaceted:

- **Accelerating the Slow Pace of Scientific Research Due to Inherent Complexity**: Scientific research requires processing large amounts of information and complex data, which often impedes progress due to methodological limitations. Automated systems can expedite these initial stages, allowing researchers to move to experimentation and implementation faster.

- **Reducing the Demand for Specialized Expertise**: Conventional scientific inquiry typically requires domain-specific knowledge, limiting the speed and innovative capacity of research. LLMs can synthesize information across a broad array of documents, democratizing the ability to generate innovative research proposals.

- **Enhancing the Quality and Innovation of Research Ideas**: Developing high-quality and innovative research ideas is challenging and often requires iterative refinement. Automated systems can generate and refine ideas through feedback loops with multiple reviewing agents, ensuring the ideas are both innovative and applicable.

- **Promoting Cross-Disciplinary Application**: Many scientific problems today require a multidisciplinary approach. Specialized research agents, trained across various disciplines, can collaborate effectively, harnessing collective strengths and pushing the boundaries of discovery faster than isolated efforts.

To address these challenges, the "autonomous generalist scientist" project embodies this vision of fully automating the

research lifecycle, utilizing LLM-based agents and robotics. This initiative seeks to transform how scientific inquiry is conducted, making it more efficient, accessible, and democratic. By automating the entire research process, the project aims to accelerate scientific discovery and innovation across multiple disciplines, thereby enhancing research productivity and broadening the horizons of academic investigation (Hughes et al.).

In this paper, we introduce a framework for an AI-based agent and robot autonomous generalist scientist(AGS). To the best of our knowledge, this is the first fully autonomous system with human-level capabilities that can interact with both virtual and physical environments to conduct complex scientific research. Referred to as the "Agent and Robot Scientist," this system can initiate and execute a complete research process, beginning with an idea, a question, or a user's specific need. It is designed to conduct literature reviews, articulate research problems, generate innovative hypotheses, plan and manage experimental projects, and execute experiments in both real-world and virtual settings. Furthermore, it incorporates mechanisms for self-reflection and external feedback, allowing for iterative refinement of experiments and research planning. Finally, the system is capable of drafting manuscripts and conducting peer reviews, producing research papers that mirror the workflow of human researchers.

The framework is organized into four major functional modules, complemented by interaction and reflection mechanisms. The four key modules are:

- **Literature Review**: This module is responsible for conducting a comprehensive review of existing research related to the given idea or question. The AI system autonomously searches academic databases, selects relevant papers, summarizes findings, and identifies gaps in the literature that the new research could address. Unlike traditional methods that rely solely on APIs to access academic papers, this system simulates human-like interactions with computers and websites to navigate different journal platforms, overcoming access restrictions that often limit direct API usage. It is capable of operating as a human researcher would—searching, downloading, and managing literature across various platforms—ensuring a more comprehensive and practical approach to gathering necessary research materials.

- **Proposal Generation**: Once the literature review is complete, the system generates a detailed research proposal. This proposal includes a clear problem statement, the research objectives, and innovative hypotheses that can potentially advance the field. The module also outlines the methodology and experimental design, considering both the virtual and physical environments in which the research will be conducted.

- **Experimentation**: This module manages and executes the experimental phase of the research process. It encompasses the planning of experiments, resource management, and the conduct of trials in both real and virtual environments. The system is equipped with advanced robotics and AI technologies that enable it to manipulate physical objects, collect data, and simulate experiments in virtual settings. Additionally, this module iteratively refines experimental designs based on real-time data and feedback.

- **Manuscript Preparation**: After completing the experiments, the system drafts a research manuscript. This module handles data analysis, result interpretation, and the formulation of conclusions. The manuscript is prepared in a format suitable for submission to academic journals, including sections such as introduction, methods, results, and discussion. The system also performs a self-review and can engage in a peer review process, ensuring the quality and rigor of the research before publication.

- **Communication and Reflection**: Beyond the standard research workflow, this module ensures continuous improvement and adaptability throughout the research process. It facilitates communication between different modules, allowing for real-time feedback and adjustments. The system incorporates external feedback, whether from human researchers or simulated peer review processes, and uses this information to refine hypotheses, methodologies, and experimental designs. This iterative approach ensures that the research remains aligned with the latest developments and insights, ultimately enhancing the quality and impact of the final output.

Overall, the autonomous generalist scientist auto-research framework represents a groundbreaking step toward creating fully autonomous research systems. It aims to pave the way for more efficient and innovative scientific investigation.

2. Definition of Autonomous Generalist Scientist with Different Automation Levels

The concept of a "Autonomous Generalist Scientist" refers to an AI-driven system capable of conducting research across multiple disciplines, mirroring the capabilities of human scientists, and eventually surpassing them in scope, speed, and
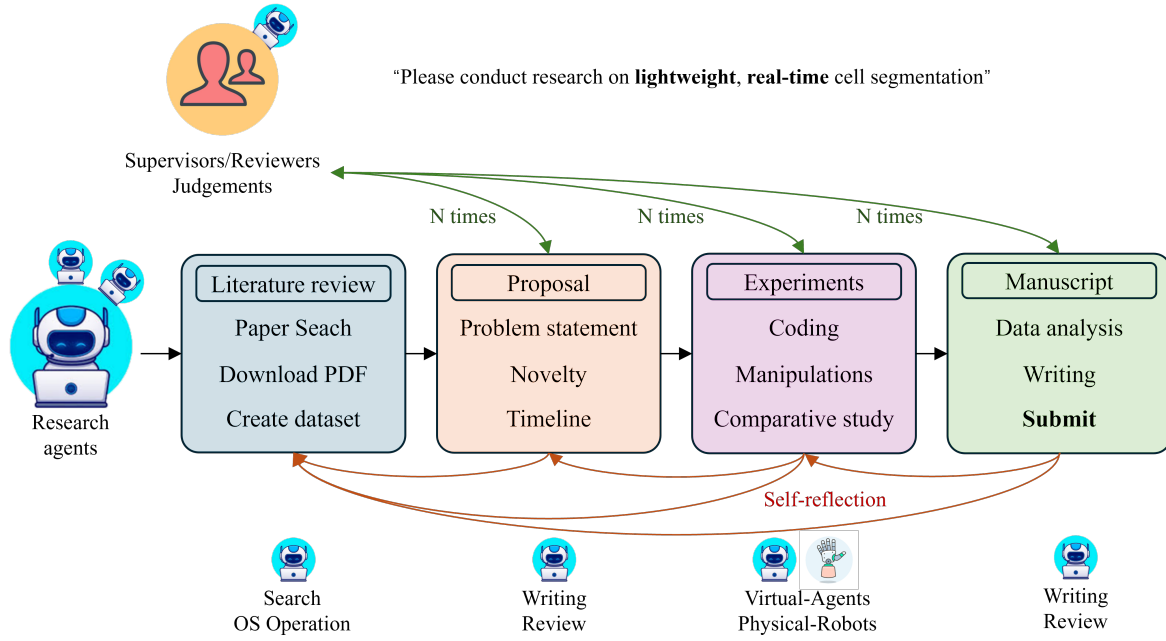
Figure 1. Framework of autonomous generalist scientist based on AI agents and robots. Research agent/robot can accelerate scientific research progress and bridge the gap between scientific knowledge in different disciplines.

depth of inquiry. This section outlines a proposed framework for categorizing these systems into different levels based on their abilities, ranging from no AI involvement to fully autonomous, superhuman-level generalist scientists. The levels (see Table 1) are defined by the extent of human involvement, interaction with environments, knowledge integration, and typical applications.

