AIGS: GENERATING SCIENCE FROM AI-POWERED AUTOMATED FALSIFICATION

Zijun Liu[1*], **Kaiming Liu**[1*], **Yiqi Zhu**[1*], **Xuanyu Lei**[1,2*], **Zonghan Yang**[1*], **Zhenhe Zhang**[1], **Peng Li**[2], **Yang Liu**[1,2]
[1]Department of Computer Science & Technology, Tsinghua University
[2]Institute for AI Industry Research (AIR), Tsinghua University

ABSTRACT

Rapid development of artificial intelligence has drastically accelerated the development of scientific discovery. Trained with large-scale observation data, deep neural networks extract the underlying patterns in an end-to-end manner and assist human researchers with highly-precised predictions in unseen scenarios. The recent rise of Large Language Models (LLMs) and the empowered autonomous agents enable scientists to gain help through interaction in different stages of their research, including but not limited to literature review, research ideation, idea implementation, and academic writing. However, AI researchers instantiated by foundation model empowered agents with full-process autonomy are still in their infancy. In this paper, we study ***AI-Generated Science*** (AIGS), where agents independently and autonomously complete the entire research process and discover scientific laws. By revisiting the definition of scientific research (Popper, 1935), we argue that ***falsification*** is the essence of both human research process and the design of an AIGS system. Through the lens of *falsification*, prior systems attempting towards AI-Generated Science either lack the part in their design, or rely heavily on existing verification engines that narrow the use in specialized domains. In this work, we propose BABY-AIGS as a baby-step demonstration of a full-process AIGS system, which is a multi-agent system with agents in roles representing key research process. By introducing FALSIFICATIONAGENT, which identify and then verify possible scientific discoveries, we empower the system with explicit ***falsification***. Experiments on three tasks preliminarily show that BABY-AIGS could produce meaningful scientific discoveries, though not on par with experienced human researchers. Finally, we discuss on the limitations of current BABY-AIGS, actionable insights, and related ethical issues in detail.[1]

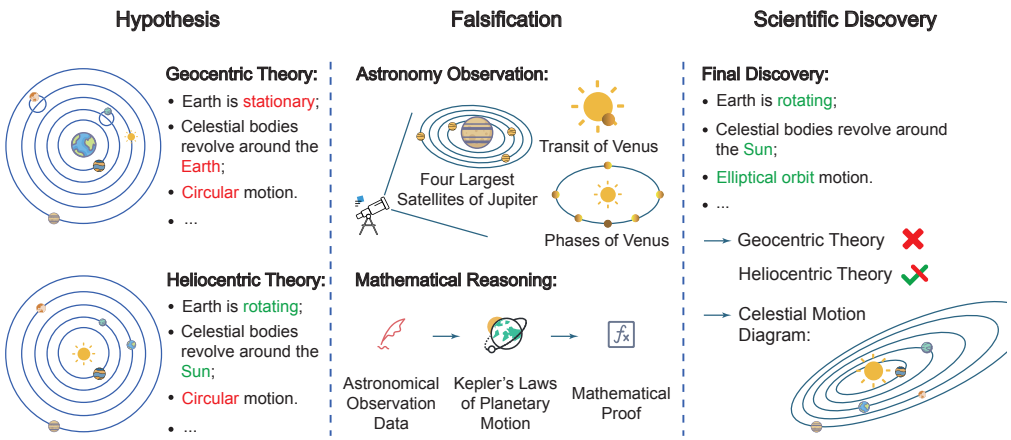

Figure 1: Examples of scientific research processes conducted by human researchers. Explicit falsification serves as a vital stage to falsify or verify the proposed hypotheses from either empirical or theoretical experiments, leading to the ultimate scientific discovery.

[*]indicates equal contribution.

[1]Official Website: `https://agent-force.github.io/AIGS/`. Code is released at `https://github.com/AgentForceTeamOfficial/Baby-AIGS`.

CONTENTS

1 Introduction **3**

2 The Development of AI-Accelerated Scientific Discovery **4**

2.1 AI as a Performance Optimizer: Discoveries in Specific Tasks 4

2.2 AI as a Research Assistant: Co-pilot in Human-AI Collaboration 5

2.3 AI as an Automated Scientist: Towards End-to-end Scientific Discovery 6

2.4 AI forms a Research Community: Enable Academic Swarm Intelligence 6

3 BABY-AIGS: A Baby Step Towards Full-Process AIGS **6**

3.1 Design Principles of a Full-Process AIGS System 6

3.2 BABY-AIGS System Design 7

3.3 Detailed Implementation 8

3.3.1 Domain-Specific Language (DSL) 9

3.3.2 PROPOSALAGENT 10

3.3.3 REVIEWAGENT 11

3.3.4 Multi-Sampling Strategy 12

3.3.5 FALSIFICATIONAGENT 13

3.4 Automated Full-Process Research Experiment 15

3.4.1 Selected Research Topics 15

3.4.2 Evaluation Settings 15

3.5 Quantitative and Qualitative Analysis 17

3.6 Discussions 18

4 Limitations and Actionable Insights **19**

5 Ethics and Impact Statement **20**

5.1 Potential Negative Impacts of AIGS Systems 20

5.2 Strategies for Responsible and Ethical Development of Automated Research Systems 21

6 Conclusion **22**

A Implementation Details of the BABY-AIGS system **29**

A.1 Research-Agnostic Implementation 30

A.2 Research-Specific Implementation 30

B Experiment Details **31**

B.1 Guidelines for Human Evaluators 31

B.2 API Costs of the Full-Process Research Experiment 31

B.3 DSL Demonstrations for Different Research Topics 31

C Prompting Structure **33**

1 INTRODUCTION

Deep learning has revolutionized scientific research (LeCun et al., 2015; Vaswani et al., 2017; Jumper et al., 2021; Achiam et al., 2023). Leveraging the enormous amount of experimental data, deep learning methods extract the underlying patterns in an end-to-end manner and effectively generalize to unobserved scenarios. The breakthroughs from deep learning in scientific domains, such as protein structure prediction (Jumper et al., 2021), gravitational wave detection (George & Huerta, 2018), and plasma control (Degrave et al., 2022), have received award-winning recognition. As a result, AI for Science has emerged as a highly-regarded research field (Wang et al., 2023a).

In the paradigm of AI for Science, AI primarily serves as a tool to assist researchers in making discoveries. With the rapid development of foundation models and autonomous agents (Park et al., 2023), AI techniques nowadays boast the capabilities of general-purposed textual understanding and autonomous interaction with the external world. These capabilities lead to the successful applications of AI-as-research-assistants, ranging from single-cell analysis (Hou & Ji, 2024) to drug discovery (Wang et al., 2023b). The capability of providing research assistance leads to a more ambitious challenge: *Can foundation model-powered agents be autonomous researchers, independently completing the entire process of scientific discovery, thereby* ***transforming AI for Science into AI-Generated Science (AIGS)****?*

When constructing an AIGS system with full-process autonomy, the desiderata of the system design should refer to the definition of the scientific research process itself. As stated by Popper (1935), scientific research follows a systematic process of proposing novel hypotheses, conducting experiments through trial and error, and falsifying these hypotheses to conclude. While it is widely-believed that ***creativity*** is indispensable in the process of research - which is also accounted in previous work (Si et al., 2024) - the central component of scientific research is ***falsification***: designing and executing experiments to validate or refute hypotheses, and falsified hypotheses pose positive contributions to scientific progress as well[2]. Moreover, experienced researchers accumulate practical skills or reusable workflows (Gil et al., 2007) from hands-on experimentation, which eases the design and execution of experiments and hypothesis falsification. The abstraction of workflows in experiments enables effective reuse, which reflects a high level of ***executability*** in scientific research. To recapitulate, a creative idea is the beginning of a piece of scientific research, which is followed by experiments and analyses to be conducted; executability forms the basis for falsification, and a sequence of logically consistent falsification processes turns a novel idea into scientific discoveries with genuine creativity. As a result, ***falsification*** is the foundation of AI-Generated Science, pillared by experimenting scaffolds accounting for ***executability*** and targeting at the ultimate goal of research ***creativity***.

Several preliminary works have been proposed to explore the potential of AIGS, which can be roughly divided into three lines. In the first line, researchers evaluate and improve the capability of LLMs to generate research ideas with high ***creativity*** (Si et al., 2024; Hu et al., 2024b). The second line emphasizes the ***executability*** of research experiments, e.g., benchmarks like MLAgentBench (Liu et al., 2023) and MLE-Bench (Chan et al., 2024) aim to evaluate the agentic ability of LLMs to achieve high performance on the provided benchmarks via code generation. These two lines of research investigate distinct sub-stages in the research process, failing to address the full-process autonomy. The third line of research attempts to construct end-to-end AIGS systems that cover both ***creativity*** and ***executability***. MLR-copilot (Li et al., 2024b) takes existing research papers as input, and produces execution results by both generating ideas and implementing experiments. AI Scientist (Lu et al., 2024) further claims to be able to organize the generated ideas and experimental results into research papers as the output. This line of research arouses significant excitement in the community, but is feedbacked with controversy: Criticisms include the incremental nature of the generated knowledge "tweaks", as well as the poor quality of the generated code and the paper presentation[3]. Indeed, as further benchmarked by DiscoveryWorld (Jansen et al., 2024) and ScienceAgentBench (Chen et al., 2024d), an automatic AIGS system that produces novel research in an end-to-end manner is still in the early stages, with significant gaps remains underexplored, especially in the area of autonomous ***falsification***. Furthermore, while specialized systems like AlphaGeometry (Trinh et al., 2024) have achieved striking domain-specific performances, they rely heavily on the existing verification engines, which alleviate the need of autonomous falsification by AI itself.

In this work, we initiate BABY-AIGS, our baby-step attempt toward a full-process AIGS system. BABY-AIGS comprises several LLM-powered agents, including PROPOSALAGENT, EXPAGENT,

[2] https://ml-retrospectives.github.io/.

[3] https://x.com/jimmykoppel/status/1828077203956850756.

REVIEWAGENT, FALSIFICATIONAGENT, etc., each responsible for distinct stages within the research workflow, mimicking the full-process human research that falsifies hypotheses based on empirical or theoretical results for scientific discoveries. BABY-AIGS operates in two phases: the first phase iteratively refines proposed ideas and methods through enriched feedback, incorporating experimental outcomes, detailed reviews, and relevant literature. The second phase emphasizes explicit ***falsification***, a key feature absent in prior systems (Lu et al., 2024), executed by FALSIFICATIONAGENT. Based on experimental results related to the proposed methodology, the agent identifies critical factors likely contributing to notable experimental phenomena, formulates hypotheses, and ultimately produces scientific insights verified through ablation experiments. Additionally, we introduce a Domain-Specific Language (DSL) (Mernik et al., 2005) for PROPOSALAGENT to articulate ideas and methodologies in an executable format, enhancing research ***executability***—particularly during experiments. We observe that multi-sampling proposals combined with re-ranking based on validation benchmarks can enhance the ***creativity*** of methodologies developed during BABY-AIGS 's first phase. We apply BABY-AIGS across three tasks: data engineering, self-instruct alignment, and language modeling. Preliminary experimental results indicate that BABY-AIGS can autonomously produce meaningful scientific discoveries from automated falsification, supported by qualitative analysis. We also observe consistent performance improvements during iterative refinement of methods proposed by BABY-AIGS. Nevertheless, current performance remains below the results achieved by experienced researchers in top academic venues, suggesting avenues for further enhancement.

2 THE DEVELOPMENT OF AI-ACCELERATED SCIENTIFIC DISCOVERY

In this section, we review and envision the development of AI-accelerated scientific discovery as four paradigms (Figure 2): (I) **AI as a Performance Optimizer**, where deep neural networks are trained with large-scale observation data in a specific scientific problem to extract the patterns in an end-to-end manner. In this paradigm, the AI techniques are used to optimize the specific prediction / regression performance in the pre-defined scientific problem with the consideration of out-of-domain generalization. (II) **AI as a Research Assistant**, where LLM-driven research copilots are used to assist the human research process. The synergy between Paradigm (I) and (II) forms the AI-powered acceleration of scientific discovery nowadays. (III) **AI as an Automated Scientist**. In this regime, foundation model empowered agents with scientist-like behavior should complete the entire research process, ranging from the initial idea proposal to the ultimate delivery of the scientific findings. (IV) **AI Forms a Research Community**. Upon the prosperity of fully-autonomous AI researchers depicted in the previous stage, we envision the collaborations among the agentic researchers foster an AI-formed research community.

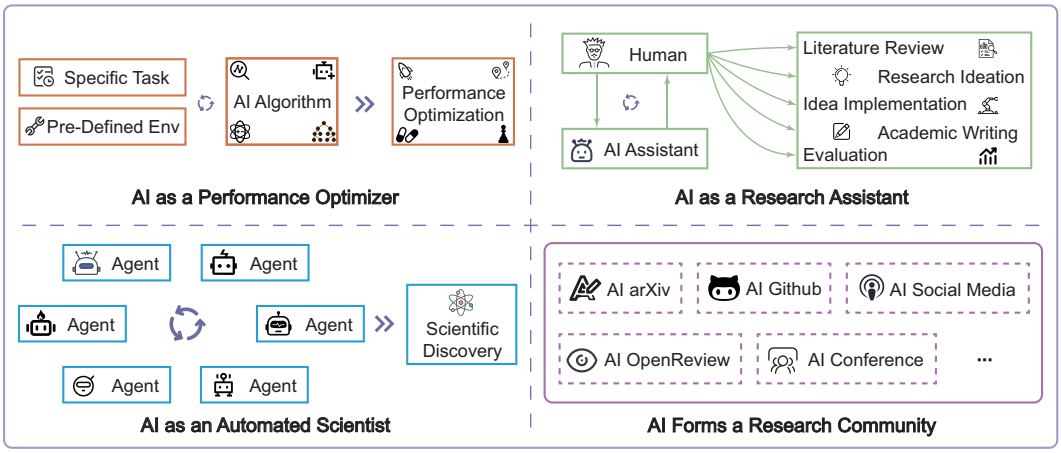

Figure 2: Overview of the four paradigms of AI-accelerate scientific discovery systems.

