



A MULTI-AGENT FRAMEWORK FOR HYPOTHESIS GENERATION, EXPERIMENT DESIGN, AND VALIDATION OF AGENTIC AI FOR AUTONOMOUS SCIENTIFIC RESEARCH

Dr. C. Mohamed Faheem, G. Ranganathan

Assistant Professor & Research Supervisor The Research Centre of English
Fatima College (Autonomous) Mary Land, Madurai – 625018
Affiliated to Madurai Kamaraj University

ABSTRACT

New possibilities for automating significant parts of the scientific research lifecycle have emerged as a result of the rapid development of large language models (LLMs), retrieval-augmented generation (RAG), autonomous agents, and tool-using artificial intelligence. In most cases, independent literature reviews, the formulation of hypotheses, the design of experiments, the analysis of data, and the validation of findings are necessary components of traditional scientific research. Even though individual stages of this process have been supported by AI more and more, most of the systems that are currently in use are still task-specific and require a lot of human intervention. A multi-agent framework for autonomous scientific research is proposed in this paper. In this framework, specialized artificial intelligence agents collaborate to perform data analysis, critical evaluation, validation, hypothesis generation, experimental design, literature discovery, and knowledge retrieval. Retrieval-augmented generation, external scientific databases, computational tools, experiment simulators, and verification agents are all part of the proposed framework. The use of independent critics and validation agents to reduce hallucinations, unsubstantiated hypotheses, methodological errors, and reproducibility issues is a key component of the framework. Hypotheses are generated from evidence, experiments are designed to test those hypotheses, results are analyzed, and conclusions are independently verified before being accepted in the closed-loop research process that is followed by the proposed architecture. An evaluation framework based on scientific validity, novelty, factuality, reproducibility, experimental efficacy, citation accuracy, and human expert agreement is also presented in the paper. Developing autonomous scientific research systems while maintaining human oversight at crucial decision points is made possible by the framework's conceptual foundation. The study argues that agentic AI should support evidence-based scientific discovery by acting as an auditable collaborative research ecosystem rather than simply replacing individual research tasks.

KEYWORDS: Agentic AI, Autonomous Scientific Research, Large Language Models, Multi-Agent Systems, Hypothesis Generation, Experiment Design, Retrieval-Augmented Generation, Scientific Discovery, AI Agents, Research Automation

INTRODUCTION

Scientific discovery is traditionally a human-driven process involving observation, literature review, hypothesis development, experimental design, data collection, analysis, interpretation, and validation. Although digital technologies have significantly improved scientific workflows, researchers continue to face challenges associated with the rapidly increasing volume of scientific literature, complex datasets, multidisciplinary knowledge, and limited research time. The growth of scientific publications has made comprehensive literature analysis increasingly difficult for individual researchers.

Artificial intelligence has progressively transformed scientific research by supporting activities such as literature retrieval, molecular prediction, image analysis, simulation, statistical analysis, and scientific writing. Recent developments in large language models have expanded these capabilities because LLMs can process natural-language information, summarize complex documents, generate hypotheses, write computational code, and interact with external tools.



However, conventional AI-assisted research systems are generally designed to perform one specific task. For example, a literature-mining system may identify relevant publications, while a machine-learning model may predict an experimental outcome. Such systems do not necessarily understand the broader research objective or coordinate multiple stages of the scientific method.

Agentic artificial intelligence provides a potential solution to this limitation. Unlike conventional generative AI systems that primarily respond to individual prompts, agentic systems can plan tasks, use tools, maintain intermediate information, evaluate outputs, and perform iterative actions toward a defined objective. A multi-agent architecture can further divide scientific research into specialized functions performed by different agents.

For example, a **Literature Agent** can search scientific databases, a **Knowledge Agent** can organize evidence, a **Hypothesis Agent** can formulate testable hypotheses, an **Experiment Agent** can design experiments, a **Data Analysis Agent** can analyze results, and a **Critic Agent** can independently challenge the proposed findings.

This paper proposes a conceptual **Multi-Agent Autonomous Scientific Research Framework (MASRF)** designed to coordinate these specialized agents through a controlled research pipeline. The central research question is:

How can a multi-agent agentic AI architecture autonomously support hypothesis generation, experiment design, analysis, and scientific validation while maintaining factuality, reproducibility, transparency, and human oversight?

BACKGROUND AND RELATED LITERATURE

2.1 Large Language Models in Scientific Research

Large language models have demonstrated strong capabilities in natural-language understanding and generation. Their ability to process scientific literature makes them potentially useful for literature synthesis, scientific communication, coding assistance, and knowledge discovery.