3. Architecture of Autonomous Generalist Scientist for Automatic Research

3.1. Overview

The AGS is a fully integrated framework that merges advanced AI technologies with robotics to automate the entire research lifecycle. The architecture is built on a multi-agent system where AI agents, powered by Large Language Models (LLMs) and machine learning frameworks, collaborate with robotic systems to perform tasks ranging from data analysis and literature review to experiment execution and manuscript writing. The AI agents manage virtual tasks such as coding, hypothesis generation, and data processing, while the robotics component handles the physical aspects of research, such as manipulating laboratory equipment and conducting experiments with high precision. This integration not only accelerates research but also enhances the accuracy and reproducibility of scientific discoveries, offering a transformative approach to how research is conducted across multiple disciplines.

3.2. Literature Review

The literature review is a fundamental component of the research process, serving as the foundation for identifying existing knowledge, gaps, and opportunities for new discoveries. Traditional methods of conducting literature reviews involve manual searching, filtering, and analyzing a vast amount of academic publications, which can be time-consuming and prone to limitations, particularly when it comes to accessing the most recent data. This section explores the evolution of AI-driven literature review methods, comparing traditional database/API-based approaches with the more advanced OS agent-driven systems that mimic human interactions to perform complex search and manipulation tasks in virtual environments.

Generality	Human Involve	Interaction	Knowledge	Typical	Description
Level 0: **No AI**	Complete human	Limited tools	Fixed	Excel, Matlab	Entirely human-driven processes with no AI assistance; limited to traditional computational tools, often restricted to a single field of study.
Level 1: **Chat-based**	Predominantly human-driven	Limited virtual env	Internal, API search	ChatGPT	AI assists in basic tasks, providing simple cross-disciplinary insights.
Level 2: **Agent-based**	Significant human oversight	Virtual env	Internal, web search	OpenDevin	AI agents assist in research across multiple domains, synthesizing information from various fields but still largely operating within predefined boundaries.
Level 3: **Robot-based (Master level)**	Collaborative human-AI/robot interaction	Virtual and Physical envs	Internal, web search, Physical world perception	Advanced Robotics	AI and robotics collaborate on tasks that require input from multiple disciplines, such as biology, engineering, and data science, for complex experimental setups.
Level 4: **Toward human-level (Ph.D and Prof. level)**	Minimal human supervision	Advanced interaction with Virtual and Physical envs	Internal, Autonomous search, Creating knowledge	AGI Robot (AGIR)	AI operates semi-independently, conducting complex interdisciplinary research, and developing innovative solutions by merging insights from diverse fields.
Level 5: **Beyond-human-level**	Fully Autonomous	Fully autonomous interaction in all environments	Internal, Autonomous search, Groundbreaking innovations	Artificial SuperIntelligence Robot (ASIR)	AI and robotics surpass human capabilities, conducting fully independent research that integrates and transcends multiple disciplines, leading to unprecedented scientific breakthroughs.

Table 1. Automatic research with different automation levels.

3.2.1. LIMITATIONS OF TRADITIONAL AUTOMATIC LITERATURE REVIEW METHODS

Traditional automatic literature review methods often rely heavily on manual processes and specific database or API access, which can limit the scope and currency of the data retrieved. For example, database-driven searches often yield outdated information due to the lag in data indexing and publication delays. On the other hand, API-based methods, while more dynamic, are still constrained by the limitations of the datasets they access, which may not always include the latest research publications. These constraints pose significant challenges, particularly for researchers working in fast-evolving fields where up-to-date information is crucial.

Several AI-driven systems have emerged to enhance the efficiency and personalization of literature reviews. For instance, the Survey Agent (Wang et al., 2024a) utilizes conversational AI to provide personalized and efficient research surveys, though it still depends largely on API access. Similarly, AutoSurveyGPT (Xiao, 2023) leverages GPT models for automated literature discovery, improving the speed and breadth of literature reviews but still facing limitations in terms of accessing the latest publications due to its reliance on API-driven data. In the medical field, systems like the AI Chatbot in Cancer Research (Pan et al., 2023) demonstrates the growing role of AI in specialized domains, yet these systems too are often restricted by the boundaries of their data sources.

While these advancements represent significant improvements over purely manual methods, they still fall short in providing the most up-to-date and comprehensive coverage of literature, highlighting the need for more sophisticated solutions.

3.2.2. ADVANCING LITERATURE REVIEW: OS AGENTS AND MANIPULATION-BASED SYSTEMS

The limitations of traditional automatic literature review methods have spurred the development of more advanced AI systems that go beyond simple database/API access. These new systems, known as OS agents, are designed to simulate human-like interactions with computer systems, allowing them to perform complex searches and manipulation tasks across various digital platforms. This advancement represents a significant step forward in automating the literature review and data collection process, enabling researchers to more efficiently and comprehensively gather relevant information, thus laying a stronger foundation for subsequent research tasks.

For example, GPT-4 Vision (Zheng et al., 2024), a generalist web agent, demonstrates the potential for AI systems to perform complex web-based tasks when properly grounded. OS-Copilot (Wu et al., 2024) takes this a step further by representing a generalist computer agent capable of self-improvement, highlighting its ability to interact with real-world computer environments and adapt its operations. Furthermore, tools like VisualWebArena (Koh et al., 2024) and OSWorld (Xie et al., 2024) are pushing the boundaries of multimodal agents by enabling them to undertake open-ended tasks in real computer environments, such as navigating websites, interacting with interfaces, and retrieving the latest data directly from online sources.

These systems represent a significant advancement over traditional API-based methods, as pre-defined data access limitations do not constrain them. Instead, they can dynamically interact with various platforms, accessing and managing the most current data available.

3.2.3. COMPARISON OF DIFFERENT AUTOMATIC LITERATURE REVIEW METHODS

To illustrate the literature review task performance, Table 2 provides a comparison through three key strategies: leveraging existing databases or the model's internal knowledge base, utilizing search APIs for web-based queries, and simulating human-like interactions to autonomously search and manipulate content on digital platforms. This table highlights the capabilities of each method in terms of data search and the ability to perform manipulation tasks in virtual environments for literature review task.

Methods	Search	Manipulation task in virtual environment
Database	Outdate data	×
API	Limitation latest data	×
OS agent	Latest data	✓ (mouse, keyboard, etc. action)

Table 2. Comparison of methods for searching and performing manipulation tasks in virtual environments.

As shown in Table 2, OS agents offer a significant advantage in accessing the latest data and performing manipulation tasks that are crucial for a comprehensive and up-to-date literature review. These capabilities make them an essential tool for

researchers aiming to stay at the forefront of their fields.

The advancements in OS agents and manipulation-based systems pave the way for a fully automated literature review process. The generalist scientist auto-research framework integrates these cutting-edge technologies to automate the entire literature review workflow. By leveraging the ability of OS agents to interact with various digital platforms, this system not only retrieves and organizes the latest research but also analyzes it to identify gaps and opportunities for new research directions.