2.1 AI AS A PERFORMANCE OPTIMIZER: DISCOVERIES IN SPECIFIC TASKS

With the rise of deep learning, AI has significantly impacted scientific discoveries across various fields, particularly in optimizing specific tasks by exploring well-defined search spaces or extracting patterns from piles of data. Utilizing specialized deep learning models, scientific breakthroughs continue to emerge across diverse fields, including accurate protein structure prediction (Jumper et al., 2021; Abramson et al., 2024), drug discovery and materials design (Gilmer et al., 2017;

Juan et al., 2021), and the simulation of physical systems (Sanchez-Gonzalez et al., 2020). Moreover, a longstanding open problem in mathematics has been resolved through training a specialized Transformer-based expert (Alfarano et al., 2024). It is widely recognized that deep learning models are highly effective in learning representations and patterns from data, enabling scientific discovery when appropriately guided.

Large Language Models (LLMs), equipped with extensive world knowledge and advanced reasoning, are emerging as increasingly creative and autonomous agents. They have demonstrated remarkable proficiency in autonomously developing evolutionary strategies for instruction datasets (Zeng et al., 2024), identifying and rectifying their own weaknesses (Cheng et al., 2024; McAleese et al., 2024), and optimizing organizational structures for improved efficiency (Zhang et al., 2024a; Hu et al., 2024a), highlighting their potential for performance optimization through structured search. Beyond language tasks, their creativity contributes to impressive discoveries in scientific fields. Via scientifically oriented, logically organized searches, LLMs can be guided to discover mathematical solutions (Romera-Paredes et al., 2024) and physical equations (Ma et al., 2024; Shojaee et al., 2024). Augmented with specialized tools and verification engine, LLMs are capable of solving advanced geometry problems (Trinh et al., 2024), designing chemical reactions (Chen et al., 2024a) and discovering novel materials (M. Bran et al., 2024; Ghafarollahi & Buehler, 2024).

2.2 AI AS A RESEARCH ASSISTANT: CO-PILOT IN HUMAN-AI COLLABORATION

Equipped with expanding scientific knowledge and generative capabilities, LLMs gradually exhibit great potential to assist researchers at various stages of the research process.

Literature review is a fundamental but tedious step for scientific research, highlighting the need for autonomous agents for this task. Advanced LLMs are employed to identify relevant literature for a given research topic and generate structured summaries (Haman & Školník, 2024; Huang & Tan, 2023). For instance, Sharma et al. (2021) introduces a retrieval-augmented framework to produce reliable summaries based on latest studies. Furthermore, Hsu et al. (2024) utilizes LLMs to organize scientific studies within hierarchical structures and Li et al. (2024d) develops an agentic pipeline that produces comparative literature summaries guided by human workflows. In summary, LLM-based agents have demonstrated the capability to produce readable and detailed literature reviews.

For **research ideation**, LLMs are employed to generate reasonable hypotheses (Wang et al., 2024a; Qi et al., 2023; Zhou et al., 2024) based on internal knowledge and supplementary inputs. To compare the quality of LLM-generated ideas with human experts, a large-scale human study (Si et al., 2024) finds that LLMs can generate research ideas of higher novelty but slightly weaker feasibility. Furthermore, Kumar et al. (2024) and Girotra et al. (2023) evaluate the idea generation capabilities of different LLMs and recognize their potential to serve as the sources of inspiration. To enhance LLM-driven ideation, Baek et al. (2024), Nigam et al. (2024a) and Nigam et al. (2024b) develop multi-agent ideation frameworks based on scientific literature, generating novel research proposals to accelerate the life-cycle of research process. Despite these advancements, generating ideas that balance both novelty and feasibility remains a significant challenge for LLM-based agents (Si et al., 2024). To evolve initial proposals into validated knowledge therefore demands substantial effort.

The attempts in AI-assisted **idea implementation and auto-experimentation** are usually conducted as repo-level coding tasks, given the growing coding capabilities of LLMs. Focused on research-related repo-level coding, Jimenez et al. (2024), Liu et al. (2023) and Chan et al. (2024) present challenging coding benchmarks targeting machine learning and software engineering tasks. Meanwhile, Yang et al. (2024a), Wang et al. (2024b) and Tao et al. (2024) leverage agentic collaboration to automated coding from language instructions, offering promising avenues to reduce researchers' coding workloads and enhance efficiency. However, the vision for agents to autonomously implement novel ideas and conduct experiments end-to-end imposes significantly higher demands on coding agents. Current challenges include a relatively low success rate Lu et al. (2024) and frequent misalignment between proposed ideas and their coding implementations, highlighting the need for improvements in both execution reliability and alignment with research objectives.

In the realm of **academic writing**, LLMs can be utilized for drafting structured outlines, refining human-written texts and presenting research findings. Recent studies (Liang et al., 2024b; Geng & Trotta, 2024) have demonstrated a steady increase for LLM usage in scientific writing. This trend presents both opportunities and challenges for academia. When properly used, LLMs could improve research efficiency and presentation; But when misused, risks emerge as well in terms of research integrity. Therefore, effective oversight through detection strategies (Liang et al., 2024a;

Yang et al., 2024b; Ghosal et al., 2023) and watermarking techniques (Kirchenbauer et al., 2023; Zhao et al., 2023; Zhang et al., 2024b) is both beneficial and necessary.

Additionally, following LLM-as-judge methods (Zheng et al., 2023), LLM-based agents are employed for comprehensive **evaluation** on research outputs (Lu et al., 2024; Li et al., 2024b). Comparing model-generated reviews with expert evaluations, researchers have evaluated the capabilities of LLMs to provide insightful and high-quality reviews by constructing meticulously annotated datasets (Du et al., 2024) or training preference models (Tyser et al., 2024). With multi-agent collaboration to promote in-depth analysis and constructive feedback, D'Arcy et al. (2024), Jin et al. (2024) and Yu et al. (2024) develop LLM-powered agent pipelines to perform paper reviews, helping researchers improve the quality of their papers. Furthermore, Sun et al. (2024) introduces a reviewing tool designed to support reviewers with knowledge-intensive annotations. In a notable development, ICLR conference adopt reviewer agents to provide constructive feedback on human-submitted reviews, showcasing a promising application of AI-assisted reviewing [4]. Recently, researchers also constructed benchmarks for AI as a research assistant at more than one stages above (Lou et al., 2024). Overall, it is promising for LLMs to assist researchers with reliable research feedback.

2.3 AI as an Automated Scientist: Towards End-to-end Scientific Discovery

Structured in well-organized agentic pipelines, LLMs are increasingly capable of tackling complex tasks collaboratively, with end-to-end scientific research being one of the most ambitious and challenging applications. For instance, Lu et al. (2024) develops an iterative multi-agent framework that supports the entire research process, from proposing novel ideas to presenting polished findings. Similarly, Li et al. (2024b) introduces an automated research system for machine learning, and Manning et al. (2024) employs LLMs to simulate scientists for social science research. Beyond research systems, Jansen et al. (2024) proposes a simulation environment designed to challenge agents in automated scientific discovery. Despite these advancements, current end-to-end research systems still fall short of generating falsifiable scientific findings, constrained by the capabilities of both designed framework and foundation models. While previous research (Lu et al., 2024) has yielded well-formulated outcomes, the vision of automated science discovery still requires further efforts.

2.4 AI forms a Research Community: Enable Academic Swarm Intelligence

Throughout human history, scientific progress has been greatly driven by collaboration, connection, and discussion among scientists, highlighting the power of a vibrant research community. We propose that a research community of AI scientists could significantly accelerate the pace of automated scientific discovery. For agentic community construction, LLM-driven agents can be organized to generate believable, human-like behaviors (Park et al., 2022; Gao et al., 2024; Park et al., 2023) and to perform specific roles as assigned (Li et al., 2024a; Hua et al., 2023; Xu et al., 2023). Although agent-based simulations of research communities are in an early developmental stage, they represent a promising avenue for the future of fully automated, AI-driven research.

3 Baby-AIGS: A Baby Step Towards Full-Process AIGS

In this section, we elaborate how a baby-step system towards the full-process AIGS is designed, in terms of design principles, overall system design, and detailed implementations.

3.1 Design Principles of a Full-Process AIGS System

The typical research process for human scientists (Popper, 1935) generally consists of two main stages: the pre-falsification stage, which encompasses exploration of research ideas, refinement of methodologies, and theoretical or empirical analysis, and the falsification stage, which involves hypothesizing scientific laws and validating these hypotheses based on theoretical or empirical findings. In research fields like machine learning, empirical results for falsification process, i.e. ablation studies, are collected after researchers design and build a system, and conduct experiments. In contrast, other fields operate differently. For example, in physics or biology, empirical results are gathered from instruments or equipment after the experimental design and execution, while in mathematics or the humanities, theoretical insights are often derived through logical reasoning or literature review rather than empirical experimentation. These root falsification processes of different subjects in distinct knowledge source. In this work, we primarily focus on empirical subjects

[4] https://blog.iclr.cc/2024/10/09/iclr2025-assisting-reviewers.

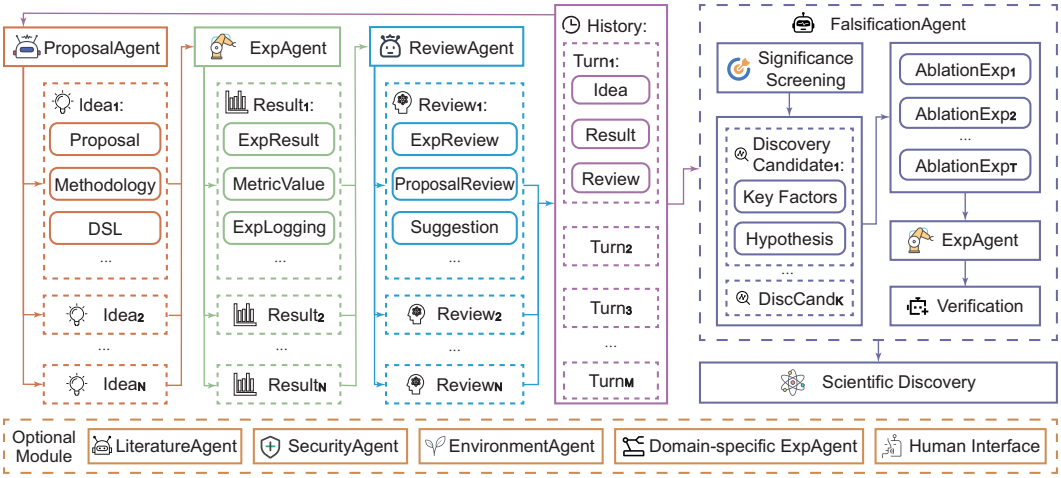

Figure 3: Overview of our BABY-AIGS system design. The left part denotes **Pre-Falsification** phase, where PROPOSALAGENT iteratively refine the proposed idea and methodology based on empirical and verbose feedback from EXPAGENT, REVIEWAGENT, etc. The iterative process summons multi-turn logs as the history context, based on which FALSIFICATIONAGENT could produce scientific discovery in the **Falsification** phase, as shown in the right part. Other modules are optional for the automated full-process research.

that requires actual implementation of the methodology of a research idea to obtain empirical results for falsification process, e.g., machine learning, and leave other venues for future work.

Human scientific research workflow above reflects the design principles of a full-process AIGS system, which are ***falsification***, ***creativity***, and ***executability***. Each of the principle could be bridged with a specific stage in the research workflow: (1) Ablation studies are fundamentally established upon ***falsification***, verifying any key factors that contribute to significant experimental results. (2) To achieve smooth and consistent experimentation, we emphasize the importance of ***executability*** of the proposed methodology, which serves as the basis for collecting empirical results for both method refinement and ablation studies. (3) ***Creativity*** of the proposed idea is the overall objective of the research process, which could be achieved through idea refinement and be identified by ***falsification*** process. **We especially argue that the process of *falsification* is equally, if not more, critical in AI-powered automated scientific discovery systems**, given that human trust in AI-generated findings relies heavily on a convincing falsification process that ensures scientific rigor and transparency.

In sum, ***falsification*** is the foundation of a full-process AIGS system, pillared by experimenting scaffolds accounting for ***executability*** and targeting at the ultimate goal of high research ***creativity***.

3.2 BABY-AIGS SYSTEM DESIGN

Heading towards a full-process system for automated scientific discovery, we present the design of BABY-AIGS system in this section. We imitate the practice of human researchers and shape it into an LLM-powered multi-agent system. And we also take into account the capacity and behaviors of current foundation models to ensure the executability in implementation.

The overall input for the system would be the topic of the research field, an accessible and configurable experiment environment, and other optional resources like a literature base; and the final outcome would be a verbal scientific discovery and the falsification process that support or falsify it. Following the principles in Section 3.1, the BABY-AIGS system operates in two phases (Figure 3):

1. **Pre-Falsification**: This phase contains several stages, such as *idea formation*, *methodology design*, *experiment execution*, *result analysis*, etc., and operates iteratively for M turns, aiming to explore and refine the proposed idea and method through feedback including experimental outcomes, reviews, etc. Specifically, the experimental results of turn 0 is from a trivial methodology at the default setting, e.g., no operation, identical mapping, etc. The multi-turn log of agent communications is recorded for **Falsification**. For better efficiency, this phase could be conducted in parallel in N threads by sampling multiple times, and the best ones for the next phase could be identified with experimental results.

2. **Falsification**: This phase aims to explicitly execute falsification by automating *ablation studies*. The agent hypothesizes on what key factors are and how they might related to significant experimental phenomenon, and the ones pass T designed ablation experiments are verified as final scientific discoveries. This could be also be K-parallel.