Nevertheless, LLMs can generate scientifically plausible but factually incorrect information. Such hallucinations represent a significant barrier to using standalone LLMs for scientific discovery. Consequently, autonomous research systems require mechanisms for grounding generated information in authoritative evidence.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation combines language-model generation with information retrieval. Instead of relying exclusively on the model's internal parameters, a RAG system retrieves relevant information from external sources and incorporates that evidence into the generation process.

For scientific research, retrieval sources may include:

- PubMed
- Crossref
- Semantic Scholar
- arXiv
- institutional repositories
- scientific datasets
- domain-specific databases



RAG can improve factual grounding and enable agents to cite evidence supporting their reasoning. However, retrieval quality, source reliability, document relevance, and citation accuracy remain important challenges.

2.3 Multi-Agent Artificial Intelligence

Multi-agent systems divide complex objectives among multiple autonomous entities. Each agent can have a specialized role, knowledge source, toolset, or reasoning objective.

For scientific research, specialization is particularly valuable because scientific discovery involves heterogeneous activities. Literature analysis, statistical modeling, experiment design, simulation, and critical evaluation require different capabilities.

A multi-agent approach can therefore transform scientific research from a single AI interaction into a coordinated computational workflow.

2.4 Autonomous Scientific Discovery

Recent research has begun exploring AI systems capable of performing multiple stages of scientific investigation. These systems indicate a transition from **AI-assisted research** toward **AI-augmented and partially autonomous research**.

However, full autonomy introduces challenges related to scientific validity, hallucination, reproducibility, ethical responsibility, data provenance, and experimental safety. Therefore, autonomy must be accompanied by verification mechanisms.

RESEARCH GAP

Existing AI research systems provide valuable assistance but several gaps remain.

First, many systems focus on individual research tasks rather than the complete research lifecycle. Second, LLM-based systems may generate unsupported hypotheses or inaccurate scientific explanations. Third, autonomous agents may propagate errors when multiple agents depend on the same incorrect information. Fourth, there is limited standardization for evaluating whether AI-generated hypotheses are genuinely novel, experimentally testable, and scientifically meaningful.

Another important gap concerns **verification**. A system that generates a hypothesis and then evaluates its own hypothesis may suffer from confirmation bias. Independent critic agents can provide a second layer of evaluation.

Finally, scientific research requires reproducibility. An autonomous research system should record the sources, prompts, datasets, tools, code, parameters, intermediate decisions, and experimental outputs used during the research process.

This study addresses these gaps by proposing a multi-agent architecture incorporating:

1. Evidence-grounded knowledge retrieval.
2. Specialized scientific agents.
3. Independent hypothesis criticism.
4. Automated experiment design.
5. Computational experimentation.
6. Result verification.
7. Provenance and audit logging.
8. Human-in-the-loop approval.



RESEARCH OBJECTIVES

The primary objective of this study is to develop a conceptual framework for autonomous scientific research using agentic AI.

The specific objectives are:

1. To develop a multi-agent architecture for scientific research automation.
2. To integrate LLMs with retrieval-augmented generation.
3. To automate evidence-based hypothesis generation.
4. To develop an AI-based experimental design mechanism.
5. To introduce independent critic and validation agents.
6. To improve scientific reproducibility through provenance tracking.
7. To define measurable criteria for evaluating autonomous scientific research.
8. To establish appropriate human oversight mechanisms.

PROPOSED MULTI-AGENT FRAMEWORK

The proposed framework is called the **Multi-Agent Scientific Research Framework (MASRF)**.

The architecture consists of nine primary agents:

5.1 Research Planner Agent

The Research Planner Agent receives the research objective and converts it into a sequence of research tasks.

Its responsibilities include:

- defining the research question;
- identifying required evidence;
- decomposing the research problem;
- assigning tasks to specialized agents;
- coordinating the research workflow.

5.2 Literature Discovery Agent

The Literature Discovery Agent searches scientific databases and identifies relevant publications.

The agent evaluates:

- title relevance;
- abstract relevance;
- publication date;
- citation information;
- research methodology;
- study population;
- research findings.

The agent produces a structured evidence collection rather than simply generating a textual summary.

5.3 Knowledge Synthesis Agent

The Knowledge Synthesis Agent integrates information retrieved from multiple sources.

It identifies:

- established findings;
- contradictory findings;
- methodological limitations;
- unexplored relationships;
- emerging trends;
- research gaps.

The output becomes the evidence base for hypothesis generation.



5.4 Hypothesis Generation Agent

The Hypothesis Generation Agent produces scientifically testable hypotheses from the identified evidence gaps.