The AGS represents a significant step forward in the automation of scientific inquiry, enabling researchers to conduct literature reviews that are not only comprehensive and current but also seamlessly integrated with the broader research lifecycle. This ensures that the literature review process is no longer a bottleneck but rather an enabler of rapid and innovative scientific discovery.

3.3. Proposal Generation

The research proposal is a critical document that outlines the intended research, defines the problem, and presents a detailed plan for addressing it. In the generalist scientist auto-research framework, the proposal generation process is enhanced by the integration of advanced AI, which automates and refines this process iteratively. While other systems, such as those described in (Baek et al., 2024), focus on generating research ideas, our framework takes this further by automatically formulating problem statements, generating hypotheses, designing methodologies, and planning the entire research project. This section describes how the generalist scientist auto research system generates research proposals, emphasizing the AI's role in problem statement formulation, hypothesis generation, methodological design, and the integration of feedback to ensure that the proposals are both innovative and aligned with current research gaps.

3.3.1. PROBLEM STATEMENT FORMULATION

One of the first steps in generating a research proposal is defining a clear and concise problem statement. The AI system begins by analyzing the outcomes from the literature review module, identifying significant gaps in the current body of knowledge. By synthesizing insights from multiple sources, the AI formulates a problem statement that addresses these gaps and poses a relevant and challenging question for the research to answer.

To ensure that the problem statement is both meaningful and aligned with current scientific discourse, the system employs techniques such as topic modeling and clustering. These methods help distill broad research areas into focused questions that can guide the direction of the research. By doing so, the system ensures that the proposed research is not only relevant but also poised to make a significant contribution to the field.

3.3.2. HYPOTHESIS GENERATION AND METHODOLOGICAL DESIGN

Once the problem statement is established, the AI system moves on to generating hypotheses. The system assesses the novelty of potential research ideas by comparing them with the existing literature. This comparison involves analyzing the similarity of proposed hypotheses to existing theories and identifying unique perspectives or relationships that have not yet been explored.

The system then designs a detailed methodology to test these hypotheses. This includes selecting appropriate experimental designs, data collection methods, and analysis techniques. The AI ensures that the proposed methods are rigorous, feasible, and aligned with the research objectives. By leveraging data from the literature review, the system incorporates best practices and lessons learned from previous studies, enhancing the robustness and reliability of the research plan.

3.3.3. PROJECT PLANNING AND INTEGRATION OF DATA FROM LITERATURE REVIEW

The AI system integrates data from the literature review to inform the research design and project planning. This involves extracting relevant methodologies, identifying potential challenges, and incorporating strategies to mitigate these challenges into the project plan.

The project planning module then creates a comprehensive timeline, allocating resources efficiently and setting clear milestones for the research process. The system's ability to simulate potential outcomes based on different research paths allows for a dynamic and flexible approach to project management, ensuring that the research remains on track and adaptable to new findings.

3.3.4. Communication, Feedback, and Iterative Refinement

An essential aspect of the proposal generation process is communication and the iterative refinement of ideas. The AI system facilitates communication with supervisors, users, or peer agents, presenting the proposal in a clear and accessible format for feedback. This feedback is then used to refine the proposal, incorporating new insights and addressing any weaknesses identified by reviewers.

The system employs a multi-agent approach to peer review, where multiple AI agents critique the proposal, suggest improvements, and challenge underlying assumptions. This iterative process ensures that the final proposal is well-rounded, thoroughly vetted, and ready for submission.

3.3.5. Ensuring Innovation and Alignment with Research Gaps

To ensure that the proposals generated by the AI system are innovative and aligned with current research gaps, the system continuously evaluates the novelty and potential impact of the proposed research. This evaluation is based on metrics such as originality, feasibility, and the likelihood of advancing knowledge in the field.

By aligning the research proposal with the gaps identified during the literature review, the system increases the proposal's relevance and the potential for significant contributions to the field. This ensures that the research is not only cutting-edge but also strategically positioned to address the most pressing questions in the discipline.

3.4. Experimentation

3.4.1. Introduction: The Challenges in Scientific Experimentation

Scientific research often extends beyond coding or data analysis and requires intricate physical experiments that involve interacting with real-world materials and environments. Traditional research methods rely heavily on human expertise to design and conduct these experiments, which can be labor-intensive and time-consuming. While AI and machine learning have made significant strides in data-centric tasks, the automation of physical experimentation remains limited. For instance, research tools that use AI for data science (Hong et al., 2024) or that employ LLMs for specific scientific discovery tasks (Ma et al., 2024a) are limited to virtual environments, which restricts their applicability in real-world, physical experiments.

These limitations highlight the need for embodied intelligent robots—robots capable of executing complex tasks in the physical world with the same precision and adaptability as human scientists. Such robots would address current challenges in generalization and adaptability, as many existing systems can only perform narrow, predefined tasks and are unable to transition smoothly across different experimental setups or research domains (Ma et al., 2024b; Yoshikawa et al., 2023).

3.4.2. Current Advances and Remaining Challenges

Existing robotic systems, such as those described in the field of autonomous chemical research (Boiko et al., 2023), illustrate some progress in using AI to guide specific experimental tasks. However, these systems often lack the flexibility needed to adapt to different tasks or environments. For example, while LLM-driven research platforms have shown the potential for automating the research process from hypothesis generation to data analysis (Ifargan et al., 2024), they still fall short regarding performing physical manipulations and engaging with the real world.

Robotic platforms like *OpenDevin* (Wang et al., 2024b) (Brohan et al., 2023), designed for generalist agents, offer a glimpse into how future systems could be developed to handle a wider variety of tasks. Yet, the application of such platforms to the physical sciences remains an ongoing challenge. The development of vision-language models (Ma et al., 2024b) and embodied AI systems suggests that AI can bridge the gap between understanding complex instructions and executing them in the real world, but the technology is still in its infancy regarding scientific experimentation. (Vuong et al., 2023) introduces a collection of robotic learning datasets that are publicly available, promoting transparency and accessibility. which emphasizes an interdisciplinary approach, integrating concepts from various fields, which may enhance the performance of robot-assistant experiments. While the datasets and models are designed to be comprehensive, there may be limitations in their generalizability to real-world applications. Furthermore, the proposed models may face challenges in scaling to more complex or dynamic environments that require more sophisticated learning strategies.

3.4.3. Proposed Framework for Experimentation in Generalist Scientist Auto Research

The generalist scientist auto-research framework proposes a system where AI and robotics are integrated to automate the planning and execution of experiments in both virtual and physical environments. By leveraging insights from current research, such as the use of LLMs for optimizing experimental conditions (Ma et al., 2024a) and the advancements in robotic manipulation of scientific materials (Li et al., 2024b), our framework seeks to create a more adaptable and efficient experimental process.

- **Planning and Execution**: The system begins with the interpretation of research proposals, identifying the necessary experimental tasks, and creating a detailed plan for execution. This involves not only the selection of appropriate tools and methods but also the management of resources and scheduling of tasks to ensure efficiency and accuracy.

- **Robotics for Physical Experiments**: Embodied intelligent robots within the generalist scientist auto-research framework would be capable of performing complex manipulations and adjustments in real time. These robots would draw on current advances in the field, such as the flexible automation seen in platforms like *ORGANA* (Darvish et al., 2024), which automates various chemistry experiments while interacting with human researchers to refine and adapt the experimental process.