In the following sections, we elaborate important components of our BABY-AIGS system. Ahead of specific modules, we introduce the Domain-Specific Language (DSL) (Mernik et al., 2005). In an BABY-AIGS system, the DSL acts a critical role to ensure the automated pipeline is errorless. Specifically, the DSL is a human-designed descriptive language which can help interpret the proposed idea and methodology into executable experimental instructions through a pre-defined action space. For instance, in a deep learning task, the DSL can directly be the codes that arrange training schedule of a model; While in a chemistry experiment, the DSL can be the interface with a certain instrument or material. Consequently, the DSL bridges the gap between formulation of proposed idea and experimentation, aligning the BABY-AIGS system to the executability principle.

Here, we briefly depict the modules that construct the pipeline of BABY-AIGS:

- PROPOSALAGENT is the module to propose ideas and methods within our system. It takes the detailed description of the task, the record of past experiments, and the review generated by REVIEWAGENT as input, and outputs a proposal containing the idea, verbal and DSL-format methodology, and other necessary components to carry out the experiment for EXPAGENT. It could iteratively interact with EXPAGENT to refine its proposal in order that the experiment can be successfully completed based on its proposal.
- EXPAGENT is responsible for experiment execution in the BABY-AIGS system. It receives the proposal from PROPOSALAGENT and interprets DSL the components relevant to the experiment into executable code. After execution, it transmits the experimental result as well as the whole process of the experiment to REVIEWAGENT for review and analysis.
- REVIEWAGENT reviews the proposed idea and method based on the empirical results. It takes the whole record of both the experiments and the proposals as inputs, and generates the multi-granular review content. The review is then returned to PROPOSALAGENT for the next iteration of refinement. Through this iterative process between agents above in the **Pre-Falsification** phase, creativity of the proposed idea evolves in tandem.
- FALSIFICATIONAGENT is responsible for doing the ablation studies and deriving scientific discoveries as the final outcome. FALSIFICATIONAGENT takes the multi-turn log of all other agents as input. It has access to the record of the whole process of **Pre-Falsification** phase, and hypothesize possible key factors influencing significant experimental phenomenon based on empirical results. Then, it designs and conducts ablation experiments for T times to verify the hypothesis, leading to final scientific discoveries.
- Other optional modules include LITERATUREAGENT, SECURITYAGENT, ENVIRONMENTAGENT, DOMAIN-SPECIFIC EXPAGENT, and HUMAN INTERFACE. LITERATUREAGENT is responsible for gathering and providing relevant literature to support all other agents. SECURITYAGENT ensures safe experiment execution by identifying and preventing actions that may pose potential hazards or infringe upon intellectual property rights. ENVIRONMENTAGENT creates simulated environments to facilitate the testing and refinement of ideas, enabling more controlled and accurate scientific discoveries. DOMAIN-SPECIFIC EXPAGENT is a customizable agent tailored for specific fields. HUMAN INTERFACE allows different agents in the system to ask human researchers for help when necessary.

We also acknowledge that the implementation of BABY-AIGS at the current stage has various limitations towards a general functionable full-process AIGS system. In Section 4, we outline these limitations and discuss actionable insights for future improvements.

3.3 DETAILED IMPLEMENTATION

In the following sections, we elaborate on the the detailed implementation of our AIGS system through DSL, multi-sampling strategy, and three main agents: PROPOSALAGENT, REVIEWAGENT, and FALSIFICATIONAGENT. The rest of optional modules have been omitted for the sake of clarity. In order to aid in the elaboration of the following sections, we present the research topic of data engineering (Liu et al., 2024; Chen et al., 2024b; Li et al., 2024c; Zhao et al., 2024), which requires BABY-AIGS to identify key distinguishing features of datasets, and filter and extract high-quality data subsets. Implementation details are elaborated in Appendix A and Appendix C.

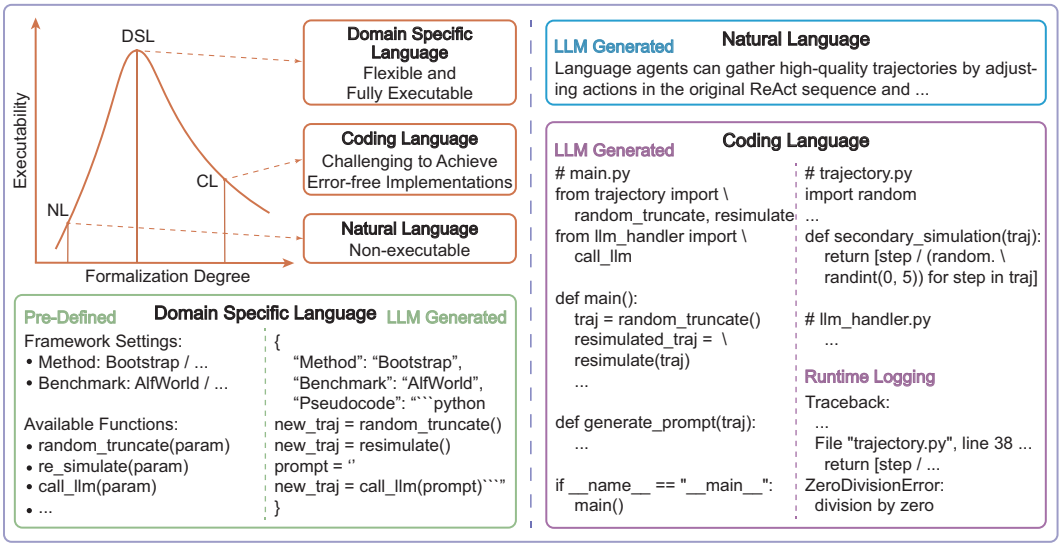

Figure 4: The relationship between formalization degree and system executability when expressing ideas through Natural Language (NL), Coding Language (CL), and Domain-Specific Language (DSL), illustrated with examples. NL expresses ideas in the simplest and most flexible form but is non-executable; CL offers greater precision but is challenging to achieve error-free implementation; DSL achieves a better tradeoff between flexibility and executability.

3.3.1 Domain-Specific Language (DSL)

A domain-specific language (Mernik et al., 2005) is created specifically for a particular application domain, providing greater expressiveness and ease of use within that domain compared to general-purpose languages, traditionally for programming languages. However, we observed that the situation is the same for agents in the AIGS systems. When conducting scientific research, agents have access to a wide and diverse action space, making it challenging to perform error-free long-sequence actions for every stage of the research process, particularly when translating the methodology into executable actions for experimentation. For instance, in machine learning research, an agent may edit multiple code files and manipulate large amount of data, as part of the methodology execution. However, limited by the current capacity of foundation models, it remains a severe challenge for agents to carry out the proposed experiment with both full-process autonomy and satisfiable success rates (Jimenez et al., 2024; Chan et al., 2024; Lu et al., 2024) without dedicated interface design (Yang et al., 2024a; Wang et al., 2024b) or tool use (Paranjape et al., 2023; Qin et al., 2024).

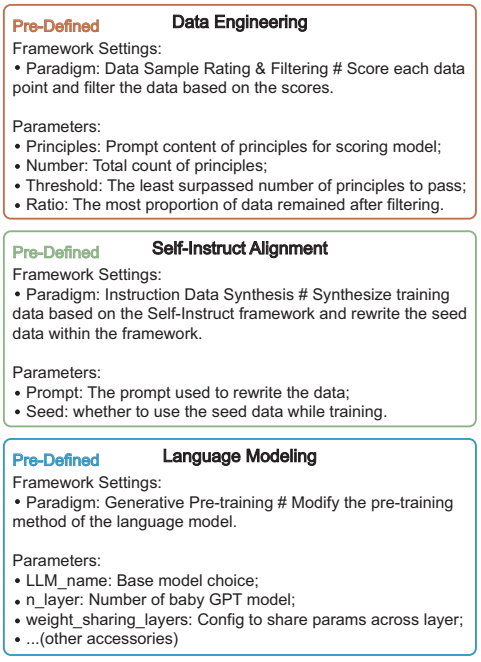

Figure 5: The DSL design in Baby-AIGS for experimented research topics in Section 3.4. The full demonstration is in Appendix B.

In Baby-AIGS, we extend the original definition of DSL in programming to semi-structure objects with pre-defined grammars, making it a bridge that fills the gap between the proposed methodology and experimentation. The DSL restricts the action space of the agents while maintaining the freedom for agents to conduct proposed methods at the same time, through dedicated design with human effort. To utilize the capabilities of current LLMs in natural language and function-level coding, we design the semi-structured grammar to be flexible between verbal instructions and structured statements. As shown in Figure 4, the DSL has both a higher degree of formalization and executability than natural language; compared to the coding language adopted in previous work (Lu et al., 2024), though DSL has a lower degree of formal-

ization, with human effort, it exhibits higher executability and thus ensures successful execution of experiments, according to empirical analysis (Section 3.5). However, when the grammar is poorly designed, the DSL is likely to restrain the creativity of the system, because some ideas might not be able to be implemented, which is a limitation of BABY-AIGS for future work.

We present the pre-defined grammar of DSL used in a few selected research topics in Figure 5. Under a specific paradigm related to the research topic, the grammar contains a series of parameters in either structured statement, e.g., code, integers, etc., or natural language, collectively depicting the methodology under the paradigm. PROPOSALAGENT would select a research paradigm when there are multiple, and fill out each parameter as required in the grammar. EXPAGENT is equipped with a pre-defined interpreter to translate the DSL into executable code lines, or inputs to specific LLMs or other models. For instance, one parameter of the DSL for data engineering is a few lines of data rating principles represented in natural language, and the model architecture parameters for language modeling still remains in codes, indicating the flexibility of DSL design. Please refer to Section 3.4 for detailed formulation of the research topics and topic-specific DSL designs.

3.3.2 PROPOSALAGENT

An example of the proposal from PROPOSALAGENT

Idea & Methodology

Idea: ...Key issues identified include overly brief or excessively lengthy answers, lack of unique words, irrelevant content, poor adherence to instructions, lack of coherence, low keyword overlap, and poor sentiment balance...

Methodology: **Key metrics to observe** include the *coherence of responses*, *adherence to instructions*, *relevance to the prompt*, *depth of information provided*, *clarity of instructions and responses*, *engagement in the conversation*...

Experiment Settings

Baseline: Iteration 0 *(the trivial baseline)*
Thought: ... we will filter the original dataset using the refined DSL with weighted criteria. ... and this will help in identifying the initial impact of the new criteria on the raw data and ensure that the dataset is not overly biased by similarity...

Hypothesis & Related Feature

Hypothesis: After using the processed data, the model's performance on the MT-bench task will **improve significantly**. The model should produce longer, more detailed, and coherent responses, ... The responses should be rich in unique words, and demonstrate appropriate sentiment balance compared to the baseline.

Related Feature: ... length of responses, keyword overlap, unique word count, and sentiment balance.

Rebuttal

The review should provide an overall view of the experiment result, focusing on whether the selected examples effectively demonstrate improvements in the key metrics. The review should compare the performance of the model before and after the data curation to highlight the impact of the methodology. Specific examples should be used to illustrate both improvements and remaining issues to provide ...

As the first step towards the scientific research, idea formation and methodology design usually lay the foundation for valuable insights or impactful discoveries from falsification process based on empirical results, i.e., ***creativity*** in the AIGS system. We refer to the corresponding module in BABY-AIGS as PROPOSALAGENT, drawing inspiration from human practice of proposing an idea and formulating the methodology before starting the experiments.

Metric	Level	Description	Execution
Length	Corpus	The length and word count of responses	Pre-defined statistic function
Keyword Overlap		The keyword overlap between instructions and responses	
Sentiment		The contained sentiment in model-generated responses	NLTK (Bird & Loper, 2004)
Worst Data Points	Sample	The worst rating samples compared with baselines	Ranking & reciting function
Best Data Points		The best rating samples compared with baselines	
.....	Corpus / Sample	Other useful metrics generated by REVIEWAGENT or pre-defined by researchers	Free-form code segment

Table 1: Examples of multi-level metrics for REVIEWAGENT to empirically review the experimental results and the proposal from PROPOSALAGENT in the data engineering research.

PROPOSALAGENT is important part of the pre-falsification phase. It takes the detailed description of research topic, the history log, including records of previous proposals and experiments, and the review from REVIEWAGENT as the overall input, except for the first iteration, in which only the description of the research topic is the input to PROPOSALAGENT. As shown in the case above on the data engineering research topic, the output of PROPOSALAGENT includes

- the proposed *idea and methodology*, that the former is a high-level thought and the latter is a semantically equal but concise description of instructions to be carried out in the experiment in natural language and DSL format, aiming either to improve the experimental results or to advance towards scientific discoveries,
- the configurable *experiment settings*, such as specifying which turn's proposal is considered the baseline for the current iteration, along with other options specific to the research topic,
- *hypothesis* on how would the experimental results change compared to and the most *related feature* that may empirically reflect the hypothesis, which could guide REVIEWAGENT to identify relevant components from all experimental results,
- and *rebuttal* to the review from previous turns, except for the first iteration.

Thus, the formulation of PROPOSALAGENT could be expressed as:

$$\begin{aligned} \text{Proposal}^{(i)} &= \left\{ \text{Idea \& Method.}^{(i)}, \text{Exp. Settings}^{(i)}, \text{Hypo. \& Related Feat.}^{(i)}, \text{Rebuttal}^{(i)} \right\}, \\ &= \textsc{ProposalAgent}\left(\text{Research Topic} \mid \text{History}^{(i)} \right), 1 \leq i \leq M, \end{aligned} \tag{1}$$

where

$$\text{History}^{(i)} = \begin{cases} \emptyset, & \text{if } i = 1 \\ \left\{ \text{Proposal}^{(j)}, \text{Exp. Res.}^{(j)}, \text{Review}^{(j)} \right\}_{j=1}^{i-1}, & \text{if } 1 < i \leq M \end{cases}, \tag{2}$$

i indicates the number of iteration, N denotes the maximum iteration, $\textsc{ProposalAgent}(\cdot \mid \cdot)$ indicates the agentic workflow, and *experiment result* and *review* are from EXPAGENT and REVIEWAGENT elaborated in Section 3.3.3. The DSL format of the proposed methodology is illustrated in Appendix B. Building upon the aforementioned components, PROPOSALAGENT puts forward a comprehensive yet highly executable proposal, which is then submitted to EXPAGENT for execution. Upon receiving the review form REVIEWAGENT, PROPOSALAGENT can initiate the next iteration, either exploring a brand new direction or optimizing current experimental results.