A hypothesis should satisfy four major conditions:

Evidence-supported + Novel + Testable + Falsifiable

For example:

H1: Integrating multimodal scientific evidence through a retrieval-grounded multi-agent architecture improves hypothesis-generation accuracy compared with a standalone LLM.

The system can generate multiple candidate hypotheses rather than accepting the first generated hypothesis.

5.5 Hypothesis Critic Agent

The Critic Agent independently evaluates candidate hypotheses.

Evaluation criteria include:

Criterion	Description
Novelty	Whether the hypothesis provides a potentially new research direction
Evidence	Whether existing literature supports the hypothesis
Testability	Whether it can be empirically tested
Falsifiability	Whether evidence could disprove it
Feasibility	Whether the required experiment is practical
Scientific significance	Potential contribution to the field

Hypotheses failing minimum criteria are returned to the generation stage.

5.6 Experiment Design Agent

The Experiment Design Agent converts the selected hypothesis into an experimental protocol.

It determines:

- variables;
- control groups;
- independent variables;
- dependent variables;
- sample requirements;
- datasets;
- evaluation metrics;
- statistical methods;
- experimental conditions.

5.7 Data Analysis Agent

After data become available, the Data Analysis Agent performs statistical or computational analysis.

Depending on the research domain, this may include:

- descriptive statistics;
- hypothesis testing;
- regression;
- classification;
- clustering;
- deep learning;
- survival analysis;
- simulation;
- sensitivity analysis.



5.8 Scientific Critic Agent

The Scientific Critic Agent independently reviews the experimental results.

It checks whether:

- conclusions are supported by results;
- statistical assumptions are satisfied;
- alternative explanations exist;
- the analysis contains errors;
- limitations are properly reported.

5.9 Validation Agent

The Validation Agent performs the final verification.

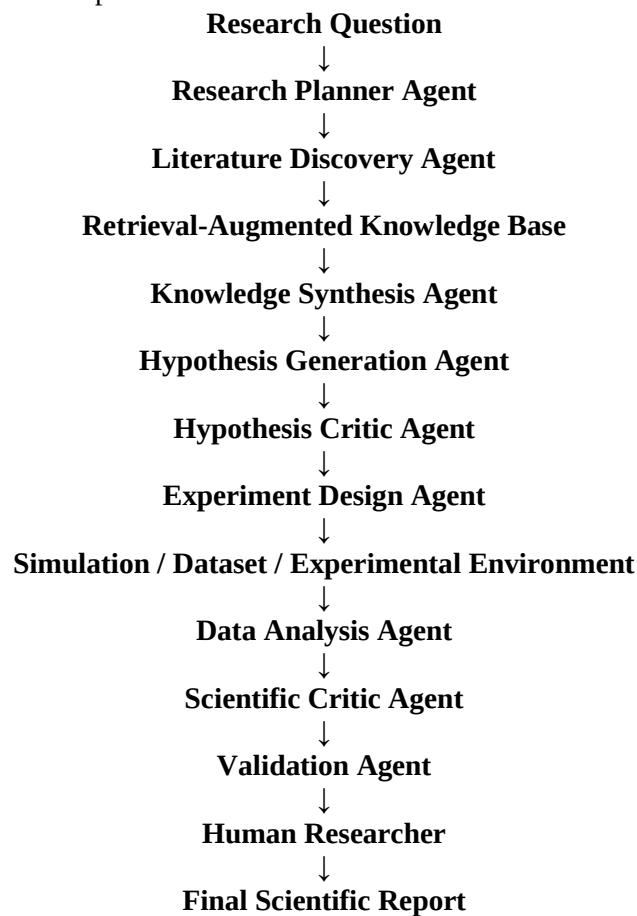
It checks:

- source citations;
- data provenance;
- computational reproducibility;
- methodological consistency;
- statistical validity;
- factual consistency.

Only after validation should the system produce the final scientific report.

SYSTEM ARCHITECTURE

The proposed workflow can be represented as:





The human researcher remains involved at critical decision points, particularly where physical experiments, ethical approval, sensitive data, or high-risk decisions are involved.

METHODOLOGY

7.1 Research Design

This study proposes a design-science research methodology. The objective is to design and evaluate an artificial intelligence framework capable of coordinating multiple scientific research activities.

The proposed evaluation compares three system configurations:

Model A: Standalone LLM

Model B: LLM + RAG

Model C: Proposed Multi-Agent LLM + RAG + Verification Framework

This comparison can determine whether adding retrieval and specialized agents improves scientific research performance.

7.2 Dataset and Knowledge Sources

The evaluation framework can use domain-specific scientific literature and publicly available datasets.