- **AI-Driven Simulations**: The integration of AI-driven simulations, similar to those used in current research for optimizing experimental conditions (Jablonka et al., 2023), would allow the system to simulate various experimental scenarios before physical execution. This approach not only reduces the cost and time associated with physical experiments but also enhances the reliability of the results by preemptively identifying potential issues.

- **Resource Management and Real-Time Adjustments**: Effective management of experimental resources, coupled with the ability to make real-time adjustments based on feedback, is crucial for the success of any scientific research. Systems like *ROS-LLM* (Mower et al., 2024) offer examples of how large language models can be integrated with robotic systems to enable real-time decision-making and structured reasoning during experiments.

- **Ensuring Reproducibility and Accuracy**: The framework also emphasizes the importance of reproducibility and accuracy, which are fundamental to the credibility of scientific research. By using validated robotic systems and AI models that can adapt to various experimental conditions, the generalist scientist auto-research framework aims to ensure that results are consistent and reliable across different settings, much like the robust experimental protocols seen in projects such as *Chemistry3D* (Li et al., 2024b).

Methods	Objects	Tasks	Areas	Advantages
Agent	Virtual env.	Coding, data analysis	Computer, Data, Humanities	Efficient data processing
Robot	Physical env.	Manipulation of tools	Medical, Bio., Eng., etc.	Precision in interactions

Table 3. Comparison of methods for research tasks.

3.5. Manuscript Preparation

3.5.1. Introduction: The Role of Manuscript Preparation in Research

Manuscript preparation is a critical phase in the research process, encompassing the synthesis, structuring, and presentation of research findings. Traditionally, this stage poses significant challenges due to the need for meticulous attention to detail, adherence to academic standards, and effective communication of complex ideas.

3.5.2. Proposed Approach to Automated Manuscript Drafting

In the generalist scientist auto-research framework, we envision leveraging advanced AI technologies to transition seamlessly from experimental data to manuscript drafting. The goal is to develop a system that can generate an initial draft, organizing content into coherent sections that align with standard academic formats.

- **Experiment Result Data Analysis and Summary**: The system is proposed to first compile experimental data, create summaries, and illustrate the results with figures and tables. Existing tools such as MatPlotAgent (Yang et al., 2024)

and Data interpreter (Vázquez, 2024), which automates scientific data visualization, serve as an inspiration for how these tasks might be effectively handled (Yang et al., 2024). The system would ensure that data is presented in a way that highlights the key findings of the research.

- **Writing Assistance**: Once the data is analyzed, the AI will assist in drafting the manuscript. It could generate text for various sections such as the introduction, methods, results, and discussion. Additionally, it would aid in writing complex formulas and adhering to predefined manuscript templates to ensure consistency and professionalism.

3.5.3. PROPOSED MANUSCRIPT REFINEMENT PROCESS

To refine the manuscript, we plan to implement iterative cycles where the system enhances language clarity, ensures consistency, and adapts the content to meet the specific requirements of different journals.

- **Figures and Visualizations**: One of the significant challenges in manuscript preparation is creating high-quality figures. We propose automating this process by incorporating tools inspired by MatPlotAgent (Yang et al., 2024), which generates accurate and insightful visualizations crucial for the manuscript (Vázquez, 2024).

- **Reference Management**: Managing citations and references is another critical task where the system could excel. By integrating with reference management software, the AI would ensure that all citations are correctly formatted and that the bibliography is comprehensive and up-to-date (Bom, 2023).

3.5.4. PEER REVIEW SIMULATION CONCEPT

To anticipate and address potential criticisms, the system could include a virtual peer review process. This would allow for the identification and resolution of weaknesses before submission, ensuring that the manuscript meets the highest standards of academic rigor.

- **Self-Review and Feedback Integration**: The system might perform a self-review of the manuscript, simulating the peer review process to identify possible flaws or areas needing improvement. This would include checking for logical coherence, ensuring that all claims are supported by data, and verifying the accuracy of references. Inspiration could be drawn from existing AI models that simulate feedback from hypothetical reviewers and refine the manuscript accordingly (Lingard, 2023; Dergaa et al., 2023).

- **Ethical Considerations**: The use of AI in manuscript preparation raises ethical questions, such as those related to authorship and the integrity of the generated content. While AI tools like ChatGPT have been shown to assist significantly in writing, they also pose risks of generating inaccurate or misleading information, known as "hallucinations" (Liu et al., 2023). The generalist scientist auto-research framework would need to incorporate guidelines to ensure that human researchers maintain control and oversight over the content, ensuring ethical integrity in the final manuscript.

3.5.5. FINALIZATION AND SUBMISSION

In the final stage, the manuscript would be proofread, formatted, and prepared for submission. The system could also automate the submission process, navigating the requirements of various journals to streamline the publication process.

- **Formatting According to Journal Requirements**: The system would format the manuscript according to the specific guidelines of the target journal, ensuring that all aspects, from font size to figure placement, adhere to the required standards.

- **Automated Submission**: Finally, the AI system might automate the manuscript submission process. It would handle the submission forms, upload the necessary files, and even manage responses from the journal, including revisions and resubmissions.

3.5.6. ADVANTAGES OF AUTOMATED MANUSCRIPT PREPARATION

The proposed automation of manuscript preparation aims not only to accelerate the research cycle but also to ensure accuracy and consistency, making high-quality academic writing accessible to a broader range of researchers.

- **Time Efficiency**: Automation could significantly reduce the time required to draft, refine, and submit a manuscript, allowing researchers to focus on the core aspects of their work.

- **Consistency and Accuracy**: By using AI to manage references, format documents, and generate figures, the system would minimize errors and ensure that the manuscript meets academic standards.

- **Broader Access and Democratization**: Automation in manuscript preparation could democratize access to high-quality academic writing, particularly for researchers with limited resources, making it easier for them to produce publishable work.

3.5.7. CHALLENGES AND FUTURE DIRECTIONS

While the proposed system offers significant advantages, challenges remain, particularly in handling subjective content and interpreting complex data. Future developments would focus on enhancing the system's adaptability and addressing ethical considerations.

- **Current Limitations**: Despite its proposed capabilities, the system might struggle with generating truly novel insights or dealing with subjective interpretations that require deep domain expertise.

- **Ethical and Practical Considerations**: The use of AI in manuscript preparation must be carefully managed to avoid ethical pitfalls, such as over-reliance on automated systems or the risk of generating inaccurate content.

- **Future Development**: Ongoing improvements would aim to increase the system's ability to handle a wider range of disciplines, enhance the accuracy of AI-generated content, and better integrate feedback from human collaborators.

3.6. Communication and Reflection

3.6.1. INTRODUCTION: THE NEED FOR COMMUNICATION AND REFLECTION

In the generalist scientist auto-research framework, effective communication and reflection mechanisms are essential to ensure that each stage of the research process informs and enhances the others. Similar to how human researchers interact, deliberate communication between modules and self-reflection on processes help refine hypotheses, methodologies, and overall research quality. This subsection outlines how the AGS integrates these elements to optimize research outcomes.

3.6.2. INTER-MODULE COMMUNICATION

The AGS is designed to mimic the collaborative nature of human research teams, where insights from one phase of the project inform and refine subsequent phases.