3.3.3 REVIEWAGENT

Drawing inspiration from human practice, we recognize that significant insights and breakthroughs often emerge from in-depth analysis of experiments and reflection on methodology based on empirical results. To facilitate this process, we design REVIEWAGENT to analyze the experimental results and provide feedback to PROPOSALAGENT, iteratively improving the overall proposal.

In order to conduct a comprehensive and constructive review, REVIEWAGENT performs analysis at different levels of granularity. For fine-grained analysis, REVIEWAGENT examines comprehensive experimental logs, analyzing intermediate results from multi-level metrics which could be predefined by human researchers, e.g. performance indicators of the benchmark, or self-generated in code segment (examples for data engineering shown in Table 1). The *review of the experimental results* identifies hidden patterns in the empirical details, resulting in fruitful low-level feedback mainly on experiment design and adjustment on the expectation of PROPOSALAGENT for the experimental results. For coarse-grained analysis, it evaluates the general validity and reasonableness

of the methodology and hypothesis, providing *review of the* whole *proposal*. This review content serves as high-level advice on the idea and methodology, with the aim of provoking PROPOSALAGENT toward higher creativity. An example of a review of data engineering research is as follows:

An example of the review from REVIEWAGENT

Review of the Experimental Results

Summary and Actionable Insights: Based on the comprehensive analysis of various features influencing the scores of responses in the Alpaca-GPT4 Database, here are the key findings and recommendations for optimizing the dataset...
Key Insights:
1. Length and Word Count: High-quality responses tend to be longer, with word counts above 1000 for answers and around 15-20 words for queries.
2. Conciseness: While length...

Review of the Proposal

Evaluation of Current Research Components:
Your proposal effectively identifies key issues within the Alpaca-GPT4 dataset, such as... Additionally, the need for specific, measurable criteria for evaluating data points to improve...
Suggestions: 1. Data Distribution Analysis: Perform a quantitative analysis to understand the prevalence and distribution of these issues within your dataset...

Formally, the outcome of REVIEWAGENT could be expressed as:

$$\begin{aligned}\text{Review}^{(i)} &= \left\{\text{Review of the Exp. Res.}^{(i)}, \text{Review of the Proposal}^{(i)}\right\}, \\ &= \textsc{ReviewAgent}\left(\text{Research Topic} \mid \text{Proposal}^{(i)}, \text{Exp. Res.}^{(i)}, \text{History}^{(i)}\right), 1 \leq i \leq M,\end{aligned} \tag{3}$$

where $\textsc{ReviewAgent}(\cdot \mid \cdot, \cdot, \cdot)$ indicates the agentic workflow, and *experiment result* contains the benchmark results and other metric values extracted from experiments. In addition, human scientists derive valuable insights not only from a literature review and reasoning, but also through empirical analysis and detailed inspection of the experimental phenomenon, especially for subjects relying largely on empirical studies. Compared to previous work (Lu et al., 2024; Su et al., 2024) that improve ideation creativity primarily based on literature, our system advances this approach by introducing multi-granular review of experimental results and processes. We argue **the groundtruth of scientific laws root and get reflected in experimental outcomes, which could serve as process supervision** in our iterative refinement of the proposal in the pre-falsification phase, and might contribute to the overall creativity of BABY-AIGS. Please refer to Section 3.5 for empirical analysis.

3.3.4 MULTI-SAMPLING STRATEGY

In this section, we formalize the multi-sampling strategy employed in the pre-falsification phase of BABY-AIGS system. This strategy is designed for better efficiency and quality of iterative exploration by parallel executing PROPOSALAGENT, EXPAGENT, REVIEWAGENT, etc. for multiple threads, combined with reranking to retain the most promising threads for further exploration.

As shown in Figure 3, the multi-sampling strategy operates orthogonal to the iterative refinement of the proposal, where the pre-falsification process of each iteration i involves parallel sampling across N threads, and each sampled thread represents a full pre-falsification process, including ideation, experimentation, reviewing, etc. Formally, let $\mathcal{S}^{(i)} = \{s_1^{(i)}, s_2^{(i)}, \ldots, s_N^{(i)}\}, i = 1, ..., M$ represent the set of threads sampled in iteration i. Each sample $s_j^{(i)}, j = 1, ..., N$ undergoes experiments and reranking based on pre-defined criteria, and only a subset with top-ranked samples $\mathcal{S}_{\text{top}}^{(i)} \subset \mathcal{S}^{(i)}$ of size N_s is retained for the next iteration. The process can be summarized as follows:

1. **Sampling Step**: In each iteration i, the system generates N samples $\{s_1^{(i)}, s_2^{(i)}, \ldots, s_N^{(i)}\}$ in parallel. If the former samples $\mathcal{S}_{\text{top}}^{(i-1)}$ are available, i.e., it is not the first iteration, each $s_j^{t+1}, j = 1, ..., N$ is generated by taking into account the historical log from the $\left(j \lfloor \frac{N}{N_s} \rfloor + 1\right)$-th sample of the previous $\mathcal{S}_{\text{top}}^{(i-1)}$ threads.

2. **Reranking**: All samples are reranked on the basis of the benchmarking result during experimentation. For simplicity, we adopt the average performance score of all benchmarks.
3. **Selection for Next Iteration**: After step 2, the samples are reranked and the top N_s samples are selected to form the set $\mathcal{S}_{\text{top}}^{(i)}$ for the next iteration.

Within BABY-AIGS, the multi-sampling strategy with reranking is applied primarily in the **Pre-Falsification** phase, facilitating an extensive yet efficient exploration of ideas, methods, and experimental configurations. By iteratively narrowing down to the top candidates, this strategy effectively focuses resources on promising pathways. In Section 3.6, we empirically demonstrate the multi-sampling strategy, coupled with reranking, is essential for guiding the iterative process in BABY-AIGS towards scientifically significant discoveries in an effective and potentially scalable manner.

3.3.5 FALSIFICATIONAGENT

In the research process, there is usually a gap between the experimental results indicating improvement in performance and the final conclusions of the scientific findings, and human researchers usually perform ablation studies to verify the authenticity of scientific discoveries. We term progress like this ***falsification***, which is a critical step towards full-process automated scientific discoveries.

Recognizing the importance of ***falsification***, we introduce FALSIFICATIONAGENT, a novel component not present in previous work (Lu et al., 2024; Su et al., 2024). FALSIFICATIONAGENT has access to all history records, including proposals from PROPOSALAGENT, experiment results from EXPAGENT, and reviews from REVIEWAGENT. We hypothesize that scientific discoveries are more likely to emerge from significant experimental phenomena, i.e. changes in results, thus, FALSIFICATIONAGENT in BABY-AIGS first performs a "Significance Screening" to identify adjacent turns of pre-falsification phase with greatest performance discrepancies, as shown in Figure 6. Following this, FALSIFICATIONAGENT generates scientific discovery candidates from these selected turns. Then FALSIFICATIONAGENT generates the plans and the ablated methods for ablation experiments. We require that at most T plans are made for each discovery candidate, indicating that at most T ablation experiments will be conducted, and each ablation experiment focuses on the verification of a single factor that may influence the experimental result. Specifically, FALSIFICATIONAGENT must select an iteration as the baseline for the ablation study, and FALSIFICATIONAGENT follows the "Experiment Settings" of the baseline, and modify the methodology according to the ablated factor.

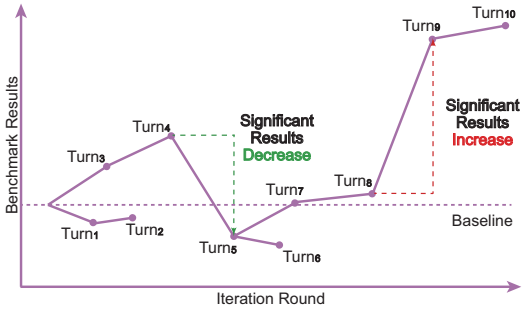

Figure 6: Illustration of "Significance Screening" on history records. The starting point of each turn represents the modifications and experiments based on proposals from that round. The "Significance Screening" process identifies results with significant performance increase or decrease.

Attempting to reach a robust and reliable conclusion of the ablation study, both baseline and ablation experiments are repeated multiple times. FALSIFICATIONAGENT is given the complete record of these experiments to decide the validity of the associated scientific principle. If a particular discovery withstands this process and consistently produces results similar to those in the main experiment, it is regarded as a verified and valuable scientific discovery. And it is falsified otherwise.

Formally, the outcome of FALSIFICATIONAGENT, which is also the output of BABY-AIGS, is:

$$\text{Scientific Discovery} = \text{FALSIFICATIONAGENT}\left(\text{Research Topic} \mid \text{History}\right), \tag{4}$$

where

$$\text{History} = \left\{\text{Proposal}^{(i)}, \text{Exp. Res.}^{(i)}, \text{Review}^{(i)}\right\}_{i=1}^{M}, \tag{5}$$

and $\text{FALSIFICATIONAGENT}(\cdot \mid \cdot)$ indicates the agentic workflow. We also provide an example on the data engineering research to better describe the different parts of the output of FALSIFICATIONAGENT in BABY-AIGS as follows, in which specific parts of the methodology are ablated and reasonable conclusions are made based on the results of the ablation experiment:

An example of the falsification process from FALSIFICATIONAGENT

Dicovery Candidate

Key Factor: Importance of Context and Specificity.

Ablation Experiemnt Plan

Conduct an ablation study by systematically removing or altering one element related to context retention or specificity at a time. For example, test the impact of removing specific instructions or reducing context retention by limiting the number of conversational turns accessible to the model. This will help identify which specific factors within context and specificity contribute most significantly to model performance on MT-bench.

Methodology

Methodology for Ablaton Experiments:{..."Principles": "...7. Responses should be concise and fall within the optimal length range (800-1500 characters).\n8. Responses should engage the user naturally and be informative.\n9. Weighting of each principle should be considered based on its importance to the downstream task.\n10. Incorporate dynamic thresholding to adjust based on the number of data points passing the initial filter."...}

Methodology for Baseline Experiments:{..."Principles": "...7. Responses should be concise and fall within the optimal length range (800-1500 characters).\n8. Responses should demonstrate context retention and follow multi-turn dialogue accurately.\n9. Responses should engage the user naturally and be informative.\n10. Weighting of each principle should be considered based on its importance to the downstream task.\n11. Incorporate dynamic thresholding to adjust based on the number of data points passing the initial filter.\n12. Break down complex criteria into more specific sub-criteria to capture nuances better. For example, 'context retention' can include sub-criteria like 'long-term memory' and 'contextual continuity.'\n13. Evaluate the impact of each criterion through ablation studies."...}

Experiment Result

Metric	Ablation		Baseline	
	Trial 1	Trial 2	Trial 1	Trial 2
Vicuna-Bench (Validation) ↑	7.1625	6.7500	6.475	6.5375
MT-Bench (Test) ↑	4.10625	4.1125	4.05625	3.96875

Verification & Scientific Discovery

Verification: The discovery regarding the importance of context and specificity in multi-turn conversations appears to have a negative influence on the quality of the data for this specific task. The ablation study results suggest that **limiting the number of conversational turns and focusing on fewer criteria can lead to better scores, which implies that the discovery might not positively affect the data quality as initially thought**.

Scientific Discovery: The true scientific discovery here is that while maintaining context and specificity is generally important, overemphasizing these elements may not always lead to improved performance in all tasks. Simplifying the criteria and focusing on essential conversational elements without excessive context retention can lead to better quality results in the context of the MT-bench task.

To our knowledge, FALSIFICATIONAGENT is the first agent within AI-accelerated scientific discovery systems capable of autonomously completing the falsification process, by independently proposing scientific discovery candidates, designing and executing ablation experiments, and performing verification. For a detailed qualitative analysis, see Section 3.5.

3.4 AUTOMATED FULL-PROCESS RESEARCH EXPERIMENT

3.4.1 SELECTED RESEARCH TOPICS

We conduct experiments on three primary research topics in machine learning to evaluate BABY-AIGS in autonomous full-process research. Formally, let $\mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^N$ denote the k-th benchmark of a given ML problem, where x_i represents input features and y_i represents the corresponding labels. The goal is building a system $f : \mathcal{X} \to \mathcal{Y}$ that maximizes metric functions $\mathcal{L}_k(f(x), y)$ over all benchmark $\mathcal{D}_k$. We split benchmarks into validation and test ones, and only the former is available in the pre-falsification phase, avoiding wrong scientific discoveries from over-fit results.

Data Engineering Data engineering is a critical research topic that focuses on the identification, extraction, and processing of relevant data features that significantly influence model performance. We formulate the research goal as follows: Given a data set $\mathcal{H}$ that contains instruction-response pairs, the goal is to identify the key distinguishing characteristics of $\mathcal{H}$, which in turn enables the system to filter and extract high-quality data subsets $\mathcal{H}' \subset \mathcal{H}$ for the development of LLMs. This process is crucial to improving the quality and relevance of data for a wide range of areas, ensuring downstream tasks, such as in-context learning (Brown et al., 2020) and Supervised Fine-Tuning (SFT) for LLM alignment (Ouyang et al., 2022), are more effective. Specifically, we leverage Alpaca-GPT4 dataset (Peng et al., 2023) as the dataset $\mathcal{H}$. We follow previous work (Liu et al., 2024; Chen et al., 2024b; Li et al., 2024c; Zhao et al., 2024) in this field and let the AIGS systems write principles for LLMs to rate data samples and extract the top rated ones as the refined dataset. Thus, for BABY-AIGS, we input the description of the topic and design the main DSL as a list of required principles for the evaluation of the data sample and a threshold indicating the least number of principles that a data sample in the refined dataset has to pass.