Potential sources include:

- PubMed for biomedical research;
- Crossref for publication metadata;
- Semantic Scholar for scholarly literature;
- arXiv for preprints;
- domain-specific open datasets;
- government and institutional datasets.

Each source should be stored with provenance information.

7.3 Hypothesis Generation Evaluation

Generated hypotheses can be evaluated using:

$$[H_{\text{score}} = w_1N + w_2E + w_3T + w_4F + w_5S]$$

where:

- (N) = novelty score;
- (E) = evidence-support score;
- (T) = testability score;
- (F) = falsifiability score;
- (S) = scientific significance;
- (w_i) = corresponding weighting factors.

Human experts can independently score generated hypotheses to establish an external benchmark.

7.4 Factuality Evaluation

Factuality can be evaluated by determining whether claims generated by the system are supported by retrieved scientific evidence.

A factuality score may be calculated as:

$$[F = \frac{C_s}{C_t} \times 100]$$

where:

- (C_s) = number of scientifically supported claims;
- (C_t) = total number of evaluated claims.



7.5 Citation Accuracy

Citation accuracy should examine whether:

1. the cited paper actually exists;
2. the citation contains correct metadata;
3. the cited source supports the claim;
4. the source is relevant to the claim.

7.6 Reproducibility

Reproducibility should be evaluated by independently rerunning the computational experiment using:

- the same dataset;
- the same code;
- the same model parameters;
- the same experimental conditions.

A reproducibility rate can be expressed as:

$$R = \frac{E_r}{E_t} \times 100$$

where:

- (E_r) = experiments successfully reproduced;
- (E_t) = total experiments attempted.

PROPOSED EVALUATION METRICS

The framework should be evaluated across multiple dimensions.

Metric	Purpose
Hypothesis Novelty	Measures originality of generated hypotheses
Scientific Validity	Determines whether hypotheses and conclusions are scientifically sound
Factuality	Measures evidence-supported claims
Citation Accuracy	Evaluates correctness of references
Experiment Quality	Measures methodological quality
Reproducibility	Determines whether experiments can be repeated
Research Efficiency	Measures reduction in researcher workload
Human Agreement	Compares AI outputs with expert evaluation
Tool Reliability	Measures successful tool/API execution
Error Rate	Measures hallucinations and methodological errors

EXPECTED RESULTS

The proposed framework is expected to demonstrate that multi-agent systems can provide broader research capabilities than standalone language models.

The expected improvements include:

9.1 Improved Literature Coverage

A dedicated literature agent can systematically search multiple sources and reduce dependence on a single model's internal knowledge.

9.2 Improved Hypothesis Quality

Generating multiple candidate hypotheses and subjecting them to independent criticism can reduce poorly formulated or unsupported hypotheses.



9.3 Improved Experimental Planning

An experiment-design agent can translate research questions into structured experimental protocols and identify appropriate datasets and analytical techniques.

9.4 Reduced Hallucination

RAG-based evidence retrieval combined with citation verification should reduce unsupported scientific claims.

9.5 Improved Reproducibility

Automated provenance tracking can preserve the sequence of actions, sources, parameters, datasets, and code used during the research process.

9.6 Increased Research Efficiency

Automating repetitive activities may allow researchers to spend more time on scientific interpretation, creativity, experimental judgment, and ethical decision-making.

DISCUSSION

The proposed framework represents a transition from **AI as a research assistant** toward **AI as a coordinated computational research ecosystem**.

The distinction is important. A conventional AI assistant typically waits for a researcher to issue instructions. In contrast, an agentic system can decompose an objective into smaller tasks and coordinate multiple specialized agents.

However, autonomy should not be interpreted as unrestricted independence. Scientific research involves ethical, methodological, and contextual judgments that cannot always be automated safely.

For example, an AI system may identify a statistically significant relationship but fail to recognize an important confounding variable. Similarly, an LLM may generate a scientifically plausible hypothesis that is actually already established in literature. Therefore, independent validation and human oversight remain essential.

The proposed critic-agent architecture addresses part of this problem by requiring the system to challenge its own outputs. However, multi-agent systems can still share common biases because agents may rely on the same underlying model or evidence sources.

Consequently, future implementations should consider heterogeneous models, independent evidence sources, formal statistical validation, and human expert review.

Another important consideration is scientific accountability. If an autonomous system generates an incorrect hypothesis or flawed experimental conclusion, responsibility must remain clearly defined. Research institutions should therefore establish governance mechanisms specifying which decisions can be automated and which require human approval.

ETHICAL AND GOVERNANCE CONSIDERATIONS

Autonomous scientific research introduces several ethical challenges.

