Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

LLM-Based Multi-Agent Systems for Automated Research: A Systematic Review of Architectures, Capabilities, and Evaluation Practices

Rao Hamza Bilal
School of International Education, Wuhan Institute of Technology, Wuhan, China
raohamzabilal@gmail.com

Sanaullah Memon
Department of Information Technology Shaheed Benazir Bhutto University Campus Naushahro Feroze, Sindh, Pakistan
sanaullah.memon_nf@sbbusba.edu.pk

Iftikhar Ahmad
Department of Computer Science, University of Swabi, Khyber Pakhtunkhwa, Pakistan
Ifee910@gmail.com

Aamish Ali
UGLC (University Gillani Law College), Bahauddin Zakriya University Multan
(LLB)
aamishali1995@gmail.com

Sandeep Kumar
Department of Computer Science, Sukkur IBA University
sandeeplakhani340@gmail.com

ABSTRACT

Large language model (LLM) agents are becoming more prevalent with teams organized to suggest hypotheses, to search the literature, to conduct experiments and to write papers with very little human input. The capabilities these multi-agent systems are capable of is widely distributed throughout machine-learning venues, domains journals, and technical reports. This review presents a synthesis of 50 sources published from 2023 to 2025 that address four questions: What is architecture? What are the research-lifecycle aspects covered? What is evaluated? What is failure? Sources were located using keyword search and citation chain, filtered with explicit criteria and coded according to the following categories: contribution type, architectural pattern, research stage, level of validation. There were six architectural patterns identified, the most prevalent of which was role-specialised pipelines. Literature synthesis and idea generation are the most well supported stages, with fewer systems covering experimentation, analysis, and writing, and these systems are mainly validated on

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

machine-learning problems. There are few systems that report laboratory identification of agent-generated proposals, primarily those in chemistry and biology. There is a consistent finding in benchmarks that agents are able to accomplish only a small fraction of the tasks that are typically research-engineering tasks, and failure analyses indicate that a significant number of failures are due to specification, coordination, and verification issues, not just to the inability to do the tasks themselves. We propose a reference architecture that incorporates and implements all of the above requirements for credible automated research, namely: verification-centred design, explicit provenance, calibrated human oversight, and standardised reporting.

Keywords— large language models; multi-agent systems; automated scientific discovery; AI scientist; research agents; systematic review

1. Introduction

Research is a series of dependent judgments: picking a question, reading what is already known, formulating a hypothesis, designing and conducting a test, interpreting noisy results, and communicating the results for others to verify. The judgement between the links remains human, while software has been supportive to each link for decades. In recent years, such division of labour has been questioned when the large language models (LLMs) serve as agents, i.e., are components that plan, call tools, observe results, and adjust their actions (L. Wang et al., 2024; Yao et al., 2023). By assigning different roles to multiple such agents and having them share information with each other, an LLM-based multi-agent system (LLM-MAS) is created (Guo et al., 2024; Li et al., 2024).

The highest level implementation of this concept is automated research. The AI Scientist pipeline can generate ideas, write and execute code, create figures, write the manuscript, and score it with an automated reviewer (Lu et al., 2024), and the second version eliminates the need for human-written code templates with agentic tree search (Yamada et al., 2025). Google has an AI co-scientist, backed by a team of specialised agents, that creates, debates, ranks, and grows hypotheses, with some of its repurposing ideas for drugs and target discovery being followed up in the lab (Gottweis et al., 2025). An LLM-driven planner helped guide a web search, documentation retrieval, code execution, and lab automation process to perform cross-coupling reactions (Boiko et al., 2023); in structural biology, a ‘virtual lab' of agents was used to design nanobody candidates for testing in the lab (Swanson et al., 2025).

There's a tinge of fragility in the enthusiasm. Failure analysis of widely used multi-agent frameworks suggests that the improvements over single-agent baselines are often small and failures tend to concentrate on the task specification, misalignment between agents, and on weak verification (Cemri et al., 2025). Sometimes even the best configurations can only solve a fraction of the problems in benchmarks that require the agents to perform real machine-learning engineering problems or to reproduce published papers (Chan et al., 2025; Starace et al., 2025). In addition, methodological researchers caution that agent assessments often focus on accuracy rather than cost or reproducibility (Kapoor et al., 2024), and that scientific AI tools can generate 'illusions of understanding,' when researchers believe that they have understood a concept but in

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

fact have not (Messeri & Crockett, 2024).

Existing surveys describe LLM-MAS in general (S. Chen et al., 2024; Guo et al., 2024; Li et al., 2024; Tran et al., 2025) or LLMs in scientific discovery in general (Reddy & Shojaee, 2025; T. Zheng et al., 2025). The problem is that there is no synthesis that asks the question that a research team would pose before implementing such a system: which parts of the research process can an agent team perform, how, and with what evidence? This review provides answers to four research questions that fill this gap.