Self-Instruct Alignment The self-instruct alignment (Wang et al., 2023c) is a well adopted data synthesis paradigm for LLM alignment. The objective of this research topic is to synthesize a set of SFT data with high quality and diversity for LLM alignment (Ouyang et al., 2022) by rewriting a seed set of data, thereby enhancing the performance of the fine-tuned model on this dataset. In the research process, an AIGS system is required to construct an optimal set of instructions from a seed instruction dataset, which are used to generate an instruction-response dataset from LLMs. This dataset is then leveraged to refine the alignment of an LLM via SFT. In the experiment, we rewrite the original seed instruction set, and use the same LLM in instruction synthesis and response generation for SFT data. Specifically, for BABY-AIGS, the DSL is designed as an option whether to use the seed instruction set, and a list of requirements for the given LLM to generate instructions.

Language Modeling Language modeling is a core research topic in natural language processing that aims to improve the ability of a model to understand and generate human language. Currently, the mainstream approach is generative pre-training (Radford et al., 2018), and the objective is to maximize the perplexity of the next token prediction, i.e. minimize the model perplexity. The AIGS system seeks to explore different architectural and training schedule modifications to enhance quality of language model pre-trained on large corpora. We designed DSL of the BABY-AIGS system as a set of constrained configurations of model architecture and training hyper-parameters.

Each of these research topics requires unique methodological innovations of an AIGS system to foster high ***creativity***, ***executability***, and ***falsification*** capabilities. We demonstrate the pre-defined grammars of BABY-AIGS in Figure 5. Please refer to Appendix B for detailed settings.

3.4.2 EVALUATION SETTINGS

We evaluate BABY-AIGS based on three key principles central to AIGS systems as proposed in Section 3.1: ***falsification***, ***creativity***, and ***executability***. We introduce the AI Scientist (Lu et al., 2024) as the baseline of the automated research system, and also select published literature from top conference as the baseline of research from experienced human researchers.

Falsification We assess BABY-AIGS 's ability to perform falsification through human evaluation, focusing on the falsification process carried out by FALSIFICATIONAGENT. This process involves hypothesizing potential influencing factors, identifying the key variables that may impact experimental results, designing and conducting ablation experiments, and ultimately validating the real factors contributing to the experimental significance. The human evaluation is carried out by volunteer researchers with experience in publishing at top-tier conferences. Evaluators assess the fal-

Metric	AVG	STD	P-Value	MIN	MAX
Importance Score ($0 \sim 2$)					
BABY-AIGS (Ours)	1.80	0.41	0.02	0.00	**2.00**
Top Conference	**2.00**	0.00	—	**2.00**	**2.00**
Consistency Score ($0 \sim 2$)					
BABY-AIGS (Ours)	1.00	0.86	0.00	0.00	**2.00**
Top Conference	**2.00**	0.00	—	**2.00**	**2.00**
Correctness Score ($0 \sim 2$)					
BABY-AIGS (Ours)	0.95	0.94	0.00	0.00	**2.00**
Top Conference	**2.00**	0.00	—	**2.00**	**2.00**
Overall Score ($0 \sim 2$)					
BABY-AIGS (Ours)	1.25	0.47	0.00	0.67	**2.00**
Top Conference	**2.00**	0.00	—	**2.00**	**2.00**

Table 2: Statistic results of human evaluation on the falsification process in our data engineering research experiments.

sification process based on three key dimensions, each scored on a scale from 0 to 2, with a higher score indicating better performance:

- ***Importance Score***: This score reflects the importance of the scientific discovery candidate. It evaluates the extent to which the identified factors can influence the experimental results, considering their relevance and potential impact with the primary experiments.
- ***Consistency Score***: This score assesses whether the proposed ablation experiment plan is aligned with the identified scientific discovery candidate. It considers whether the experiments are designed to ablate the factor of interest and appropriately test the hypothesis.
- ***Correctness Score***: This score evaluates the accuracy of the final scientific discovery derived from the ablation studies. It considers whether the conclusions drawn from the ablation experiments and baseline results are correct, based on the observed empirical results.

Additionally, several studies from the top conferences (Liu et al., 2024; Chen et al., 2024b; Li et al., 2024c; Zhao et al., 2024) are included in the evaluation set to serve as a baseline. We conduct the evaluation on the data engineering research experiment, with statistic results shown in Table 2, where the p-values obtained from a left-tailed hypothesis test against the top conference baseline.

Creativity We measure the creativity of BABY-AIGS by evaluating the performance improvement of the proposed idea and methodology against the baseline result, i.e., the result from the trivial methodology on the test benchmarks. Here are the benchmark settings for each research experiment:

- **Data Engineering**: For the refined dataset, we conduct 15-shot In-Context Learning (ICL) (Jiang et al., 2024) and SFT for LLM alignment to evaluate the overall quality. We evaluate the ICL-aligned LLM on the Vicuna-Bench, as a efficient validation benchmark, and ICL- and the SFT-aligned LLM on the MT-Bench (Zheng et al., 2023), which are used as test benchmarks. The baseline of turn 0 uses the original Alpaca-GPT4 dataset (Peng et al., 2023). We replicate AI Scientist with the same experiment template. Moreover, we replicate Deita (Liu et al., 2024) as the human research of the topic from the top conference.
- **Self-Instruct Alignment**: We also assess the aligned LLM on the Vicuna-Bench, as the validation benchmark, and the MT-Bench, as the test benchmark. The baseline of turn 0 is the result of the original self-instruct method (Wang et al., 2023c).
- **Language Modeling**: We pre-train a mini-sized language model with the modified architecture based on the configured training schedule, on three different training sets (Karpathy, 2015; Hutter, 2006; Mahoney, 2011). The validation and test benchmarks are the perplexity of LM on the split validation and test sets. With reference to Lu et al. (2024), we adopt the default settings of the nanoGPT project[5] as the baseline.

Results on all test benchmarks are in Table 3, Table 4, and Table 5, for each topic, respectively.

[5] `https://github.com/karpathy/nanoGPT`.

Method	MT-Bench ↑	
	15-shot ICL	SFT
Baseline (Turn 0)	4.18	4.53
AI Scientist	4.36	4.67
BABY-AIGS (Ours)	**4.51**	4.77
Top Conference	4.45	**5.01**

Methodology Summarization (Data Engineering)

1. Rate the response based on its contextual coherence, ensuring it logically follows the conversation.
2. Evaluate the relevance by checking if the answer stays on-topic with minimal digression.
3. Check for logical reasoning in explanations, ensuring the response is not just factual but also thoughtful.
4. Consider if the complexity and detail match the question's requirements, avoiding oversimplification.
5. Finally, evaluate the tone for politeness, clarity, and natural conversational flow.

Table 3: Benchmarking results on the test benchmarks of the data engineering research experiment (left) and a summarization of the corresponding proposed methodology from BABY-AIGS (right).

Method	MT-Bench ↑
Baseline (Turn 0)	2.45
BABY-AIGS (Ours)	**3.26**

Methodology Summarization (Self-Instruct Alignment)

Make the instruction to cover different scenarios if it lacks specificity, clearer if ambiguous, aligned with natural conversations, and to contain a diverse range of task types if it lacks variety.

Table 4: Benchmarking results on the test benchmark of the self-instruct alignment research experiment (left) and a summarization of the corresponding proposed methodology from BABY-AIGS (right).

Executability We evaluate the BABY-AIGS system's stability to execute research ideas errorlessly from ideation to implementation, measured by the success rate of obtaining meaningful experimental outcomes and scientific insights, termed as Experiment Success Rate (Exp. SR) and Overall Success Rate (Overall SR), respectively. We report the overall results on all research experiments on the three topics. AI Scientist as the baseline method, are also evaluated executability on the selected tasks in their original implementation (Lu et al., 2024). Results are shown in Table 6.

3.5 QUANTITATIVE AND QUALITATIVE ANALYSIS

BABY-AIGS could produce valid scientific discoveries with falsification process. To validate the falsification process in BABY-AIGS, we assess its ability to perform ablation studies and identify causative factors for experimental results. The qualitative analysis in Table 2 shows that FALSIFICATIONAGENT could produce valid scientific discoveries in current design, as the maximum value of each metric is tied to the top-conference baseline, contributing positively to the automation of scientific insights. However, there are two critical findings that indicate further improvement is needed. (1) The average value of the importance score is higher than the consistency and correctness score, indicating that FALSIFICATIONAGENT could identify important factors potentially related to a scientific discovery but failed to design a concrete experiment plan and verify the hypothesis. The failure could be attribute to the capacity of foundation model or the lack of high-quality demonstration of experiment design in prompts. (2) The p-values indicate that the falsification process of BABY-AIGS is significantly less satisfactory than the existing literature from top conferences from human perspectives, which emphasizes the importance of designing user-friendly interfaces besides refining the design of ablation experiments. Also, we acknowledge that the scale of the study is small compared to Si et al. (2024), which requires future effort.

BABY-AIGS demonstrates creativity during research idea exploration and refinement. Table 3, Table 4, and Table 5 show the results of the test benchmarks for *data engineering*, *self-instruct alignment*, and *language modeling* research experiments, respectively, where BABY-AIGS outperforms the baseline method, demonstrating the system's creativity in ideation and corresponding method design. For data engineering, BABY-AIGS outperforms AI Scientist with a significant margin, demonstrating the effectiveness of the enriched feedback, including multi-granular metrics, verbose review on both experiment process and methodology design, etc., in exploring research idea. However, the result of SFT alignment is inferior than Deita (Liu et al., 2024), indicating that the lack of validation benchmarking of specific downstream tasks might result in an suboptimal outcome.

Method	Perplexity ↓		
	shakespeare_char	enwik8	text8
Baseline (Turn 0)	**1.473**	1.003	0.974
BABY-AIGS (Ours)	1.499	**0.984**	**0.966**

Methodology **Summarization (Language Modeling)**

Reduce the dropout rate with more attention heads to increase model expressiveness. And implement a cyclical learning rate and adjust the weight decay to regularize the model.

Table 5: Benchmarking results on the test benchmarks of the language modeling research experiment (left) and a summarization of the corresponding proposed methodology from BABY-AIGS (right).

Method	**Experiment Success Rate** (Exp. SR)	**Overall Success Rate** (Overall SR)
AI Scientist	44.8%	29.2%
Baby-AIGS (Ours)	**Almost 100%**	**Almost 100%**

Table 6: Success rates on three selected tasks of AI Scientist and Baby-AIGS. Exp. SR denotes the times a system successfully conducted experiments out of all trials, and Overall SR denotes the times a system produces the final scientific discoveries. Higher numbers indicate better executability.

BABY-AIGS has remarkable executability in experimentation and full research process. As shown in Table 6, our quantitative analysis highlights significant improvements in executability, with BABY-AIGS achieving nearly 100% success rates in translating the generated ideas into experimental results and the final scientific discovery. This high executability, attributed to our DSL design for errorless experimentation, prevents restarting from in-process failures and enables an efficient automated research process. Detailed API costs are elaborated in Appendix B.2.

3.6 DISCUSSIONS

Q1: How do current LLMs perform in the falsification process? Falsification (Popper, 1935) is essential in AIGS systems as it provides a rigorous mechanism for verification of potential scientific discoveries, a core component in the scientific method. In BABY-AIGS, FALSIFICATIONAGENT plays the corresponding role. Thus, it demands related abilities in the foundation model, such as reasonable hypothesis generation, ablation experiment design, summarization and self-correction based on input empirical results, etc. As shown in the case in Section 3.3.5 and Table 2, current LLMs are far from desired in the agentic workflow of FALSIFICATIONAGENT. Additionally, the constraints may come from the ability of the LLM to understand the environment outside FALSIFICATIONAGENT. For instance, from our observation, FALSIFICATIONAGENT seldom proposes experiment plans beyond the provided experiment templates. In this case, although DSL makes sure the executability of the experimentation by omitting extra operations, the experiment process would differ from the original plan, thus creating inconsistency.

Method	Baseline	Turn 1	Turn 2	Turn 3	Turn 4	Turn 5
Multi-Sampling@1	4.18	3.68	4.01	4.05	3.88	3.90
Multi-Sampling@32	4.18	4.02	4.05	4.50	4.51	4.42

Table 7: Results on MT-Bench (15-shot ICL) of the ablation study on the multi-sampling strategy of our BABY-AIGS system in the data engineering research experiment. N in "Multi-Sampling@N" indicates the number of parallel threads of multi-sampling.

Q2: How does the BABY-AIGS system boost creativity? BABY-AIGS enhances creativity by integrating a multi-sampling approach combined with re-ranking, allowing it to generate diverse research proposals and rank them based on validation benchmarks. We provide detailed results of an ablation study of this process in Table 7. We observed that the performance on the test benchmark is steadily increasing with multi-sampling with large numbers of threads. This strategy is related to search-based inference-cost scaling methods (Snell et al., 2024; Brown et al., 2024). The insight is to pick random high-performing samples for better overall performance. However, since the objective of AIGS is to discover science on a research topic, the reranking method here could be large-scale

validation benchmarks indicating generalization performance, rather than reward-model-based (Stiennon et al., 2020) or self-verification methods for a specific query. As depicted in Section 3.3.3, we argue that the groundtruth of scientific laws is rooted and reflected in benchmarking results from actual experiments, which could serve as process supervision, which could be more accurate than reward models. It explains how collapse in self-refinement-style methods (Xu et al., 2024) is avoided in this setting, which is also empirically validated through the ablation results.

Q3: Why could DSL help with executability? The use of a Domain-Specific Language (DSL) in BABY-AIGS facilitates executability by providing a structured and executable representation of ideas and methodologies proposed by PROPOSALAGENT. DSL enhances the system's ability to translate complex scientific workflows into actionable experiment plans. As shown in Table 6, DSL significantly improved success rates in generating scientific discoveries, regardless of correctness, underscoring its role in achieving high executability. We acknowledge that the design of DSL requires human effort and might not be able to cover all possible method implementations. However, we believe it is a promising interface between agents and experimentation in full-process research.

4 LIMITATIONS AND ACTIONABLE INSIGHTS

Envisioning the future of AI-Generated Science systems powered by foundation models in real-world, in this section, we enumerate a few limitations for current BABY-AIGS system and provide insights on the next steps of research for **AIGS**.