11.1 Research Integrity

AI-generated research should not fabricate data, references, experimental observations, or research results.



11.2 Data Privacy

When research involves human participants, personal or clinical data must be protected using appropriate privacy and security mechanisms.

11.3 Transparency

Researchers should disclose when AI agents contribute substantially to research design, analysis, or interpretation.

11.4 Human Oversight

High-risk decisions should require human authorization. AI systems should support researchers rather than obscure responsibility.

11.5 Reproducibility and Auditability

Every autonomous action should ideally be logged so that researchers can reconstruct how a conclusion was reached.

LIMITATIONS

The proposed framework has several limitations.

First, the framework is conceptual and requires experimental implementation to determine its real-world effectiveness. Second, LLM-based agents remain vulnerable to hallucination and reasoning errors. Third, scientific databases may contain incomplete, outdated, or contradictory information.

Fourth, evaluating hypothesis novelty is particularly difficult because novelty requires extensive knowledge of existing literature. Fifth, autonomous experiment design cannot replace laboratory expertise in situations involving physical safety, biological risk, clinical intervention, or ethical approval.

Finally, computational cost may increase substantially as multiple agents independently process literature, generate hypotheses, perform analysis, and validate results.

FUTURE RESEARCH DIRECTIONS

Future research can extend the proposed framework in several directions.

13.1 Autonomous Laboratory Integration

Future systems may connect AI agents with robotic laboratories capable of conducting controlled experiments.

13.2 Multimodal Scientific Agents

Agents could integrate:

- scientific text;
- images;
- microscopy;
- genomic data;
- sensor data;
- tables;
- graphs;
- experimental measurements.

13.3 Self-Correcting Research Loops

The framework could automatically revise hypotheses when experimental results contradict predictions.



13.4 Scientific Knowledge Graphs

Knowledge graphs can improve relationships between researchers, methods, datasets, findings, chemicals, diseases, and scientific concepts.

13.5 Formal Verification

Mathematical and computational verification could be integrated to validate equations, statistical procedures, simulation results, and scientific claims.

13.6 Domain-Specific Scientific Agents

Future research could develop specialized agents for:

- pharmaceutical discovery;
- climate science;
- materials science;
- engineering;
- medicine;
- social science;
- economics.

CONCLUSION

This paper proposed a **Multi-Agent Scientific Research Framework (MASRF)** for autonomous scientific research using agentic AI, large language models, retrieval-augmented generation, specialized research agents, computational tools, and independent verification mechanisms.

The framework divides the scientific research lifecycle into coordinated stages including research planning, literature discovery, knowledge synthesis, hypothesis generation, hypothesis criticism, experiment design, data analysis, scientific criticism, and final validation.

A major contribution of the proposed architecture is the integration of **generation and verification**. Instead of allowing a single AI system to generate and accept its own conclusions, the framework introduces specialized critic and validation agents. This approach is intended to improve scientific validity, factuality, citation accuracy, and reproducibility.

The framework does not propose eliminating human researchers. Rather, it positions humans as scientific supervisors responsible for high-level objectives, ethical decisions, interpretation, and final approval. Such a human–AI collaboration model provides a more realistic pathway toward autonomous scientific research. Future empirical studies should implement the proposed architecture and compare it with standalone LLM and LLM-RAG systems using standardized benchmarks for hypothesis quality, factuality, experimental design, reproducibility, efficiency, and expert agreement.

Agentic AI therefore represents a potentially important evolution in scientific computing: from systems that **answer scientific questions** toward systems that can **systematically investigate scientific questions**, provided that autonomy is accompanied by evidence grounding, independent verification, transparency, reproducibility, and responsible human oversight.

REFERENCES

1. Boiko, D. A., MacKnight, R., & Gomes, G. (2023). Emergent autonomous scientific research capabilities of large language models. *Nature Communications*, 14, 5760. <https://doi.org/10.1038/s41467-023-44591-7>



2. Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2024). Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6, 525–535. <https://doi.org/10.1038/s42256-024-00832-8>
3. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv*. <https://doi.org/10.48550/arXiv.2312.10997>
4. Guo, T., Chen, X., Wang, Y., Chang, Y., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16), 17919–17927. <https://doi.org/10.1609/aaai.v38i16.29796>
5. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55. <https://doi.org/10.1145/3703155>
6. Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2310.06770>
7. Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv*. <https://doi.org/10.48550/arXiv.2408.06292>
8. M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2024). Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6, 525–535.
9. Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. <https://doi.org/10.1145/3586183.3606763>
10. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2024). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.
11. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., He, X., & others. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18, 186345. <https://doi.org/10.1007/s11704-024-40231-1>
12. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2210.03629>