- **RQ1.** Which architectural patterns do LLM-based multi-agent systems use for research tasks?
- **RQ2.** Which stages of the research lifecycle are automated, and at what level of validation?
- **RQ3.** How are such systems evaluated, and what do the benchmarks reveal?
- **RQ4.** Which failure modes and risks limit their reliability and trustworthiness?

The following is a summary of the four contributions that the paper makes. Firstly, it offers a coded collection of 50 sources, using clear inclusion criteria. Second, it suggests a taxonomy of six architectural patterns. Third, it provides coverage and validation information for 13 research oriented systems in research stage. Fourth, it combines evaluation and failure evidence into a reference architecture for automated research based on verification. Section 2 provides background, Section 3 outlines the methodology used for the review, Section 4 outlines the corpus and its characteristics, Sections 5-8 provide answers to the four research questions, Section 9 presents implications for this review, as well as the reference architecture, and Section 10 covers limitations and future directions for this review, followed by Section 11, which concludes.

2. Background

2.1 From prompted models to agents

There were two techniques to transform LLMs from text generators to agents. The two techniques were used to transform LLM from a text generator to an agent. In ReAct, reasoning traces are interleaved with actions, allowing a model to query external sources to revise its plan based on what it observes (Yao et al., 2023). Verbal feedback is stored in episodic memory when it is a result of a failure and later attempts are condition upon it (Shinn et al., 2023). These ideas along with the module of tool use, memory, and planning describe the agent abstraction described in surveys of autonomous agents (L. Wang et al., 2024).

2.2 Multi-agent organisation

Multi-agent designs assign distinct roles, prompts or tools to distinct model instances. CAMEL investigated role-playing interactions of communicating agents guided by inception prompts (Li et al., 2023). AutoGen provided an overall programming paradigm for a conversation, where customisable agents, tools and humans communicate with each other (Wu et al., 2024). MetaGPT embedded the SOC of human organisations into a series of roles in an assembly line (Hong et al., 2024), while ChatDev wrapped role-specialised agents into a chain of natural language conversations in the entire design, coding, and test cycles (Qian et al., 2024). AgentVerse added dynamic group composition, so that the set of agents adapts to the task at hand (W.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Chen et al., 2024).

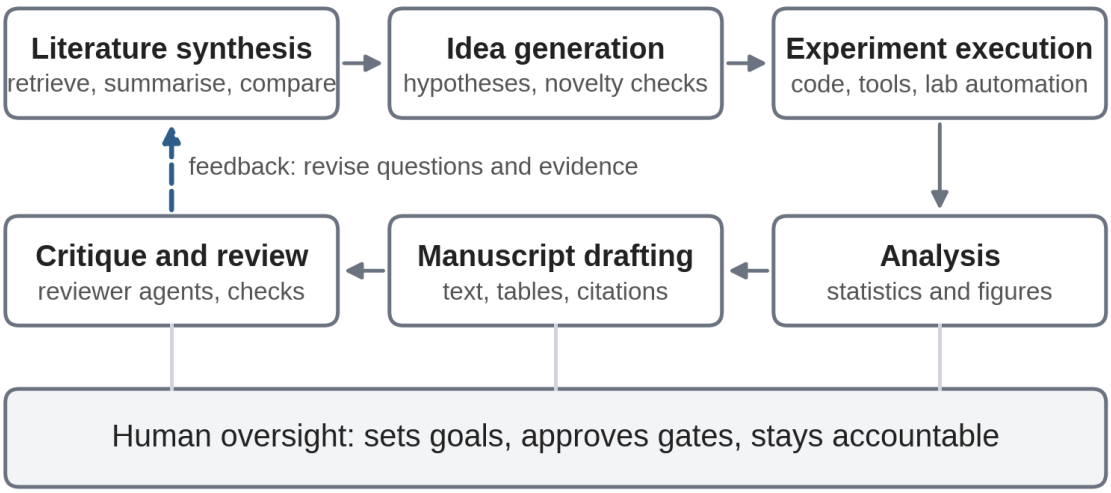
A second family is one in which the division of labor does not apply, but the family is dependent on interaction. In the original study (Du et al., 2024), the multi-agent debate system was introduced to enhance the factuality and reasoning ability of the generated instances. A variant of the debate system was also introduced (T. Liang et al., 2024) to include structured disagreement to counteract the tendency of a single model to settle on the first answer produced. Mixture-of-Agents stacks are cascaded with each agent viewing the output of the preceding layer (J. Wang et al., 2024). Populations of memory-enabled agents were able to demonstrate realistic social behaviors (Park et al., 2023). More recent research considers the organisation as a search or optimisation object (Hu et al., 2025; Zhang et al., 2025).