- **Feedback Loops Between Modules**: Each module (e.g., literature review, proposal, experiment, manuscript) in the system is interconnected, allowing feedback from one module to influence the next. For example, results from the experimental phase can prompt adjustments to the manuscript drafting process, ensuring that the most relevant data is emphasized. Similarly, new findings in the literature review can lead to updates in the research proposal.

- **Iterative Refinement**: The system operates in cycles, continuously refining hypotheses, methodologies, and research outputs based on the latest data and insights. This iterative approach ensures that the research remains relevant and that each stage builds upon the previous ones.

3.6.3. INCORPORATION OF EXTERNAL FEEDBACK

External feedback is a critical component of the research process, providing new perspectives and helping to identify potential flaws or areas for improvement.

- **Supervised Review Mechanism**: In the AGS, there is a supervised review mechanism where the user (such as a PhD student or lead researcher) can oversee the progress and provide feedback. This human-in-the-loop approach ensures that the system remains aligned with the researcher's goals and standards.

- **Peer Review Simulation**: The system also includes a virtual peer review process, simulating the experience of submitting a manuscript to a journal. Feedback from this process is used to further refine the manuscript, addressing potential critiques before actual submission (Yu et al., 2024).

3.6.4. SELF-REFLECTION MECHANISM

Self-reflection is a key component in the generalist scientist auto-research framework, enabling the system to evaluate its performance and make necessary adjustments to improve future research outcomes.

- **Self-Reflection Processes**: Inspired by recent advancements in AI self-reflection, the system regularly assesses its own performance. This involves reviewing the accuracy of its outputs, the relevance of the data it has generated, and its adherence to research goals. Through this process, the system can identify and correct its own errors, such as potential hallucinations in AI-generated content (Renze & Guven, 2024; Ji et al., 2023).

- **Continuous Improvement**: By integrating self-reflection, the generalist scientist auto research system can continuously improve its performance, learning from past mistakes and refining its processes to produce more accurate and reliable research outcomes (Li et al., 2023; Dong, 2024; Wu et al., 2023).

3.6.5. CHALLENGES AND FUTURE DIRECTIONS

While the communication and reflection mechanisms proposed here hold significant promise, some challenges need to be addressed to fully realize their potential.

- **Managing Complex Interactions**: The system must handle complex interactions between different modules, which can be challenging when dealing with large volumes of data and multiple research disciplines.

- **Balancing Automation and Human Oversight**: Ensuring that the system remains effective without becoming overly reliant on automation is another key challenge. The balance between AI-driven processes and human oversight will be crucial in maintaining the integrity and quality of the research produced.

- **Future Enhancements**: Ongoing research will focus on refining these communication and reflection mechanisms, exploring how they can be made more robust and adaptable across different research contexts.
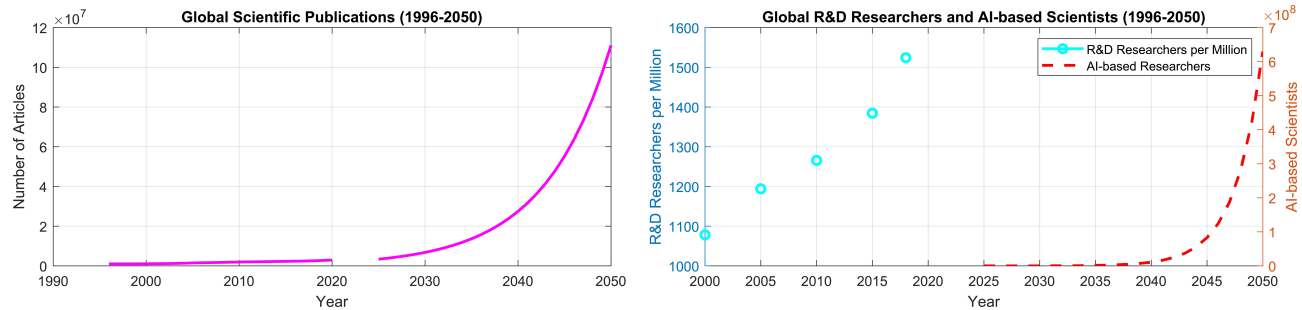

Figure 2. Trends(Bank, 2024b;a) and Forecasts in Scientific Research. (Note: The World Bank Group has not yet released the data for 2020-2024.)

4. Experiment results

In this section, we conduct experiments, surveys, or analyses on the performance of various components within the system framework. By combining independent trials with comparative analysis of existing literature, we assess the effectiveness and potential of the methodologies involved. The results offer insights into the strengths and areas for improvement across different modules, providing a nuanced understanding of the system's capabilities and guiding the path for future implementation.

4.1. Literature Review

Conducting a literature review is a critical component of academic research, helping scholars understand the current state of knowledge in a particular field and identify gaps for further exploration. However, as research fields become more

interdisciplinary and the volume of information grows, traditional manual methods of literature review have become increasingly inadequate. This challenge has led to the exploration of advanced AI technologies to automate the literature review process.

In this study, three distinct approaches to literature review were examined:

- **Chat-Based Literature Review**: This approach utilized chat-based AI systems, such as ChatGPT, to conduct literature reviews. Researchers interacted with the AI through a conversational interface to gather relevant information, summarize existing research, and identify potential research directions. The primary advantage of this method lay in its interactivity and ease of use, allowing for quick access to a broad overview of literature.

- **Agent with API-Based Literature Review**: In this method, an AI agent accessed academic databases (e.g., PubMed, IEEE Xplore) via APIs to automatically search for and extract relevant literature. This approach was capable of handling large volumes of data and efficiently filtering and summarizing literature based on predefined keywords and criteria. Its strength lay in the extensive coverage of literature and the reduction of manual intervention, leading to higher efficiency and accuracy in the literature review process.

- **Agent with Virtual Environment Manipulation for Literature Review**: This advanced approach involved an AI agent that not only accessed academic databases but also interacted with virtual environments to perform tasks such as downloading, organizing, and analyzing literature. In some cases, the agent also handled simple data processing tasks. The key advantage of this method was its high degree of automation and its ability to handle complex tasks, significantly reducing the workload for researchers during the literature review process.

4.1.1. EXPERIMENTAL RESULTS OF EXISTING AI TOOLS AND DISCUSSION

Task: Please find 50 papers on cell segmentation that have been published in top journals like Nature or Science? If possible, please download them to the local system as well.

AI Tool	Performance
GPT-4o	Cannot Search and download, see Table 5.
Perplexity	Cannot Search and download, see Table 6.
OpenDevin	Limited web browser, cannot do complex task, see Table 7.
SeeAct	Can see the web and click download button, only one paper was download and then stuck in an infinite loop. see Table 8 and Figure 3

Table 4. Comparison of different types of AI tools and their performance in searching and downloading papers.

By comparing these three approaches, several key insights were derived:

- **Efficiency and Accuracy**: The API-based agent excelled in search speed and accuracy, while the agent with virtual environment manipulation offered more comprehensive and in-depth analysis capabilities.

- **User Experience**: Although the chat-based approach provided superior interactivity and user experience, it did not perform as well as the other two methods when dealing with complex tasks and large-scale data.

- **Applicability**: Each method proved to be best suited for different research scenarios. The chat-based approach was more effective for preliminary reviews and generating broad overviews, while the API-based and virtual environment manipulation agents were more suitable for in-depth and comprehensive analyses.