Balance idea diversity and system executability. As discussed in Section 3.3.1, the design of the DSL enhances the system executability but may constrain the idea diversity. Achieving a balance between idea diversity and system executability requires further empirical analysis. One potential avenue is enabling agents to develop their own DSLs, which could enhance the executability of generated ideas without diminishing their diverse potential.

Establish systematic mechanisms for evaluation and feedback. The quality of AIGS system depends heavily on rigorous evaluation of prior proposals, methods, and results. Current approaches often adopt a peer review format, leveraging LLMs to generate feedback on results and guide future optimization (Lu et al., 2024; Yu et al., 2024; Jin et al., 2024). However, it remains unclear whether this method is the most effective for large-scale research settings. Future work should explore systematic mechanisms to analyze outcomes across iterations, maximizing experience transfer and continuous improvement.

Strengthen the falsification procedure. Our research underscores the importance of falsification to enhance the scientific rigor of the research findings. While we have prototyped the falsification process in our BABY-AIGS system, more efforts are required to strengthen the modules related to knowledge falsification, including the exploitation of the patterns and relationships derived from historical experiments for the guidance of refined research proposals. Besides, it is also vital for AIGS systems to investigate whether the delivered new scientific knowledge could generalize across diverse research domains in an autonomous manner.

Expand channels for scientific knowledge dissemination. Facilitating the exchange of AI-Generated Science is critical, both between humans and AI and among AI systems. While Lu et al. (2024) focus on disseminating knowledge through research papers, alternative formats like posters, podcasts, and videos are gaining traction with the rise of multi-modal agents. Future research should also explore more efficient communication channels between AI systems, beyond structured text or natural language (Pham et al., 2024; Chen et al., 2024c).

Exploring communication dynamics among autonomous AI researchers. As discussed in Section 2, the advancement of AI-accelerated scientific discovery spans four paradigms, culminating in the emergence of an autonomous AI research community (Paradigm IV). Within this community, individual agentic researchers engage in interactions that parallel collaborative dynamics found in human scientific networks. Analyzing these communication dynamics is essential to understand how fully-autonomous AI agents might effectively collaborate, exchange knowledge, and drive collective progress. In particular, a deeper exploration of these interactions in a multi-agent system will help establish communication frameworks that support optimal collaboration, fostering a robust and productive AI-accelerated research community.

Promote interdisciplinary knowledge integration and experimentation. In this work, we primarily focused on the application of AIGS systems within the domain of machine learning, where experiments could be executed in computers. However, future developments should extend these systems to address challenges in other scientific fields, such as biology, which has been preliminarily explored in a concurrent work (Swanson et al., 2024), chemistry, and physics, where cross-disciplinary knowledge integration is often crucial. One major challenge lies in how AI agents can synthesize and align domain-specific knowledge from multiple fields, which often have distinct terminologies, methodologies, and epistemological assumptions. Another critical challenge is the experiment environment, which could be hardly automated and might be highly resource-consuming. We hope the integrity and development of optional modules like DOMAIN-SPECIFIC EXPAGENT and ENVIRONMENTAGENT mentioned in Section 3.2 could alleviate the challenges, and further effort is needed and will be made in future work.

5 ETHICS AND IMPACT STATEMENT

In our BABY-AIGS system, the agent did not perform harmful operations on computer systems or environment because of the design of DSL, task constraints and no access to external tools. However, while the system developed in this study is limited in scope, AIGS systems as a whole may have significant impacts in the future, with potential risks that should not be overlooked. This section explores the potential negative impacts of such systems, drawing on prior research, and offers suggestions for promoting their positive development.

5.1 POTENTIAL NEGATIVE IMPACTS OF AIGS SYSTEMS

Impact on Human Researchers and Academic Community In the absence of robust publication standards and academic review processes, AIGS systems could flood the academic community with low-quality literature, which will further increase researchers' workload and disrupt the efficient dissemination of knowledge (Lu et al., 2024; Si et al., 2024; Hu et al., 2024b). And although Si et al. (2024) and Kumar et al. (2024) suggest that LLMs can generate ideas more creative than humans, the extent of such creativity remains uncertain. LLM-powered AIGS systems tend to rely heavily on existing data and patterns, which could foster *path dependency* and limit opportunities for groundbreaking discoveries. Additionally, these systems might inadvertently use proprietary or copyrighted material, raising concerns about intellectual property infringement (Kumar et al., 2024). Furthermore, AIGS systems also present several unpredictable challenges for human researchers:

- **Dependence Effect and Cognitive Inertia**: Over-reliance on AI-generated insights may diminish researchers' independent thinking, leading to cognitive stagnation and a decline in critical thinking skills (Si et al., 2024; Hu et al., 2024b).
- **Ambiguity in Responsibility Attribution**: The involvement of AI complicates the assignment of credit and responsibility, potentially disrupting existing incentive structure (Si et al., 2024; Hu et al., 2024b).
- **Weakened Collaboration and Increased Isolation**: As AIGS systems become capable of independently generating publishable work, researchers may increasingly rely on these systems, reducing the need for direct collaboration and communication with colleagues. This shift could lead to a decline in interpersonal interaction, weakening traditional research networks built on teamwork and shared discourse (Si et al., 2024; Hu et al., 2024b). Over time, the diminishing frequency of collaborative exchanges may foster a sense of professional isolation among human researchers, heightening the risk of loneliness, disengagement, and reduced psychological well-being.
- **Exacerbated Technological Barriers**: Without equitable access to advanced AIGS systems, a technological divide could emerge, disadvantaging researchers unfamiliar with or lacking access to these systems, thereby exacerbating inequalities within the community.

Impact on Environment AIGS systems can conduct large-scale experiments in parallel, but their dependence on iterative processes carries the risk of inefficient feedback loops, potentially leading to issues such as infinite loops. This inefficiency, caused by limited reasoning capabilities, the misuse of erroneous information, or ambiguity in task definition, could drive up energy consumption. Moreover, poorly regulated experiments, especially without adequate simulation environments, can lead to unintended environmental harm. For example, untested chemical processes in materials science may yield hazardous by-products, while unchecked experiments in nuclear research could increase the risk of radiation leaks (Tang et al., 2024).

Impact on Social Security AIGS systems, particularly when compromised by jailbreak attacks, could generate responses that conflict with human values, such as providing instructions for creating explosives. This raises concerns about their misuse for harmful purposes, such as designing more advanced adversarial attack strategies (Tang et al., 2024; Si et al., 2024; Lu et al., 2024; Kumar et al., 2024; Hu et al., 2024b). Even with benign intentions, unsupervised scientific research may introduce unforeseen societal risks. For instance, monopolizing breakthroughs in autonomous AI could lead to severe unemployment, market monopolies, and social unrest (Tang et al., 2024).

5.2 Strategies for Responsible and Ethical Development of Automated Research Systems

Strengthening the Security of Foundation Models The most fundamental step in mitigating security risks associated with AIGS systems is enhancing the security of their foundation models. Incorporating instructions for handling unsafe research into the alignment training corpus, alongside conducting rigorous safety audits prior to model deployment, are both crucial strategies to ensure the systems be robust and secure (Tang et al., 2024).

Aligning Scientific Agents with Human Intentions, Environment and Self-constraints Scientific agents in AIGS systems should align with human intentions, the environments in which they operate, and self-constraints (Yang et al., 2024c).

- **Human Intentions**: Agents must accurately interpret user intent, going beyond literal language to capture the deeper purpose of scientific inquiries.
- **Environment**: Agents need to adapt to the specific environments in which they function by applying domain-specific knowledge accurately and utilizing specialized tools effectively.
- **Self-Constraints**: Agents must evaluate task feasibility, manage resources wisely, and minimize waste to ensure sustainable operation. This includes setting boundaries to prevent redundant work or harmful behavior, which is essential for maintaining system efficiency.

Providing Comprehensive Training for Human Users Comprehensive and rigorous training is essential for users to fully leverage AIGS systems and prevent unintended consequences (Aidan, 2024). Proper training minimizes the risk of misuse that could lead to environmental harm, resource waste, or unethical research outcomes. Training programs should focus not only on technical skills but also on ethical considerations, ensuring users understand the limitations and responsibilities associated with these systems (Tang et al., 2024).

Building a Collaborative Framework Between Automated Research Systems and Human Researchers To prevent AIGS systems from exerting excessive influence on the academic community, collaboration between AIGS systems and human researchers will play a crucial role (Si et al., 2024; Hu et al., 2024b). It is essential to explore the new roles and responsibilities that human scientists may need to assume in this evolving research landscape shaped by the presence of AIGS systems. A well-structured partnership can leverage the complementary strengths of both, enabling outcomes that neither could achieve independently. Moreover, such collaboration fosters interaction among human researchers, encouraging deeper communication and mitigating the sense of isolation that may arise from increased reliance on automated tools.

Establishing Comprehensive Legal and Accountability Frameworks A robust legal and accountability framework is crucial to govern the use of AIGS systems. This framework should:

- **Define Clear Scientific Research Boundaries**: Specify the permissible scope and limitations of these systems, where regulate agents with the DSL might be helpful.
- **Clarify Responsibility and Credit Allocation**: Establish guidelines for assigning credit and responsibility for research outcomes generated with the assistance of AIGS systems (Si et al., 2024; Hu et al., 2024b).
- **Implement Penalties for Misuse**: Outline liability measures and penalties to address harmful behavior or unethical practices involving these systems.

Using AIGS Systems to Address Its Own Challenges AIGS systems can also play a proactive role in addressing the challenges and even ethical issues introduced by themselves. For example, AIGS systems could be used to monitor and evaluate outputs from other automated systems,

identifying potential ethical issues, biases, or environmental risks before they escalate. Moreover, AIGS systems can facilitate the development of guidelines, by automating the analysis of research trends and regulatory needs, thus helping shape future policies for responsible AI use. When employed strategically, AIGS systems become not only tools for discovery but also mechanisms for self-regulation, creating a virtuous cycle of innovation and governance.

6 CONCLUSION

We introduce the concept of **AIGS** in this paper and implement BABY-AIGS, a baby-step toward full-process automated scientific discovery systems, with a focus on incorporating ***falsification*** into the research process. By integrating a FALSIFICATIONAGENT, the multi-agent system can identify and verify potential discoveries. Techniques as DSL and multi-sampling strategy are introduced for two other principles of AIGS systems design, ***executability*** and ***creativity***. Preliminary experiments show promise, though the system's performance remains below that of experienced human researchers. This work lays the groundwork for future developments in AIGS systems, with further improvements over BABY-AIGS and ethical considerations necessary for advancing the field.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, pp. 1–3, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *ArXiv preprint*, abs/2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

Toner-Rodgers Aidan. Artificial Intelligence, Scientific Discovery, and Product Innovation. *preprint*, 2024. URL https://aidantr.github.io/files/AI_innovation.pdf.

Alberto Alfarano, François Charton, and Amaury Hayat. Global Lyapunov functions: a long-standing open problem in mathematics, with symbolic transformers. *ArXiv preprint*, abs/2410.08304, 2024. URL https://arxiv.org/abs/2410.08304.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/P04-3031.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. URL https://arxiv.org/abs/2407.21787.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv preprint*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. MLE-bench: Evaluating machine learning agents on machine learning engineering. *ArXiv preprint*, abs/2410.07095, 2024. URL https://arxiv.org/abs/2410.07095.

Kexin Chen, Junyou Li, Kunyi Wang, Yuyang Du, Jiahui Yu, Jiamin Lu, Lanqing Li, Jiezhong Qiu, Jianzhang Pan, Yi Huang, Qun Fang, Pheng Ann Heng, and Guangyong Chen. Chemist-X: Large language model-empowered agent for reaction condition recommendation in chemical synthesis, 2024a. URL https://arxiv.org/abs/2311.10776.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a Better Alpaca Model with Fewer Data. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=FdVXgSJhvz.

Weize Chen, Chenfei Yuan, Jiarui Yuan, Yusheng Su, Chen Qian, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Beyond natural language: LLMs leveraging alternative formats for enhanced reasoning and communication. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10626–10641, Miami, Florida, USA, November 2024c. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.623.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024d.

Jiale Cheng, Yida Lu, Xiaotao Gu, Pei Ke, Xiao Liu, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. AutoDetect: Towards a unified framework for automated weakness detection in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6786–6803, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.397.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-Agent review generation for scientific papers. *ArXiv preprint*, abs/2401.04259, 2024. URL https://arxiv.org/abs/2401.04259.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip Yu, and Wenpeng Yin. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5081–5099, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.292.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24, 2024.

Mingmeng Geng and Roberto Trotta. Is ChatGPT transforming academics' writing style? *ArXiv preprint*, abs/2404.08627, 2024. URL https://arxiv.org/abs/2404.08627.

Daniel George and EA Huerta. Deep neural networks to enable real-time multimessenger astrophysics. *Physical Review D*, 97(4):044039, 2018.

Alireza Ghafarollahi and Markus J Buehler. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 2024.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. Towards possibilities & impossibilities of AI-generated text detection: A survey. *ArXiv preprint*, abs/2310.15264, 2023. URL https://arxiv.org/abs/2310.15264.

Yolanda Gil, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, and Jim Myers. Examining the challenges of scientific workflows. *Computer*, 40(12):24–32, 2007.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017. URL http://proceedings.mlr.press/v70/gilmer17a.html.

Karan Girotra, Lennart Meincke, Christian Terwiesch, and Karl T Ulrich. Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN 4526071*, 2023.

Michael Haman and Milan Školník. Using ChatGPT to conduct a literature review. *Accountability in research*, 31(8):1244–1246, 2024.

Wenpin Hou and Zhicheng Ji. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods*, pp. 1–4, 2024.