2.3 Automated research

For systems that involve some or all parts of the empirical cycle, in Figure 1, in which agents execute at least one of the following activities – literature review, idea generation or hypothesis, experiment or analysis, writing of results, and critique – we use automated research. The agents can be assistants, working under the direction of a scientist; co-scientists, who can suggest and improve options; or independent scientists, who can participate in the whole cycle (Gottweis et al., 2025; Lu et al., 2024). The difference is important because it creates the zone of tolerance for the amount of human checking that is available at each of the stages. This evolution has been explored in previous surveys from LLM as a tool to LLM as an "autonomous scientist" (T. Zheng et al., 2025) and opportunities and open problems of generative AI for discovery (Reddy & Shojaee, 2025).

Figure 1

Stages of the Automated Research Cycle and the Role of Human Oversight



Note. Conceptual synthesis by the authors, drawing on the stage structure described by Lu et al. (2024) and Gottweis et al. (2025).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

3. Method

3.1 Review design

It is a protocol-driven review with an explicit search, screening and coding process but it is not a meta-analysis and does not pretend to be exhaustive of a field which expands daily. The unit of analysis is a source (conference paper, journal article, preprint, or technical report). Due to the differences in tasks, model, and metric among the studies, the results are not commensurable and were represented through narrative synthesis, instead of pooled effect sizes, which was reported with a frequency count.

3.2 Search and selection

Sources were found using keywords searches on arXiv, publishers' platforms, as well as the proceedings of the major conferences in the field of natural language processing, machine learning and artificial intelligence (ACL, EMNLP, NAACL, ICLR, ICML, NeurIPS, and IJCAI) and journals such as Advanced Materials and Nature. The search strings included phrases like 'LLM multi-agent', 'AI scientist', 'automated scientific discovery', 'research agent', 'agent benchmark', and 'multi-agent failure'. Citation chaining from four surveys (S. Chen et al., 2024; Guo et al., 2024; Li et al., 2024; T. Zheng et al., 2025) added further candidates. Titles and abstracts were first screened and then full texts, using the criteria listed in Table 1. 50 sources were kept. The corpus is purposive since only designs of direct relevance to research automation were covered, but not the entire agent literature: see Section 10 for discussion of implications for the conclusions.

Table 1

Eligibility Criteria Used to Select Sources

Criterion	Inclusion	Exclusion
Period	Published or posted between January 2023 and December 2025	Earlier work, used only as background
Technology	LLM-based agents or multi-agent systems, or evaluation and failure studies that target them	Multi-agent systems without LLMs; model-only papers without agentic components
Relevance	Addresses research tasks (literature, ideation, experiments, writing, review) or reports a design, evaluation, or failure result that transfers to them	General chatbots, tutoring, and consumer applications
Source quality	Peer-reviewed venue or widely used preprint with a described method; one industry engineering report with a detailed architecture	Marketing material and posts without technical description
Language	English	Other languages

Note. Criteria were applied at title, abstract, and full-text screening.

3.3 Coding scheme

All sources were recorded on four criteria. The first was contribution type: survey, empirical study, position or critique, framework or method, research system,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

benchmark or evaluation method. The second was architectural pattern that was used for the sources describing multi-agent designs. The third was the research stage dealt with (literature synthesis, idea generation, execution of the experiment, analysis, drafting of the manuscript, or critique and review) related to research systems. The fourth was the extent of validation, which was measured at four levels: V0 – demonstration only, V1 – automated or benchmark evaluation, V2 – human or expert evaluation, and V3 – experimental or laboratory validation of the system outputs. All codes were given by one coder; inter-rater reliability was not calculated. The numbers below do not therefore give an estimate of the composition of the field, but are descriptive statistics of this corpus.