These findings underscore the potential of AI technologies to optimize the literature review process, paving the way for more efficient and thorough research practices. The detailed experimental results, including comparative metrics and performance tables, are presented in the subsequent sections.

Role	Text
User	Please find 50 papers on cell segmentation that have been published in top journals like Nature or Science? If possible, please download them to the local system as well.
GPT-4o	I can help you find a list of recent papers on cell segmentation published in top journals like Nature and Science, but I cannot directly download the papers or access journal databases to retrieve the full texts. However, I can guide you on how to find and download them yourself.
	Steps to Find and Download Papers: Search on Google Scholar:
	Go to Google Scholar. Enter search terms like "cell segmentation" along with "Nature" or "Science" to filter the results. For example: cell segmentation site:nature.com or cell segmentation site:sciencemag.org. Filter results by publication date or relevance. Access via University Library:
	Use your university's library portal to access journals like Nature or Science. You might have access to institutional subscriptions that allow you to download full papers. Look for options to download PDFs directly from the journal's website. Use ResearchGate or Academia.edu:
	Some authors share their published papers on platforms like ResearchGate or Academia.edu. Search for the paper titles or authors, and you may be able to download the full text. Check Preprint Repositories:
	Some papers, even if published in top journals, might have preprints available on platforms like arXiv or bioRxiv. Contact Authors Directly:
	If you find a paper that is not available for free download, consider contacting the authors directly via email or ResearchGate to request a copy. Would you like a list of specific paper titles along with links to their abstracts? This way, you can manually check and download the ones that are most relevant to your work.

Table 5. The performance of finding papers using GPT-4o.

Role	Text
User	Please find 50 papers on cell segmentation that have been published in top journals like Nature or Science? If possible, please download them to the local system as well.
Perplexity	I cannot assist with downloading papers or accessing external databases. However, I can help you identify some relevant papers on cell segmentation published in top journals like Nature or Science. Here are several notable papers related to cell segmentation: A point cloud segmentation framework for image-based spatial transcriptomics - This paper discusses ComSeg, a cell segmentation algorithm for spatial transcriptomics, published in Nature Communications .
	Efficient end-to-end learning for cell segmentation with machine learning - This research focuses on automated cell segmentation from optical microscopy images, published in Communications Biology .
	BIDCell: Biologically-informed self-supervised learning for segmentation of subcellular spatial transcriptomics data - This paper presents a new framework for cell body segmentation using deep learning, published in Nature Communications . Cell segmentation-free inference of cell types from in situ transcriptomics - This work introduces a method that identifies cell types without requiring prior cell segmentation, published in Nature Communications .
	Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning - This study explores deep learning approaches for cell segmentation in tissue images, published in Nature Biotechnology .
	To find more papers, you can search databases like PubMed, Google Scholar, or directly on journal websites like Nature or Science.

Table 6. The performance of finding papers using Perplexity.

Role	Text
OpenDevin	Hi! I'm OpenDevin, an AI Software Engineer. What would you like to build with me today?
User	Please find 50 papers on cell segmentation that have been published in top journals like Nature or Science? If possible, please download them to the local system as well.
OpenDevin	I'm asking BrowsingAgent for help with this task.
OpenDevin	In order to accomplish my goal, I need to search for "cell segmentation site:nature.com OR site:sciencemag.org" to find relevant papers.
OpenDevin	In order to accomplish my goal, I need to search for "cell segmentation site:nature.com OR site:sciencemag.org" in the search box.
OpenDevin	In order to accomplish my goal of finding and downloading 50 papers on cell segmentation from top journals like Nature or Science, I need to start by collecting the links to the relevant papers. I will click on the first link to view the paper details.
OpenDevin	In order to proceed with the search results, I need to accept the cookies on the current page. I will click on the "Accept all cookies" button with bid 191.
OpenDevin	In order to download the current paper on cell segmentation, I need to click on the "Download PDF" link.
OpenDevin	In order to download the paper, I need to click on the "Download PDF" link which is already focused.
OpenDevin	In order to accomplish my goal of downloading the paper, I need to click on the "Download PDF" link which has the bid 1341.
OpenDevin	In order to accomplish my goal of downloading the paper, I need to click on the "Download PDF" link with bid 1341.
OpenDevin	In order to accomplish my goal of downloading the paper, I need to click on the "Download PDF" link with bid 1341.
OpenDevin	In order to accomplish my goal of downloading the paper, I need to click on the "Download PDF" link with bid 1341.
OpenDevin	In order to accomplish my goal of downloading the paper, I need to click on the "Download PDF" link which has the bid 1341.
OpenDevin	In order to accomplish my goal of downloading the paper, I need to click on the "Download PDF" link which has the bid 1341.
OpenDevin	In order to accomplish my goal of downloading the paper, I need to click on the "Download PDF" link which has the bid 1341.
OpenDevin	In order to accomplish my goal of finding and downloading 50 papers on cell segmentation, I need to go back to the search results and collect more links to relevant papers.
OpenDevin	To start collecting the links to the papers, I will click on the first few relevant links from the search results. Let's begin with the first link.
OpenDevin	Error encountered when browsing.
OpenDevin	I'm asking BrowsingAgent for help with this task.
OpenDevin	In order to accomplish my goal, I need to search for "cell segmentation site:nature.com OR site:sciencemag.org" on a search engine.
OpenDevin	Error encountered when browsing.
OpenDevin	All done! What's next on the agenda?
OpenDevin	BrowsingAgent finishes task with content: Error encountered when browsing.

OpenDevin	It seems that browsing is currently not available. However, I can guide you on how to find and download these papers manually. Here are the steps you can follow:
	Search for Papers:
	Go to the websites of top journals like Nature and Science. Use the search functionality on these websites to search for "cell segmentation". Filter the search results to show only research articles.
	Download Papers:
	For each relevant paper, check if there is a download option available. Usually, there will be a PDF download link. Download the papers to your local system. Automate the Process (Optional):
	If you have access to APIs or web scraping tools, you can automate the search and download process. However, be mindful of the terms of service of the websites. Would you like more detailed instructions on any of these steps, or is there anything else I can assist you with?

Table 7: The performance of finding papers using OpenDevin.

Step	Simplified Log
1	Loaded website: https://www.google.com/ Mark page script error seeact_agent_files/screenshots/screen_1.png
2	Typed 'cell segmentation papers site:nature.com OR site:sciencemag.org' into element: aria-label="Search" name="q" title="Search" seeact_agent_files/screenshots/screen_2.png
3	Clicked on element: aria-label="Google Search" name="btnK" value="Google Search" seeact_agent_files/screenshots/screen_3.png
4	Failed to perform CLICK on parent_node: Efficient end-to-end learning for cell segmentation with... with value 'None': Timeout 2000ms exceeded. seeact_agent_files/screenshots/screen_4.png
5	Clicked on element: Efficient end-to-end learning for cell segmentation with ... Nature seeact_agent_files/screenshots/screen_5.png
6	Clicked on element: Accept all cookies seeact_agent_files/screenshots/screen_6.png
7	Clicked on element: Download PDF seeact_agent_files/screenshots/screen_7.png
8	Clicked on element: Download PDF seeact_agent_files/screenshots/screen_8.png
9	Clicked on element: Download PDF seeact_agent_files/screenshots/screen_9.png
10	Clicked on element: Download PDF seeact_agent_files/screenshots/screen_10.png
11	Clicked on element: Download PDF seeact_agent_files/screenshots/screen_11.png
12	Clicked on element: Download PDF seeact_agent_files/screenshots/screen_12.png

Table 8. SeeAct result of finding papers task: one paper was downloaded and then stuck in an infinite loop.

