

AI for Scientific Discovery: A Comprehensive Review of AI-Driven Reasoning, Autonomous Agents, and Evolutionary Search

Shinde Bhumika Haridas, Prof.Bhosale S.B, Prof.Bangar A.P, Prof.Khatri A.A

Department of Computer Engineering
Jaihind College of Engineering SPPU,Pune,India

Abstract — Scientific research usually begins with a question, followed by a review of existing work, the development of possible The reasons for this, the planning of the experiments or simulations, and a careful verification of the results. Now artificial intelligence (AI) is being used can be used to help with a number of these steps. Large language models (LLMs) assist researchers in working with scientific information, AI agents can Tasks can be carried out using external tools, and evolutionary methods can continuously generate and test programs. The paper looks at four studies. that look at different parts of AI-assisted discovery: the wider influence of AI on science, autonomous agent systems, the TeLLAgent and the use of large language models in the search for evolutionary programming. Together, these studies demonstrate how AI could be of assistance with literature reviews and idea generation. They also highlight important issues, such as incorrect model outputs, and involve generation, computational testing, and program improvement. data that is biased or incomplete, the limited ability to generalize, poor reproducibility, reliance on tools, and data leakage that occurs during evaluation. The The paper examines the four studies and suggests a possible workflow for using AI responsibly in scientific research. Overall, AI can be It is a useful research assistant, but its suggestions have to be checked by way of transparent analysis and independent validation.

Keywords— Artificial Intelligence, Scientific Discovery, Large Language Models, Autonomous AI Agents, Evolutionary Search, Scientific Research

I. INTRODUCTION

Scientific discovery involves much more than making a A prediction; in order for researchers to understand what is already known, identify unanswered questions, develop possible explanations, select suitable methods, analyze the results, and Determine if the evidence supports their conclusions. Computing tools have helped scientists with many of these tasks, but the growing volume of research publications and the size of It is now very difficult to carry out each step by hand due to the presence of modern datasets. AI is therefore being explored as a way to support literature analysis, prediction, program generation, and research workflows [1], [2]. The four studies reviewed here approach this change from Various angles; Bianchini et al. look at how the use of AI relates Regarding scientific output in various research fields [1], Ferrag et al. carry out a review. the development of autonomous AI agents, including their The architectures, tools, benchmarks, and reliability issues [2]. Sun et We introduce TeLLAgent, a system that combines language By using a two-agent design [3], Mao focuses on using LLMs with evolutionary methods to search for better The papers on programs and algorithms [4]; when taken together, these papers offer a broad picture of both the opportunities and the limits of AI assisted discovery. This

review considers four main questions: where AI can help during scientific research, how agents and external tools extend the abilities of language models, how evolutionary search can support the discovery of programs and algorithms, and what safeguards are needed to make the resulting work reliable and repeatable.

This is a focused narrative review of the four research papers given for analysis; each paper was then assessed in accordance with its main contribution: large-scale evidence about AI in science [1], a survey of autonomous AI agents [2], the TeLLAgent dual-agent system [3], and also evolutionary program search based on LLMs [4]. The papers were compared in terms of their methods, applications, the benefits and the limitations that have been reported. Since the review relates only to these four sources, it should not be read as a systematic review of the whole field of AI for science; the discussion that follows, however, is limited to what these studies report..

II. REVIEW METHODOLOGY

It is a focused narrative review drawn from the four research papers provided. I read the papers and They are grouped according to their main contribution: field-level evidence on AI in science [1], a general survey of autonomous AI agents [2],

an implemented dual-agent scientific discovery system [3], and LLM-driven evolutionary program search [4]. A comparative synthesis was carried out with regard to the methods, the scope of application, the strengths, and the reported limitations. Since the review only includes these four sources, it does not constitute a systematic review of the entire AI-for-science literature; the claims and examples listed below are limited to the evidence given in those papers.

III. AI ACROSS THE SCIENTIFIC PROCESS

AI is capable of assisting with a number of linked phases in research. When it comes to the knowledge stage language models can aid in understanding questions and in obtaining or summarising information. It is necessary to verify the statements that they generate against sources. At the hypothesis stage AI can suggest explanations, molecular structures, algorithms or experimental approaches. At the analysis stage predictive models and scientific software can assess the candidates by means of data, simulations or specialised tools for the field. Lastly automated workflows can compare the results with the constraints. Modify the plan if a candidate fails. Bianchini and others examined than 80 million publications in 172 different scientific fields from 2005 to 2023 and found that the use of AI is linked to scientific novelty and citation impact although there are variations between fields [1]. This large-scale analysis is significant since it goes beyond considering the accuracy of models. The impact of AI depends on the fields in which and the way in which researchers use it. The associations reported should not be seen as evidence that AI by itself leads to science. Other factors such as the characteristics of the field the available research resources and the patterns of adoption can also influence the results.