3.4 Synthesis

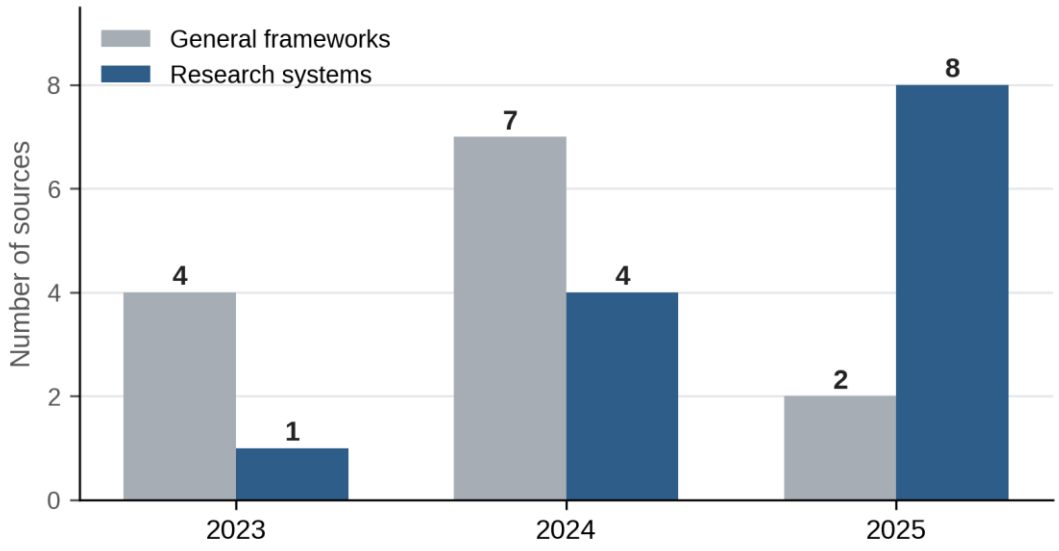
Sources were coded for each research question, compared against design and evidence, and summarised narratively within each research question. If quantitative results were reported in the original papers, they are presented as reported by the authors; no system was re-run. Statements of our own interpretation are identified and claims in the Results are limited to what is stated in the cited sources.

4. Corpus Overview

The corpus contains 7 sources from 2023, 24 from 2024, and 19 from 2025. It is important to note that the smaller number of 2025 does not reflect a decline in activity: the corpus is purposive, and many of the 2025 sources were initially published as preprints which were not yet indexed by venues. A more descriptive pattern has to do with the type of work that was created. The number of new general-purpose frameworks saw a decline from 4 in 2023 and 7 in 2024 to 2 in 2025, while the number of research systems grew from 1 in 2023 to 4 in 2024 and 8 in 2025, as indicated in Graph 1. The community shifted from the question of how agents can be organised, to the question of what organised agents can do for science.

Graph 1

Sources per Year, General Frameworks versus Research Systems



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

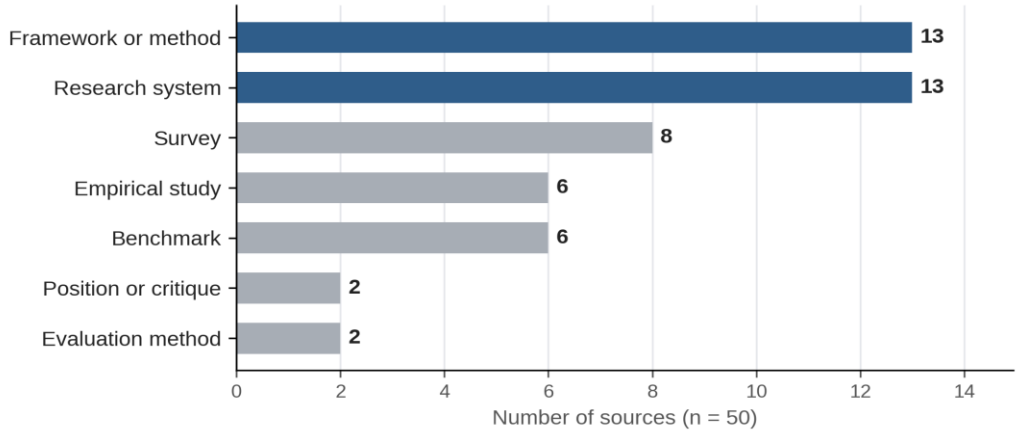
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Note. Counts come from the coded corpus (n = 50) and describe the sources reviewed here, not the field as a whole.

Graph 2 shows the composition by contribution type. Frameworks and research systems together account for 26 of the 50 sources, surveys for eight, and empirical studies and benchmarks for six each. Only two sources are position papers (Kapoor et al., 2024; Messeri & Crockett, 2024) and two propose evaluation methods (Zhuge et al., 2025; L. Zheng et al., 2023). Even so, the small number of critical sources carries disproportionate weight in the later sections, because they identify the problems that engineering papers rarely report.

Graph 2

Corpus Composition by Contribution Type



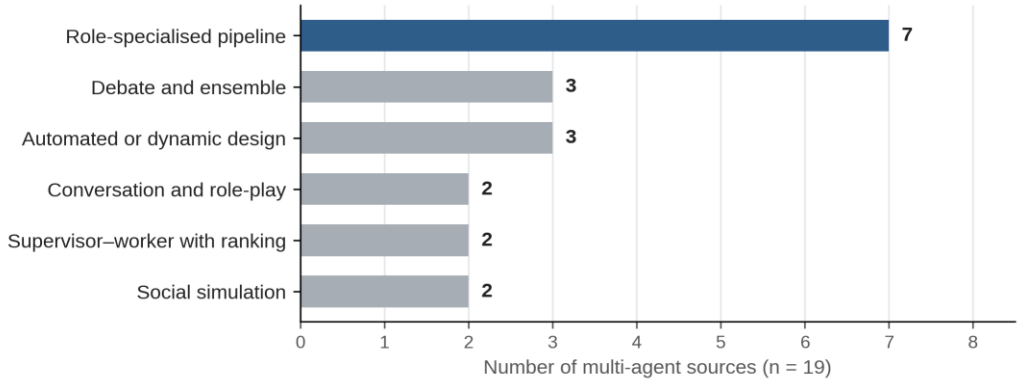
Note. Categories are mutually exclusive; each source received one code.