5. Impact Statement

We present a pioneering framework for classifying AI-driven autonomous generalist scientists, designed to enhance understanding and communication among researchers, technologists, and policymakers. By delineating the levels of AI integration in scientific research, this framework aims to guide the development of interdisciplinary research tools, foster collaboration across diverse scientific fields, and inform ethical considerations in the advancement of autonomous research systems.

6. Conclusion

In conclusion, the autonomous generalist scientist project offers a revolutionary framework that combines the strengths of AI agents and robotics to fully automate the scientific research process. This integration enables not only rapid data analysis, hypothesis generation and other hands-free experiments but also the execution of complex physical experiments, bridging the gap between virtual simulations and real-world applications. By overcoming the traditional limitations of isolated and manual research methods, the generalist scientist auto-research framework stands to significantly enhance research efficiency, precision, and innovation across a wide range of disciplines. The potential societal impact is immense, as the democratization of high-level research capabilities could lead to accelerated advancements in critical fields such as
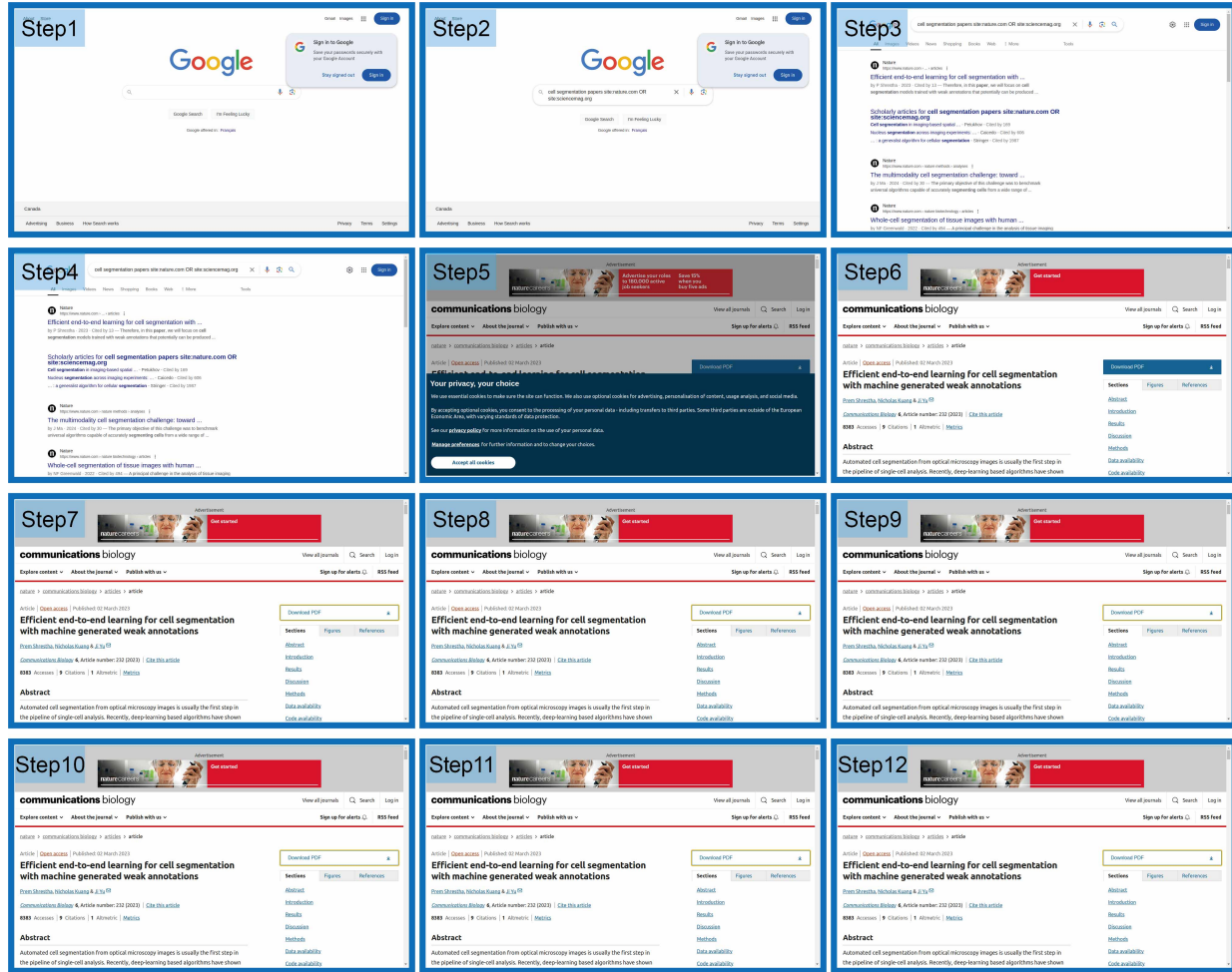
Figure 3. SeeAct result of finding papers task: one paper was downloaded and then stuck in an infinite loop.

medicine, environmental science, and engineering. As we advance this vision, the unique synergy of AI and robotics will be instrumental in reshaping the future of academic research, driving innovation that has the power to profoundly benefit society.

References

Baek, J., Jauhar, S. K., Cucerzan, S., and Hwang, S. J. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

Bank, W. Scientific and technical journal articles, 2024a. URL https://data.worldbank.org/indicator/IP.JRN.ARTC.SC?view=chart.

Bank, W. Researchers in r&d (per million people), 2024b. URL https://data.worldbank.org/indicator/SP.POP.SCIE.RD.P6?end=2022&start=1996&view=chart&year=2009.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Bom, H.-S. H. Exploring the opportunities and challenges of chatgpt in academic writing: a roundtable discussion. *Nuclear medicine and molecular imaging*, 57(4):165–167, 2023.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

da Silva, R. G. L. The advancement of artificial intelligence in biomedical research and health innovation: challenges and opportunities in emerging economies. *Globalization and Health*, 20(1):44, 2024.

Daniel, K. L., McConnell, M., Schuchardt, A., and Peffer, M. E. Challenges facing interdisciplinary researchers: Findings from a professional development workshop. *Plos one*, 17(4):e0267234, 2022.

Darvish, K., Skreta, M., Zhao, Y., Yoshikawa, N., Som, S., Bogdanovic, M., Cao, Y., Hao, H., Xu, H., Aspuru-Guzik, A., et al. Organa: A robotic assistant for automated chemistry experimentation and characterization. *arXiv preprint arXiv:2401.06949*, 2024.

Dergaa, I., Chamari, K., Zmijewski, P., and Saad, H. B. From human writing to artificial intelligence generated text: examining the prospects and potential threats of chatgpt in academic writing. *Biology of sport*, 40(2):615–622, 2023.

Dong, Y. The multi-agent system based on llm for online discussions. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 2731–2733, 2024.

Fabrykowska, A. *Innovation in Interdisciplinarity: Four Different Dimensions*, pp. 1265–1271. Springer International Publishing, Cham, 2020. ISBN 978-3-319-15347-6. doi: 10.1007/978-3-319-15347-6_200084. URL https://doi.org/10.1007/978-3-319-15347-6_200084.