IV. AUTONOMOUS AI AGENTS FOR DISCOVERY

By incorporating a language model with the ability to interact with tools, data sources, code environments or other agents an LLM-based agent is obtained. A normal workflow consists of interpreting a goal forming a plan choosing the tools carrying out the actions checking the results and altering the plan when needed. The survey by Ferrag et al. Looks at the development from LLM reasoning to agents covering areas such as agent architectures, evaluation benchmarks, tool protocols, applications and the reliability issues that remain unresolved [2].

Tool use is so important in work since a language model on its own might produce explanations or calculations that are

plausible but wrong. A scientific agent can hand over computations to software ask a structured database run a simulation or make use of a domain-specific predictor. Although such tools can make the outputs more reliable they do not ensure correctness by themselves. The tool might be inappropriate the data it provides might be out of date. The agent might misinterpret the result. Reliable systems therefore require actions, explicit checks and human review.

An important issue in terms of architecture is whether one model should plan and carry out all the steps or whether the responsibilities should be divided. Since complex tasks usually involve both high-level decisions and accurate use of tools separating these functions can make it easier to manage and assess the workflows. However it might increase cost and coordination overhead [2] [3].

V. TELLAGENT: DUAL-AGENT SCIENTIFIC DISCOVERY

Sun and others have introduced TeLLAgent a framework consisting of a supervisor and an executor which separates reasoning from the carrying out of low-level tools [3]. The supervisor is responsible for formulating and updating the plan whereas the executor carries out specific operations by means of a curated toolkit. This separation is intended to reduce the failures that occur when a single agent has to alternate between long-range planning and making tool calls.

Furthermore the framework makes use of feedback and iterative correction to deal with steps. The paper shows how the framework can be applied in fields such as the development of organic solar-cell materials and in the area of molecular discovery. In the case of the materials example an acceptor that had been designed by AI was actually. Attained a power conversion efficiency of 16.44% [3]. This serves as an instance of a candidate suggested by computational methods being verified by experiment.

The paper also looks at drug-discovery processes using screening and molecular evaluation tools. While computational confirmation is helpful merely predicting binding or drug-like properties does not prove efficacy. Independent experimental tests are still required.

The TeLLAgent identifies three design principles for AI-assisted science: it should separate planning from execution in the case of tasks link the models reasoning to specialized tools and employ iterative evaluation rather than settling on the first answer produced. Its limitations consist of dependence on the

underlying LLM and toolkit overhead and difficulty in dealing with targets that lie outside the range of distributions supported by the tools or the models [3].

These limitations illustrate the reason why autonomous execution should not be mistaken, for scientific authority.

VI. LLM-DRIVEN EVOLUTIONARY PROGRAM SEARCH

Mao examines the combination of large language models with evolutionary algorithms (EAs) for the purpose of program search and for automated scientific discovery [4]. In this approach, the large language model suggests or alters possible programs, the evaluator assesses their performance, and an evolutionary process then selects the candidates which are to be improved further. The use of memory-based semantic mutation, hierarchical evaluation feedback, and population-management strategies enables a more systematic exploration of the search space than is possible with a single generation attempt. The paper in question looks at various systems and application fields, such as combinatorial optimisation, the discovery of mathematical algorithms, and software engineering, with a view to algorithm and GPU-kernel optimisation [4]. Its method combines the generative flexibility of LLMs together with the objective feedback provided by executable tests or performance measures; rather than just asking the model to explain a solution, the process involves testing out candidate implementations and retaining those which satisfy the specified objective. Yet if the evaluator is incomplete evolutionary search may end up optimising the wrong objective, and a high score does not mean that the programme is necessarily novel, interpretable, safe or generally correct. Mao identifies credibility and hallucination, large search spaces, black-box interpretability, and data leakage as important challenges and points to formal verification and more interpretable search histories [4]. It follows that strong evaluation criteria and reproducible test environments are essential.

VII. COMPARATIVE ANALYSIS

Table I compares the four papers that have been reviewed with regard to their focus, main contribution, and key limitation. The contributions of these papers are complementary rather than directly interchangeable since one of them examines field-level patterns, another surveys agent systems, a third evaluates a particular dual-agent framework and the fourth concentrates on evolutionary program search.

Table 1. Comparison of the Four Reviewed Papers

Paper	Focus, contribution, and key limitation
Bianchini et al. [1]	Focus: AI across science fields. Contribution: study of many papers that looks at how new ideas appear and how often papers are cited. Caution: I find that just because two things happen together does not mean one causes the other; the impact can differ from one area to another.
Ferrag et al. [2]	Focus: agents. Contribution: review of how large language models reason, design, tools, tests and trustworthiness. Caution: I find that how safe autonomous agents are in life and how to test autonomous agents properly remain unsolved questions.
Sun et al. [3]	Focus: TeLLAgent. Contribution: system, with a supervisor agent and an executor agent, plus tools and a loop that fixes mistakes; also shows a demo of finding materials. Caution: I find that the quality of the tools and models is important; computer checks cannot replace lab experiments.
Mao [4]	Focus: program search. Contribution: repeated creation, testing and choosing of programs. Caution: I find that the amount of computing required, the ease of understanding the results the checking of the work and accidental data leaks remain worries.