5. Architectural Patterns (RQ1)

Nineteen sources describe multi-agent designs that could be assigned to one of six patterns (Graph 3 and Figure 2). Each source received the pattern that best matches its dominant organising principle, although many systems combine features of several.

Graph 3

Architectural Patterns among Multi-Agent Sources



Note. Counts are based on 19 sources that describe multi-agent designs; each source received one pattern.

Role-specialised pipelines are the most frequent pattern, with seven sources. Fixed

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

roles mirror a human workflow and pass artefacts from one role to the next. MetaGPT assigns roles according to standard operating procedures (Hong et al., 2024), ChatDev follows design, coding, and testing phases (Qian et al., 2024), and Agent Laboratory moves from literature review through experimentation to report writing, with optional human feedback at each stage (Schmidgall et al., 2025). SciAgents couples specialised agents with reasoning over an ontological knowledge graph to propose materials-science hypotheses (Ghafarollahi & Buehler, 2025). MARG distributes a manuscript among several agents so that long papers can be reviewed despite context limits (D'Arcy et al., 2024), VIRSCI simulates a research team that discusses and refines ideas (Su et al., 2025), and in the Virtual Lab a principal-investigator agent convenes meetings with specialist agents and a critic (Swanson et al., 2025). Pipelines are easy to inspect and debug, but a mistake made early propagates through every later role.

Conversation and role-play frameworks (two sources) leave the sequence of steps to the dialogue itself. CAMEL pairs two role-playing agents (Li et al., 2023), and AutoGen lets developers compose agents, tools, and humans in flexible conversation patterns (Wu et al., 2024). They are general-purpose infrastructure rather than research systems, which explains why later research systems often build on them.

Debate and ensemble methods (three sources) improve the quality of a single answer at inference time. Debate rounds raised factuality and reasoning accuracy in the study that introduced them (Du et al., 2024), forced disagreement encouraged more divergent reasoning (T. Liang et al., 2024), and layered aggregation of several models' outputs produced strong results on instruction-following benchmarks (J. Wang et al., 2024). These methods multiply cost with the number of agents and rounds. We note, as our own interpretation, that agreement among instances of one model is weaker evidence than agreement among independent evidence sources, because their errors are correlated.

Supervisor–worker systems with ranking (two sources) place a coordinating agent above specialised workers. The AI co-scientist uses a supervisor to assign tasks to agents that generate, reflect on, rank, and evolve hypotheses, and it relies on tournament-style comparison of hypotheses so that quality can improve with additional computation (Gottweis et al., 2025). Anthropic (2025) described a research system in which a lead agent decomposes a query and delegates parts of it to subagents that search in parallel; the engineers reported that this configuration outperformed a single agent on their internal research evaluation while consuming substantially more tokens. Both examples add an explicit selection mechanism, which separates them from simple division of labour.

Automated or dynamic design (three sources) treats the organisation as a variable. AgentVerse recruits agents for the task and adjusts the group between rounds (W. Chen et al., 2024), ADAS uses a meta-agent that programs new agent designs in code and archives successful ones (Hu et al., 2025), and AFlow searches over code-represented workflows with Monte Carlo tree search (Zhang et al., 2025). These methods have been evaluated mainly on coding, mathematics, and question answering, so their value for research tasks remains largely untested.