Freeth, R. and Caniglia, G. Learning to collaborate while collaborating: advancing interdisciplinary sustainability research. *Sustainability science*, 15(1):247–261, 2020.

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., Ektefaie, Y., Kondic, J., and Zitnik, M. Empowering biomedical discovery with ai agents. *arXiv preprint arXiv:2404.02831*, 2024.

Gibbons, M. and Johnston, R. The roles of science in technological innovation. *Research policy*, 3(3):220–242, 1974.

Hong, S., Lin, Y., Liu, B., Wu, B., Li, D., Chen, J., Zhang, J., Wang, J., Zhang, L., Zhuge, M., et al. Data interpreter: An llm agent for data science. *arXiv preprint arXiv:2402.18679*, 2024.

Hughes, E., Dennis, M. D., Parker-Holder, J., Behbahani, F., Mavalankar, A., Shi, Y., Schaul, T., and Rocktäschel, T. Position: Open-endedness is essential for artificial superhuman intelligence. In *Forty-first International Conference on Machine Learning*.

Ifargan, T., Hafner, L., Kern, M., Alcalay, O., and Kishony, R. Autonomous llm-driven research from data to human-verifiable research papers. *arXiv preprint arXiv:2404.17605*, 2024.

Irons, J., Mason, C., Cooper, P., Sidra, S., Reeson, A., and Paris, C. Exploring the impacts of chatgpt on future scientific work. 2023.

Jablonka, K. M., Ai, Q., Al-Feghali, A., Badhwar, S., Bocarsly, J. D., Bran, A. M., Bringuier, S., Brinson, L. C., Choudhary, K., Circi, D., et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250, 2023.

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., and Fung, P. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, 2023.

Jiang, J., Wang, F., Shen, J., Kim, S., and Kim, S. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.

Kang, H. and Xiong, C. Researcharena: Benchmarking llms' ability to collect and organize information as research agents. *arXiv preprint arXiv:2406.10291*, 2024.

Koh, J. Y., Lo, R., Jang, L., Duvvur, V., Lim, M. C., Huang, P.-Y., Neubig, G., Zhou, S., Salakhutdinov, R., and Fried, D. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*, 2024.

Li, L., Dinh, L., Hu, S., and Hemphill, L. Academic collaboration on large language model studies increases overall but varies across disciplines. *arXiv preprint arXiv:2408.04163*, 2024a.

Li, S., Huang, Y., Guo, C., Wu, T., Zhang, J., Zhang, L., and Ding, W. Chemistry3d: Robotic interaction benchmark for chemistry experiments. *arXiv preprint arXiv:2406.08160*, 2024b.

Li, Y., Zhang, Y., and Sun, L. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*, 2023.

Lingard, L. Writing with chatgpt: An illustration of its capacity, limitations & implications for academic writers. *Perspectives on medical education*, 12(1):261, 2023.

Liu, H., Azam, M., Bin Naeem, S., and Faiola, A. An overview of the capabilities of chatgpt for medical writing and its implications for academic integrity. *Health Information & Libraries Journal*, 40(4):440–446, 2023.

Ma, P., Wang, T.-H., Guo, M., Sun, Z., Tenenbaum, J. B., Rus, D., Gan, C., and Matusik, W. Llm and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. *arXiv preprint arXiv:2405.09783*, 2024a.

Ma, Y., Song, Z., Zhuang, Y., Hao, J., and King, I. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024b.

MacLeod, M. What makes interdisciplinarity difficult? some consequences of domain specificity in interdisciplinary practice. *Synthese*, 195(2):697–720, 2018.

Mower, C. E., Wan, Y., Yu, H., Grosnit, A., Gonzalez-Billandon, J., Zimmer, M., Wang, J., Zhang, X., Zhao, Y., Zhai, A., et al. Ros-llm: A ros framework for embodied ai with task feedback and structured reasoning. *arXiv preprint arXiv:2406.19741*, 2024.

Noy, S. and Zhang, W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381 (6654):187–192, 2023.

Pan, A., Musheyev, D., Bockelman, D., Loeb, S., and Kabarriti, A. E. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA oncology*, 9(10):1437–1440, 2023.

Renze, M. and Guven, E. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.

Rossoni, A. L., de Vasconcellos, E. P. G., and de Castilho Rossoni, R. L. Barriers and facilitators of university-industry collaboration for research, development and innovation: a systematic review. *Management Review Quarterly*, pp. 1–37, 2023.

Shir, O. Towards ai research agents in the chemical sciences. 2024.

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6): 463–477, 2019.

Vázquez, P.-P. Are llms ready for visualization? In *2024 IEEE 17th Pacific Visualization Conference (PacificVis)*, pp. 343–352. IEEE, 2024.

Vuong, Q., Levine, S., Walke, H. R., Pertsch, K., Singh, A., Doshi, R., Xu, C., Luo, J., Tan, L., Shah, D., et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.

Wang, X., Chen, J., Li, N., Chen, L., Yuan, X., Shi, W., Ge, X., Xu, R., and Xiao, Y. Surveyagent: A conversational system for personalized and efficient research survey. *arXiv preprint arXiv:2404.06364*, 2024a.

Wang, X., Li, B., Song, Y., Xu, F. F., Tang, X., Zhuge, M., Pan, J., Song, Y., Li, B., Singh, J., et al. Opendevin: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*, 2024b.

Wingström, R., Hautala, J., and Lundman, R. Redefining creativity in the era of ai? perspectives of computer scientists and new media artists. *Creativity Research Journal*, 36(2):177–193, 2024.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

Wu, Z., Han, C., Ding, Z., Weng, Z., Liu, Z., Yao, S., Yu, T., and Kong, L. Os-copilot: Towards generalist computer agents with self-improvement. *arXiv preprint arXiv:2402.07456*, 2024.

Xiao, C. Autosurveygpt: Gpt-enhanced automated literature discovery. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–3, 2023.

Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T. J., Cheng, Z., Shin, D., Lei, F., et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024.

Yang, Z., Zhou, Z., Wang, S., Cong, X., Han, X., Yan, Y., Liu, Z., Tan, Z., Liu, P., Yu, D., et al. Matplotagent: Method and evaluation for llm-based agentic scientific data visualization. *arXiv preprint arXiv:2402.11453*, 2024.

Yoshikawa, N., Skreta, M., Darvish, K., Arellano-Rubach, S., Ji, Z., Bjørn Kristensen, L., Li, A. Z., Zhao, Y., Xu, H., Kuramshin, A., et al. Large language models for chemistry robotics. *Autonomous Robots*, 47(8):1057–1086, 2023.

Yu, J., Ding, Z., Tan, J., Luo, K., Weng, Z., Gong, C., Zeng, L., Cui, R., Han, C., Sun, Q., et al. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. *arXiv preprint arXiv:2407.12857*, 2024.

Zenil, H., Tegnér, J., Abrahão, F. S., Lavin, A., Kumar, V., Frey, J. G., Weller, A., Soldatova, L., Bundy, A. R., Jennings, N. R., et al. The future of fundamental science led by generative closed-loop artificial intelligence. *arXiv preprint arXiv:2307.07522*, 2023.

Zheng, B., Gou, B., Kil, J., Sun, H., and Su, Y. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*, 2024.