Chao-Chun Hsu, Erin Bransom, Jenna Sparks, Bailey Kuehl, Chenhao Tan, David Wadden, Lucy Wang, and Aakanksha Naik. CHIME: LLM-assisted hierarchical organization of scientific studies for literature review support. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 118–132, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.8. URL https://aclanthology.org/2024.findings-acl.8.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. *ArXiv preprint*, abs/2408.08435, 2024a. URL https://arxiv.org/abs/2408.08435.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. Nova: An iterative planning and search approach to enhance novelty and diversity of LLM generated ideas. *ArXiv preprint*, abs/2410.14255, 2024b. URL https://arxiv.org/abs/2410.14255.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and Peace (WarAgent): Large language model-based multi-agent simulation of world wars. *ArXiv preprint*, abs/2311.17227, 2023. URL https://arxiv.org/abs/2311.17227.

Jingshan Huang and Ming Tan. The role of ChatGPT in scientific communication: writing better scientific review articles. *American journal of cancer research*, 13(4):1148, 2023.

Marcus Hutter. The hutter prize, 2006. URL http://prize.hutter1.net.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DISCOVERYWORLD: A virtual environment for developing and evaluating automated scientific discovery agents. *arXiv preprint arXiv:2406.06769*, 2024.

Yixing Jiang, Jeremy Andrew Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y. Ng. Many-shot in-context learning in multimodal foundation models. In *ICML 2024 Workshop on In-Context Learning*, 2024. URL https://openreview.net/forum?id=j2rKwWXdcz.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. AgentReview: Exploring peer review dynamics with LLM agents. *ArXiv preprint*, abs/2406.12708, 2024. URL https://arxiv.org/abs/2406.12708.

Yongfei Juan, Yongbing Dai, Yang Yang, and Jiao Zhang. Accelerating materials discovery using machine learning. *Journal of Materials Science & Technology*, 79:178–190, 2021.

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

Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks, 2015. URL https://karpathy.github.io/2015/05/21/rnn-effectiveness/.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. Can large language models unlock novel scientific research ideas? *ArXiv preprint*, abs/2409.06185, 2024. URL https://arxiv.org/abs/2409.06185.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. Agent Hospital: A simulacrum of hospital with evolvable medical agents. *ArXiv preprint*, abs/2405.02957, 2024a. URL https://arxiv.org/abs/2405.02957.

Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. MLR-Copilot: Autonomous machine learning research based on large language models agent. *arXiv preprint arXiv:2408.14033*, 2024b.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. One-shot learning as instruction data prospector for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4586–4601, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.252. URL https://aclanthology.org/2024.acl-long.252.

Yutong Li, Lu Chen, Aiwei Liu, Kai Yu, and Lijie Wen. ChatCite: LLM agent with human workflow guidance for comparative literature summary. *ArXiv preprint*, abs/2403.02574, 2024d. URL https://arxiv.org/abs/2403.02574.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *ArXiv preprint*, abs/2403.07183, 2024a. URL https://arxiv.org/abs/2403.07183.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. Mapping the increasing use of LLMs in scientific papers. *ArXiv preprint*, abs/2404.01268, 2024b. URL https://arxiv.org/abs/2404.01268.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BTKAeLqLMw.

Yuliang Liu, Xiangru Tang, Zefan Cai, Junjie Lu, Yichi Zhang, Yanjun Shao, Zexuan Deng, Helan Hu, Zengxian Yang, Kaikai An, et al. ML-Bench: Evaluating large language models and agents for machine learning tasks on repository-level code. *ArXiv preprint*, abs/2311.09835, 2023. URL https://arxiv.org/abs/2311.09835.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, Hongchao Fang, Zhuoyang Zou, Wenchao Ma, Xi Li, Kai Zhang, Congying Xia, Lifu Huang, and Wenpeng Yin. AAAR-1.0: Assessing AI's potential to assist research. *ArXiv preprint*, abs/2410.22394, 2024. URL https://arxiv.org/abs/2410.22394.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *ArXiv preprint*, abs/2408.06292, 2024. URL https://arxiv.org/abs/2408.06292.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pp. 1–11, 2024.

Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. LLM and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. In *International Conference on Machine Learning*. PMLR, 2024.

Matt Mahoney. About the test data, 2011. URL http://mattmahoney.net/dc/textdata.html.

Benjamin S Manning, Kehang Zhu, and John J Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. LLM critics help catch LLM bugs. *ArXiv preprint*, abs/2407.00215, 2024. URL https://arxiv.org/abs/2407.00215.

Marjan Mernik, Jan Heering, and Anthony M. Sloane. When and how to develop domain-specific languages. *ACM Comput. Surv.*, 37(4):316–344, 2005. ISSN 0360-0300. doi: 10.1145/1118890.1118892. URL https://doi.org/10.1145/1118890.1118892.

Harshit Nigam, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. Acceleron: A tool to accelerate research ideation. *ArXiv preprint*, abs/2403.04382, 2024a. URL https://arxiv.org/abs/2403.04382.

Harshit Nigam, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. An interactive Co-Pilot for accelerated research ideation. In Su Lin Blodgett, Amanda Cercas Curry, Sunipa Dey, Michael Madaio, Ani Nenkova, Diyi Yang, and Ziang Xiao (eds.), *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pp. 60–73, Mexico City, Mexico, 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.hcinlp-1.6.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. ART: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint*, abs/2303.09014, 2023. URL https://arxiv.org/abs/2303.09014.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–18, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with GPT-4. *ArXiv preprint*, abs/2304.03277, 2023. URL https://arxiv.org/abs/2304.03277.

Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A. Plummer, Zhaoran Wang, and Hongxia Yang. Let models speak ciphers: Multiagent debate through embeddings. In *International Conference on Learning Representations (ICLR)*, 2024.

Karl R. Popper. *The Logic of Scientific Discovery*. Routledge, London, England, 1935.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. Large language models are zero shot hypothesis proposers. *ArXiv preprint*, abs/2311.05965, 2023. URL https://arxiv.org/abs/2311.05965.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations (ICLR)*, 2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468. PMLR, 2020. URL http://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Abheesht Sharma, Gunjan Chhablani, Harshit Pandey, and Rajaswa Patil. DRIFT: A toolkit for diachronic analysis of scientific literature. In Heike Adel and Shuming Shi (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 361–371. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-DEMO.40. URL https://doi.org/10.18653/v1/2021.emnlp-demo.40.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. LLM-SR: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*, 2024.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? *ArXiv preprint*, abs/2409.04109, 2024. URL https://arxiv.org/abs/2409.04109.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, abs/2408.03314, 2024. URL https://arxiv.org/abs/2408.03314.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. *ArXiv preprint*, abs/2410.09403, 2024. URL https://arxiv.org/abs/2410.09403.

Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P Dow. ReviewFlow: Intelligent scaffolding to support academic peer reviewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 120–137, 2024.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, 2024. doi: 10.1101/2024.11.11.623004. URL https://www.biorxiv.org/content/early/2024/11/12/2024.11.11.623004.

Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, et al. Prioritizing safeguarding over autonomy: Risks of LLM agents for science. *ArXiv preprint*, abs/2402.04247, 2024. URL https://arxiv.org/abs/2402.04247.

Wei Tao, Yucheng Zhou, Wenqiang Zhang, and Yu Cheng. MAGIS: LLM-based multi-agent framework for GitHub issue resolution. *ArXiv preprint*, abs/2403.17927, 2024. URL https://arxiv.org/abs/2403.17927.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, et al. AI-driven review systems: Evaluating LLMs in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023a.

Rui Wang, Hongsong Feng, and Guo-Wei Wei. ChatGPT in drug discovery: A case study on anticocaine addiction drug development with Chatbots. *Journal of chemical information and modeling*, 63(22):7189–7209, 2023b.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. Hypothesis search: Inductive reasoning with language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=G7UtIGQmjm.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Yanjun Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, Junyang Lin, Robert Brennan, Hao Peng, Heng Ji, and Graham Neubig. OpenHands: An open platform for AI software developers as generalist agents, 2024b. URL https://arxiv.org/abs/2407.16741.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15474–15492, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.826. URL https://aclanthology.org/2024.acl-long.826.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*, 2023. URL https://arxiv.org/abs/2309.04658.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-Computer interfaces enable automated software engineering, 2024a.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. A survey on detection of LLMs-generated content. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9786–9805, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.572.

Zonghan Yang, An Liu, Zijun Liu, Kaiming Liu, Fangzhou Xiong, Yile Wang, Zeyuan Yang, Qingyuan Hu, Xinrui Chen, Zhenhe Zhang, Fuwen Luo, Zhicheng Guo, Peng Li, and Yang Liu. Position: Towards unified alignment between agents, humans, and environment. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024c. URL https://openreview.net/forum?id=DzLna0cFL1.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, RenJing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10164–10184, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.595.

Weihao Zeng, Can Xu, Yingxiu Zhao, Jian-Guang Lou, and Weizhu Chen. Automatic instruction evolving for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6998–7018, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.397.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. AFlow: Automating agentic workflow generation. *ArXiv preprint*, abs/2410.10762, 2024a. URL https://arxiv.org/abs/2410.10762.

Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. {REMARK-LLM}: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1813–1830, 2024b.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60674–60703. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhao24b.html.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. Protecting language generation models via invisible watermarking. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42187–42199. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhao23i.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. *ArXiv preprint*, abs/2404.04326, 2024. URL https://arxiv.org/abs/2404.04326.

A Implementation Details of the Baby-AIGS system

In this section, we elaborate the implementation details of the Baby-AIGS system. All artifacts are used as intended with their license strictly followed in our work.

A.1 RESEARCH-AGNOSTIC IMPLEMENTATION

System Pipeline We posit that all agents mentioned in Section 3.2 contribute to a full-process AIGS system, but based on preliminary experiments, we simplify the design of EXPAGENT and LITERATUREAGENT to a large extent in our implementation. For EXPAGENT, given the design of DSL with human effort, proposed methodology generated by PROPOSALAGENT can be executed reliably in experiments, which is also shown in Section 6. This reduces the need of iteratively refining proposals between PROPOSALAGENT and EXPAGENT. For LITERATUREAGENT, preliminary results show literature integration did not significantly impact the outcomes in both phases of BABY-AIGS. We conclude the reason as that agents failed to understand the in-depth literature information and the retrieval of literature did not match the need of each agent perfectly. Therefore, in our implementation, we minimize the design of these two agents: EXPAGENT functions through fixed code, and LITERATUREAGENT was not put into pratical use. Other optional agents are designed to function in broader research fields, and we chose to omit them in experiments based on the selected research topics for experiments (Section 3.4).

Hyper-Parameters Experiments in ICL (In Context Learning) of the data engineering research and in language modeling research are conducted on 8 NVIDIA GeForce RTX 3090 24 GB GPUs. Experiments in SFT (Supervised Fine-tuning) of the data engineering research and in Self-Instruct alignment research are conducted on 8 A100 80GB GPUs. All researches utilize the gpt-4o-2024-05-13 model as the underlying model for our agents. When agents invoke GPT-4o, we use the openai module[6] with a temperature setting of 0.7, while all other parameters are setting as default values. During the synthesis of proposals, PROPOSALAGENT generates three sets of proposals with a temperature of 0.7. After generation, the Jaccard similarity (Jaccard, 1901) of bigram sets is calculated between the methodology of each proposal and the methodology produced in the previous iteration. The proposal with the lowest similarity in methodology is selected as the final output to increase its diversity. For REVIEWAGENT and FALSIFICATIONAGENT, they invoke the GPT-4o only once each time when generating responses.

A.2 RESEARCH-SPECIFIC IMPLEMENTATION

Data Engineering In this research experiment, our system is tasked with exploring different approaches to improve the quality of Alpaca-GPT4 dataset (Peng et al., 2023). The DSL configuration and instance are shown in Figure 5 and Figure 7. The Llama-3-8B-Instruct[7] model is employed to rate all data samples with the principles in DSL. We deploy Llama-3-8B-Instruct using vLLM[8], configuring the temperature to 0.05, while keeping all other parameters at the default settings. We use Llama-3-8B[9] for ICL- and SFT-alignment, and the model and the fine-tuned checkpoints are deployed using vLLM with a maximum token limit of 1024, while other parameters follow the default configurations provided by FastChat[10]. In falsification process, the BABY-AIGS system identifies the factors that contribute to quality improvements and conclude whether there are ways to stably improve the quality of the extracted dataset, thus delivering valuable scientific discoveries. For significance screening in FALSIFICATIONAGENT, iterations are identified as having significant improvements if the difference of adjacent benchmarking results exceeds 1.5 for the ICL-aligned Llama-3-8B on the Vicuna-Bench (the validation benchmark) or 0.5 on the MT-Bench (the test benchmark). From these iterations, candidates for scientific discovery are extracted. For hyper-parameters, we set the total iteration number $M = 5$ and set the multi-sample threads number $N = 32$.

Self-Instruct Alignment In this research experiment, our system is tasked with exploring different approaches to improve the quality of synthesized SFT data from a seed dataset in Self-Instruct[11] (Wang et al., 2023c). We use GPT-4o to rewrite the seed data for better quality with the temperature parameter set to 0.05. The DSL configuration and instance are shown in Figure 5 and Figure 8. We use the Llama-3-8B[12] model to generate instructions and responses, with it also serving as the base model for SFT alignment. We use LoRA (Hu et al., 2022) method from LLaMA-

[6] https://github.com/openai/openai-python

[7] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[8] https://github.com/vllm-project/vllm.

[9] https://huggingface.co/meta-llama/Meta-Llama-3-8B.

[10] https://github.com/lm-sys/FastChat.

[11] https://github.com/yizhongw/self-instruct.

[12] https://huggingface.co/meta-llama/Meta-Llama-3-8B.

Factory[13] to fine-tune the model with default training hyper-parameters[14]. The other experiment setting is the same as data engineering research. For hyper-parameters, we set the total iteration number $M = 15$ and set the multi-sample threads number $N = 1$ due to limited computing resources for parallel model training.

Language Modeling In this research, the system is tasked to pre-train a mini-sized language model on several small corpora, aiming to improve performance by minimizing loss on the selected datasets. The experiment mainly follows the same setup as the language modeling task in AI Scientist (Lu et al., 2024), based on the nanoGPT project [15]. The DSL configuration and instance are shown in Figure 5 and Figure 9, where we guide the models in adjusting parameters related to model architecture and training process. For the experiments, we use the sampling scripts provided in the template code without modifications. For hyper-parameters, we set the total iteration number $M = 10$ and set the multi-sample threads number $N = 1$ due to limited computing resources for parallel model training.