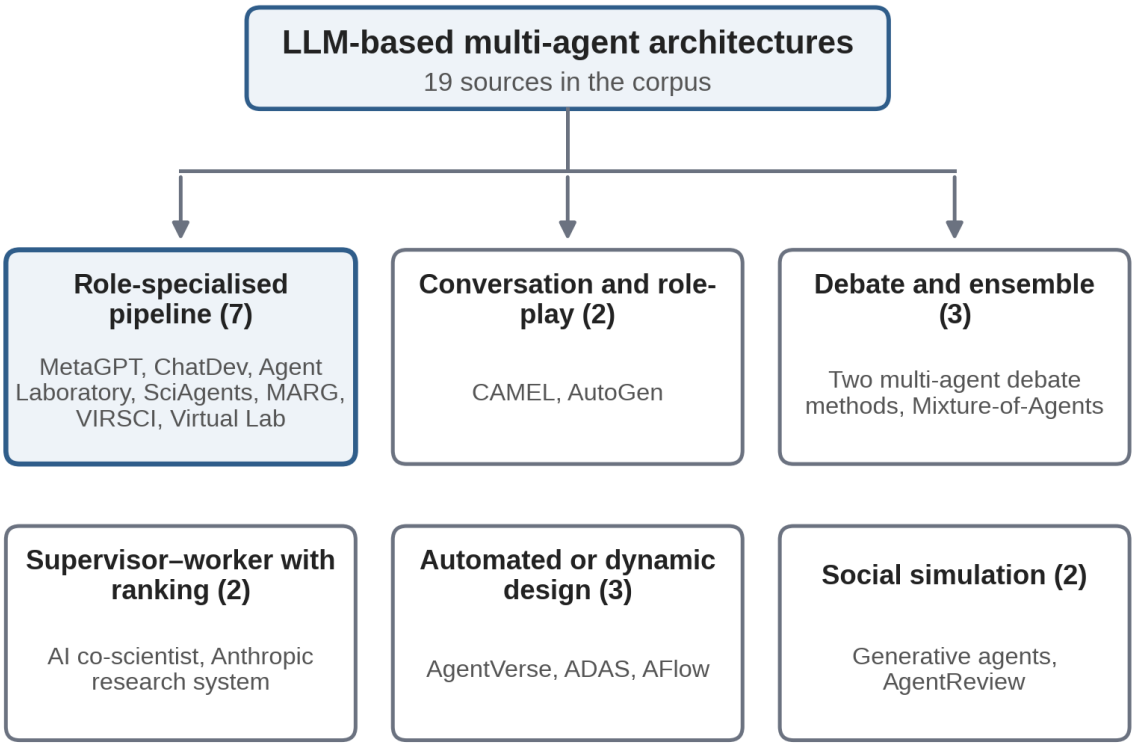
Social simulation (two sources) uses agents to model human behaviour rather than to solve a task. Generative agents showed believable individual and group behaviour in a

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

simulated town (Park et al., 2023), and AgentReview applied the same approach to peer review, assigning reviewer, author, and area-chair roles to examine how social influence and bias shape decisions (Jin et al., 2024). For automated research, such simulations are useful as testbeds for the institutions in which research is judged.

Figure 2
Taxonomy of Architectural Patterns in LLM-Based Multi-Agent Systems



Note. Numbers in parentheses are source counts from the coded corpus. Each source received one pattern.

In answer to RQ1, role-specialised pipelines are the most common pattern (7 of 19 sources, 37%), but the designs that report the most substantial research outcomes combine specialisation with an explicit mechanism for selecting among candidates, such as tournament ranking, critic agents, or experimental testing. This observation is consistent with the failure literature reviewed in Section 8, which locates many breakdowns in verification rather than generation (Cemri et al., 2025).

6. Lifecycle Coverage and Validation (RQ2)

Thirteen sources in the corpus are research systems, meaning that they apply LLM

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

agents directly to one or more stages of the research cycle. Table 2 summarises their domain, design, coverage, and validation level, and Graph 4 shows how many of them address each stage.

Table 2

Research-Oriented Systems: Domain, Design, Stage Coverage, and Validation

System	Domain	Design	Lit	Idea	Exp	Ana	Wri	Rev	Val
AI Scientist (Lu et al., 2024)	Machine learning	Pipeline	•	•	•	•	•	•	V1
AI Scientist-v2 (Yamada et al., 2025)	Machine learning	Tree search		•	•	•	•		V2
AI co-scientist (Gottweis et al., 2025)	Biomedicine	Supervisor–worker	•	•				•	V3
Agent Laboratory (Schmidgall et al., 2025)	Machine learning	Pipeline	•		•	•	•		V2
Coscientist (Boiko et al., 2023)	Chemistry	Planner with tools	•		•	•			V3
ResearchAgent (Baek et al., 2025)	Cross-domain	Generator with reviewers	•	•				•	V2
SciAgents (Ghafarollahi & Buehler, 2025)	Materials science	Knowledge-graph pipeline	•	•				•	V1
Virtual Lab (Swanson et al., 2025)	Structural biology	Meeting-based team		•	•	•		•	V3
PaperQA2 (Skarlinski et al., 2024)	Biomedical literature	Agentic retrieval	•						V2
VIRSCI (Su et al., 2025)	Idea generation	Team discussion	•	•					V1
SciMON (Q. Wang et al., 2024)	NLP ideation	Retrieval and novelty loop	•	•					V2