Across the four papers, a common theme is that useful AI discovery depends on more than fluent generation. AI needs appropriate data, suitable domain tools, measurable evaluation, and a process for rejecting weak candidates. Agent frameworks emphasize tool orchestration and planning [2], [3], while evolutionary search emphasizes repeated generation and objective selection [4]. Field-level analysis adds that scientific benefits are not uniform across disciplines [1]. Together, these findings favour workflows that record their decisions and preserve human responsibility for interpretation.

VIII. PROPOSED CONCEPTUAL FRAMEWORK

Based on the four papers Figure 1 shows a workflow for AI-assisted scientific discovery. This is a summary that is

presented in this review. It is not a system that was put into practice or tested through experiments by the four papers together. The process starts with a research question and information from studies. An AI model or an AI agent then comes up with a plan. Uses tools, databases, code or simulations from the field to create possible results. These possible results are checked against standards. If they don't pass the results are used to improve the planning or search. Good results are looked at by experts, in the field. If needed, tested in real experiments. Only findings that have been proven should be shared with details so others can do the samework again.

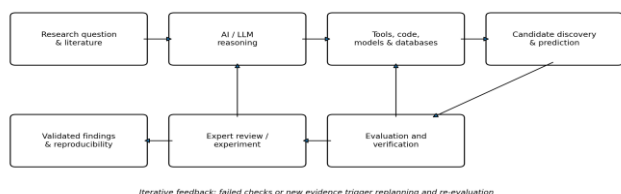


Fig. 1. Proposed conceptual workflow for AI-assisted scientific discovery (synthesized from [1]–[4]; not experimentally validated as a combined system).

Challenges and Future Directions

Hallucination. Scientific systems need to tell the difference between things that sound right and things that are actually proven. Future systems should connect claims to the data they come from show what tools did and check calculations or predictions on their own [2] [3] [4].

Reproducibility and evaluation. Agent workflows should keep track of prompts, model versions, tool versions, settings what happened along the way and what didn't work. Program-search systems also need test sets that're safe from data leaks and evaluation methods that match the real scientific goal instead of just a small part of it [4].

Data quality. Results from one area one dataset or one computer setup might not work elsewhere. The big differences found by Bianchini et al. Show that AIs help depends on the situation [1]. Experts in the field should check if the data covers what it needs if there are biases and if the results still make sense outside the data used to train the AI.

Human. Experimental validation. AI can help find ideas and test them faster but a result, from AI should not be called a real scientific discovery unless it is checked properly. Experimental science needs measurements and control groups; computational science needs code that can be run again standards and checks

done by others. Future studies should compare AI tools with ways and share both the failures and the successes.

Efficiency and interpretability. Multi-agent systems and search methods that evolve over time can use a lot of computing power and tool uses. Research should talk about the costs how long things took, how often they. What kinds of errors happened. It should also look into whether simpler ways can do the job. Clear search records and formal checks can help scientists see why a certain idea was picked [3] [4].

IX. CONCLUSION

This review examined four ways to use AI in science. It looked at how AI connects to results how to build AI agents a dual-agent tool using a collaborative approach and how large language models can drive evolutionary program search in science. The studies show that AI helps scientists organize knowledge generate candidate ideas run multi-step workflows and fine-tune programs or algorithms in science. The useful systems link AI generation to domain tools and measurable feedback instead of relying only on unverified text. Still AI still faces challenges: AI can hallucinate AI does not always generalise well AI may leak data AI can be hard to understand AI costs a lot of compute. Ai needs independent checks. A reliable approach should mix exploration with evaluation reproducible records, domain expert oversight and real experimental or computational verification in science. AI works best as a speed booster and partner in discovery not as a replacement, for evidence.

REFERENCES

1. S. Bianchini, V. Di Girolamo, J. Ravet and D. Arranz "AI in science: When and where it makes a difference " Research Policy, vol. 55 Art. No. 105478 2026 Doi: 10.1016/j.respol.2026.105478.
2. M. A. Ferrag, N. Tihanyi and M. Debbah "From LLM Reasoning to AI Agents: A Comprehensive Review," IEEE Access, vol. 14, Pp. 84237–84278, 2026 Doi: 10.1109/ACCESS.2026.3698694.
3. J. Sun et al. "TeLLAgent: a dual-agent framework for scientific discovery with tool-enhanced LLMs " Chemical Science, vol. 17 Pp. 13101–13117, 2026 Doi: 10.1039/d5sc09883a.
4. J. Mao, "LLM-Driven Evolutionary Program Search: From FunSearch to Automated Scientific Discovery " Proceedings of the International Conference, on Computing and Data Science, 2026 doi: 10.54254/2755-2721/2026.34758.