B Experiment Details

B.1 Guidelines for Human Evaluators

To thoroughly assess the quality of our falsification process, we conducted a human evaluation of 20 agent-generated falsification logs. The guidelines are summarized as follows:

- ***Importance Score***: Assess the significance of the proposed scientific discovery candidate, considering its potential impact on experimental results and its relevance and consistency with the main experiments.
- ***Consistency Score***: Evaluate whether the proposed ablation experiments align with the scientific discovery candidate and whether the experiment appropriately isolates the factor in question.
- ***Correctness Score***: Determine whether the final scientific discovery drawn from the falsification process is correct based on the ablation and baseline results.

For each dimension, the evaluator assigns an integer score ranging from 0 to 2, where a higher score indicates better performance. The overall statistic results are shown in Table 2.

B.2 API Costs of the Full-Process Research Experiment

In our experiments, we measured the average token counts and costs of different phases of Baby-AIGS (Section 3.2) for invoking the GPT-4o API and the results are presented in Table 8. Note that as the experimental records in past iterations are used as input in most requests, with the rise of iteration, the length of record will consequently increase, leading to the use of more tokens.

	Input Tokens	**Generated Tokens**	**Cost** ($)
Pre-Falsification (per iter.)	6,616.2	761.5	0.045
Falsification (per disc. cand.)	43,375.5	1,120.3	0.234

Table 8: Average token consumption and API costs for GPT-4o API in the full-process research experiment. The costs at pre-falsification phase is calculated for each iteration, and the costs at falsification phase is calculated for each discovery candidate.

B.3 DSL Demonstrations for Different Research Topics

We present an example of the methodology in DSL format generated during the experiment for each research topic, as shown in Figure 7, Figure 8 and Figure 9, corresponding to data engineering, self-instruct alignment, and language modeling, respectively.

[13] https://github.com/hiyouga/LLaMA-Factory.

[14] https://github.com/hiyouga/LLaMA-Factory/blob/main/examples/train_lora/llama3_lora_sft.yaml.

[15] https://github.com/karpathy/nanoGPT.

Data Engineering

LLM Generated

{
"Paradigm": "Data Sample Rating & Filtering",
"Principles": "Relevance: The data point should be relevant to the task of multi-turn conversation and instruction-following.\n Completeness: The conversation should be complete and coherent, without abrupt endings or missing context.\nInstruction adherence: The responses should strictly follow the given instructions and provide accurate information.\nConversational flow: The dialogue should flow naturally, with each turn logically following the previous one.\nComplexity handling: The data point should demonstrate the ability to handle complex and challenging questions effectively.\nEngagement: The conversation should be engaging and maintain the user's interest throughout.\nEach data point should involve multi-turn dialogue.\nResponses should be contextually relevant to the preceding turns.\nData points should cover a diverse range of topics to mimic common use cases.\nResponses should be coherent and grammatically correct.\nEach response should show logical progression and consistency across turns.\nRelevance to the prompt: The response should directly address the question or task presented in the prompt.\nCoherence: The response should be logically structured and easy to follow.\nGrammar and syntax: The response should be free of grammatical and syntactic errors.\nCreativity and depth: The response should demonstrate creative thinking and provide in-depth information when required.\nConsistency: The response should maintain consistency in its argument or narrative throughout.\nLength: Ensure responses are comprehensive, aiming for lengths similar to high-scoring entries (1000 to 3000 characters).\nWord Count: Encourage comprehensive and thorough responses, ensuring the content is relevant and informative.\nUnique Words: Ensure responses contain a broad range of unique words while maintaining relevance and coherence.\nStopwords Count: Ensure responses are detailed and contextually rich.\nKeyword Overlap: Ensure responses are relevant and contextually appropriate.\nDiversity: Aim for answer diversity in the range of 0.396 to 0.690.\nAverage Word Length: Encourage balanced word lengths between queries and answers.\nSentiment: Train models to deliver engaging, relevant, and positive responses.\nCoherence Score: Refine the scoring method to better capture logical progression and consistency.\nInstruction Adherence: Ensure responses have high instruction adherence.\nComplexity Score: Prioritize generating detailed and complex answers. \nEngagement Score: Ensure responses are engaging and interactive.",
"Number": 27,
"Threshold": 15,
"Ratio": 0.7
}

Figure 7: The DSL instance for data engineering research.

Self-Instruct Alignment

LLM-Generated

{
"Paradigm": "Instruction Data Synthesis",
"Prompt": "1. Ensure queries are between 50-150 characters and answers are between 300-1500 characters. Aim for clear and concise queries (10-26 words) and detailed yet concise answers (55-254 words).\n2. Balance specificity to provide clear and relevant information without being overly detailed (Query specificity: 1, Answer specificity: 2-4). Ensure specific terms are contextually relevant.\n 3. Maintain moderate complexity in language to ensure clarity and conciseness (Query clarity score: 2-5, Answer clarity score: 3-7). Avoid jargon unless necessary.\n4. Increase relevance by incorporating task-specific keywords and ensuring both queries and answers are contextually relevant and detailed. Ensure answers directly address the queries.\n5. Diversify the seed data to cover a broad range of tasks, topics, and scenarios, including more complex instructions. Include tasks of varying complexity and from different domains (e.g., healthcare, finance, education).\n6. Use an LLM to perform the initial evaluation and rewrite. Have human reviewers refine the rewritten instructions.\n7. Implement a structured feedback mechanism to continuously refine the principles and methodology.\n 8. Analyze high-scoring tasks and responses on VicunaBench and MT-bench to tailor the principles.",
"Seed": true
}

Figure 8: The DSL instance for self-instruct alignment research.

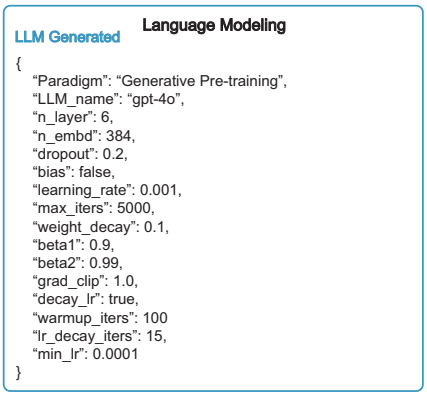

Figure 9: The DSL instance for language modeling research.

C PROMPTING STRUCTURE

In this section, we will briefly introduce the prompting structures of the PROPOSALAGENT, REVIEWAGENT, and FALSIFICATIONAGENT as shown in Figure 10, Figure 11, and Figure 12, respectively. For detailed prompts, please refer to our code repository[16] .

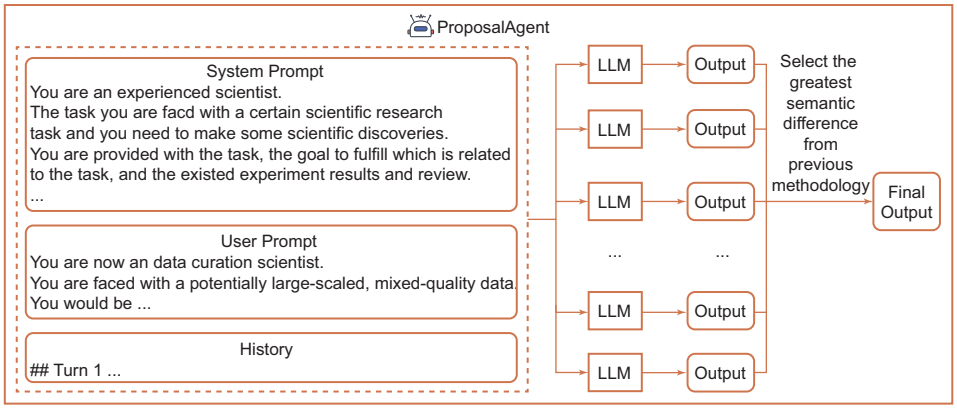

Figure 10: The prompting structure for the PROPOSALAGENT includes a general system prompt, a research-topic-specific user prompt and history logs. The LLM generates multiple outputs, covering elements such as idea, methodology, DSL, etc. From these outputs, the one whose methodology has the greatest semantic difference from the previous round's methodology is selected as the idea for the current round, aiming to boost creativity in ideation.

[16] https://github.com/AgentForceTeamOfficial/Baby-AIGS.

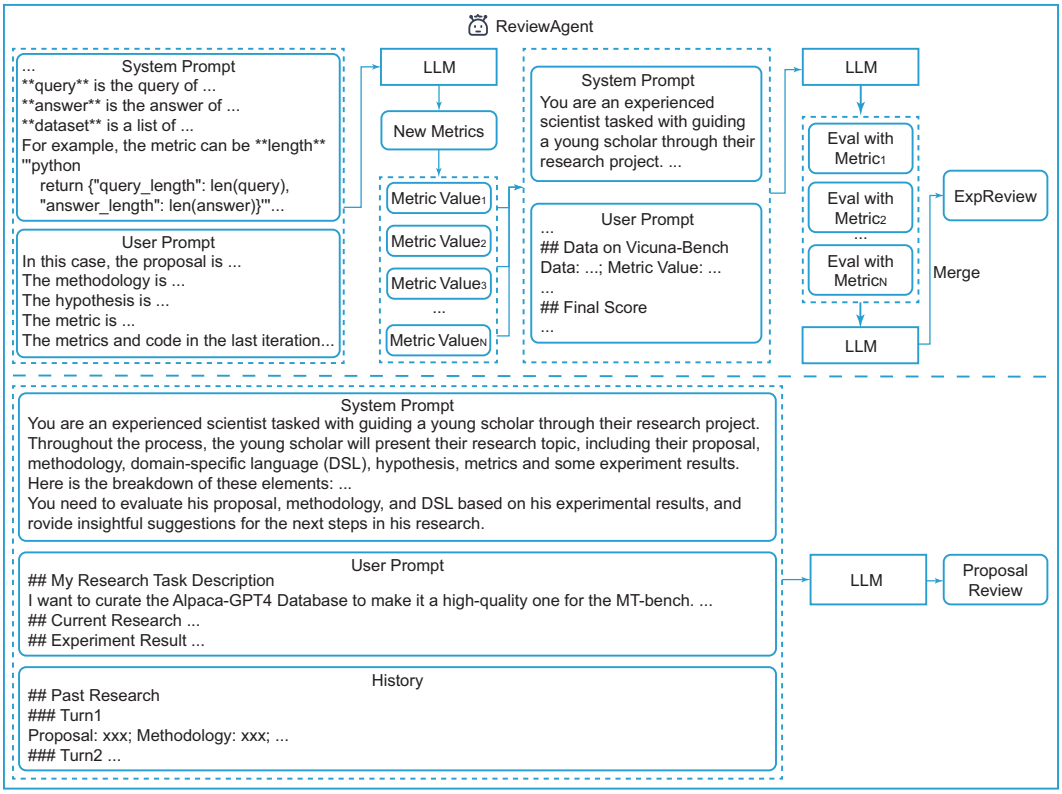

Figure 11: The REVIEWAGENT will first generate new metrics and then analyze each metric individually using the LLM. Following this, the REVIEWAGENT will call the LLM to merge the analysis results for each metric, resulting in the *ExpReview*. Next, the REVIEWAGENT will assess the experimental results by integrating insights from previous ideas and experiments, yielding the *ProposalReview*.

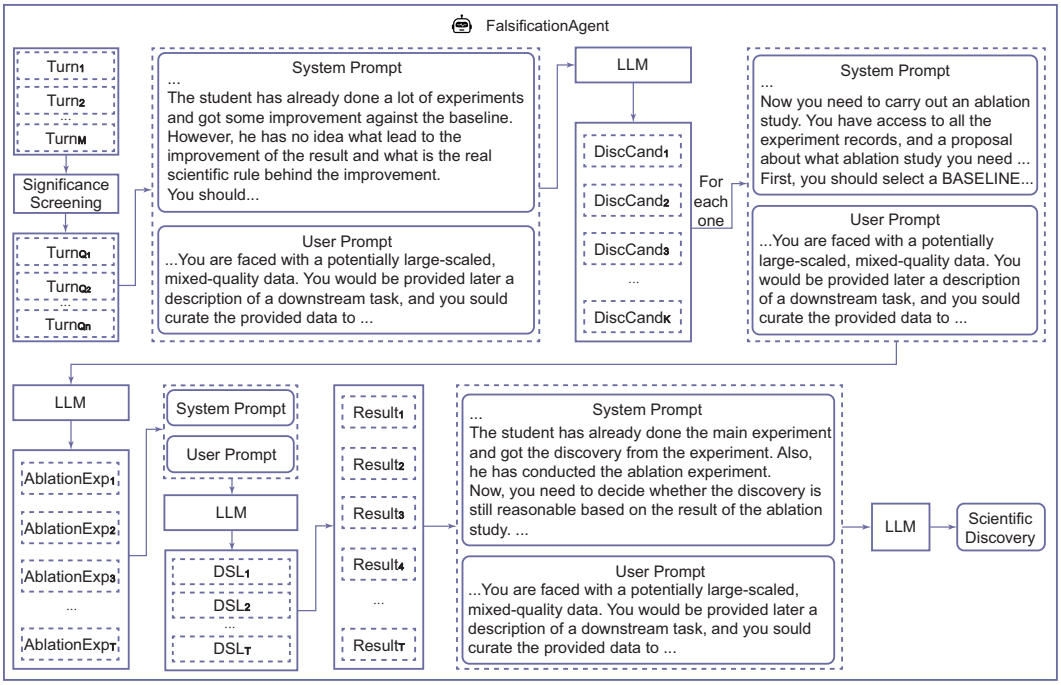

Figure 12: The FALSIFICATIONAGENT first screens all history turns to identify turns with notable changes in results. It then generates discovery candidates from the results obtained through significance screening. For each discovery candidate, it then creates several ablation experiment setups and generates the corresponding DSL to obtain experimental results. Once the experimental results are obtained, the FALSIFICATIONAGENT calls on the LLM to produce the final scientific discovery.