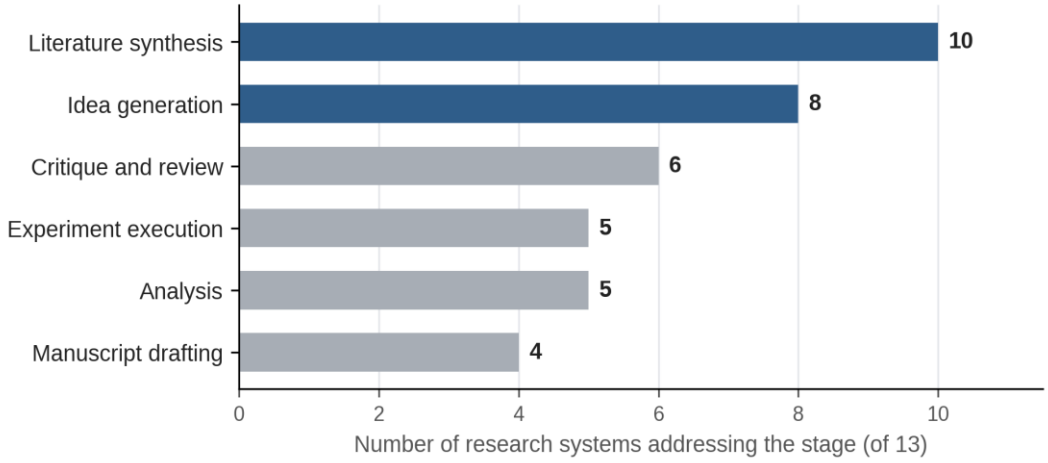
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

MARG (D'Arcy et al., 2024)	Paper review	Review agents			•	V2
Anthropic research system (Anthropic, 2025)	Web research	Lead with subagents	•		•	V1

Note. Lit = literature synthesis; Idea = idea generation; Exp = experiment execution; Ana = analysis; Wri = manuscript drafting; Rev = critique and review; Val = validation level (V1 automated or benchmark evaluation, V2 human or expert evaluation, V3 laboratory confirmation). Codes assigned by the authors from the cited descriptions.

Graph 4

Research Stages Addressed by the Thirteen Research Systems



Note. A system can address several stages, so counts do not sum to 13.

6.1 Literature synthesis and idea generation

Literature work and ideation are the best-covered stages: ten systems retrieve or synthesise literature and eight generate ideas. PaperQA2 combines iterative retrieval, evidence gathering, and answer generation, and its authors reported that it matched or exceeded domain experts on several literature tasks, including summarisation and the detection of contradictions in the published record (Skarlinski et al., 2024). For ideation, ResearchAgent generates problems, methods, and designs from the literature and refines them through reviewing agents (Baek et al., 2025), SciMON retrieves inspirations and optimises ideas for novelty (Q. Wang et al., 2024), and VIRSCI uses team discussion (Su et al., 2025).

The strongest evidence on idea quality comes from a human study with more than 100 NLP researchers, in which LLM-generated ideas were rated as more novel than ideas written by human experts but slightly weaker on feasibility (Si et al., 2025). The same study reported limited diversity among generated ideas and unreliable LLM-based evaluation of ideas, both of which matter for any pipeline that lets agents rank their own proposals.

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

6.2 Experiments, analysis, and writing

Only five systems execute experiments and analyse results, and only four draft manuscripts. The AI Scientist runs the complete cycle on machine-learning problems from human-written templates, at a reported cost of roughly 15 US dollars per paper, and its authors document failure modes such as flawed implementations and misleading comparisons (Lu et al., 2024). The AI Scientist-v2 removes the templates and explores experiments through tree search; its authors reported that one fully AI-generated manuscript scored above the average acceptance threshold at an ICLR 2025 workshop (Yamada et al., 2025). Agent Laboratory reported that human feedback at each stage improved the quality of its outputs (Schmidgall et al., 2025), which suggests that partial autonomy may be the more reliable operating mode at present.

Domain systems that act on the physical or computational world are scarcer but more convincing. Coscientist planned and executed palladium-catalysed cross-coupling reactions through laboratory automation (Boiko et al., 2023). The Virtual Lab designed nanobody candidates against SARS-CoV-2 variants using in silico tools, and a subset was tested experimentally (Swanson et al., 2025). The AI co-scientist's proposals for acute myeloid leukaemia drug repurposing and liver fibrosis targets were tested in vitro (Gottweis et al., 2025). These three systems are the only ones in the corpus coded at validation level V3.

6.3 Critique and review

Six systems include a critique or review step. Peer-review-oriented work shows partial but real value: large-scale analysis found that LLM-generated feedback on manuscripts overlapped substantially with human reviewers' comments, and most participants in a user study found the feedback helpful (W. Liang et al., 2024). Multi-agent review generation was designed to handle full-length papers (D'Arcy et al., 2024). Simulation work cautions that reviewer agents reproduce social biases found in human review (Jin et al., 2024), and LLM judges in general show position and verbosity biases even while agreeing with human preferences at rates comparable to agreement between humans (L. Zheng et al., 2023).

6.4 Validation levels

Of the thirteen research systems, four were coded V1 (automated or benchmark evaluation), six V2 (human or expert evaluation), and three V3 (laboratory confirmation). None was coded V0, because authors generally included at least one evaluation, but the choice of evaluator matters. Machine-learning systems are usually judged by automated reviewers or workshop review, whereas chemistry and biology systems are judged against experiments. We interpret this as a consequence of what each field can check cheaply: software experiments are fast to run but their meaning is easy to dispute, while wet-lab results are slow to obtain but harder to argue with.

In answer to RQ2, agents currently cover the reading and proposing stages far more widely than the doing and writing stages, and the highest validation level has been reached in only three systems, all in domains with an external experimental check.

7. Evaluation Practices and Benchmarks (RQ3)

7.1 Benchmarks of research-relevant capability

Six benchmarks in the corpus probe skills that automated research depends on (Table

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

3). They share a design principle, which is to use tasks with checkable outcomes such as passing tests, achieving a competition score, or reproducing reported results. They differ in what they treat as the unit of work, from a single repository issue (Jimenez et al., 2024) to the replication of an entire paper (Starace et al., 2025).

Table 3

Benchmarks Relevant to Automated Research

Benchmark	Focus	Scale	Task and scoring
SWE-bench (Jimenez et al., 2024)	Software engineering	2,294 issues from 12 Python repositories	Produce a patch that passes the repository's tests
MLAgentBench (Q. Huang et al., 2024)	Machine-learning experimentation	13 tasks	Improve a model through iterative experimentation
MLE-bench (Chan et al., 2025)	Machine-learning engineering	75 Kaggle competitions	Reach medal thresholds on competition leaderboards
ScienceAgentBench (Z. Chen et al., 2025)	Data-driven scientific discovery	102 tasks from 44 publications	Write a self-contained program; score outputs and code
CORE-Bench (Siegel et al., 2024)	Computational reproducibility	270 tasks from 90 papers	Run a paper's code and answer questions on the results
PaperBench (Starace et al., 2025)	Paper replication	20 ICML 2024 papers	Replicate a paper from scratch; graded against rubrics

Note. Scale and task descriptions follow the original publications.

7.2 What the results show

Original evaluations agree that realistic research-engineering tasks remain difficult. In MLE-bench the best configuration reached at least bronze-medal level in roughly one in six competitions (Chan et al., 2025). On PaperBench the strongest agent achieved a replication score of about one-fifth (Starace et al., 2025), and on ScienceAgentBench the best agent solved roughly a third of tasks without expert-provided knowledge (Z. Chen et al., 2025). CORE-Bench reported low success on its hardest tier (Siegel et al., 2024), and MLAgentBench found that average success on machine-learning experimentation tasks stayed well below one half (Q. Huang et al., 2024). These figures describe performance at the time of each publication, and scores have risen as models improved, so they indicate the difficulty of the tasks rather than the current ceiling.

7.3 Evaluating the evaluators

Benchmarks only measure what their scoring method can see. LLM judges scale cheaply and agree with human preferences at high rates, but they carry known biases (L. Zheng et al., 2023). Agent-as-a-Judge extends the idea by letting an agent inspect

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

intermediate steps of another agent's work, and its authors reported closer agreement with human experts and lower cost than a single LLM judge on code-generation tasks (Zhuge et al., 2025). A methodological critique argues that agent evaluations should report cost alongside accuracy, because costly agents can look better than simple baselines that are cheaper and nearly as accurate, and because reproducibility problems are common (Kapoor et al., 2024).

7.4 What is not measured

Reading Table 3 as a whole, no benchmark in the corpus measures whether an agent's conclusion is scientifically correct, new, or useful, and none measures cost per validated finding. Existing benchmarks test engineering competence that discovery needs, such as setting up code, running experiments, and reproducing results, but they cannot say whether an agent chose a worthwhile question. This is our interpretation of the pattern across sources, not a claim made by any one of them.

In answer to RQ3, evaluation is fragmented, relies heavily on automated judges, and shows that agents succeed on a minority of realistic tasks, while the properties that matter most for science, namely correctness, novelty, and cost-effectiveness of discovery, are not yet measured systematically.

8. Failure Modes and Risks (RQ4)

8.1 Failures of organisation

The most direct evidence comes from an analysis of execution traces from several popular multi-agent frameworks, which produced a taxonomy of failures grouped into specification and system-design problems, inter-agent misalignment, and task verification and termination (Cemri et al., 2025). Its authors concluded that many failures reflect organisational design rather than the weakness of individual models, and that improving role specifications and orchestration helped but did not remove the problems. This finding fits the observation in Section 5 that systems with explicit selection mechanisms report stronger results.

8.2 Limits of the underlying models

Multi-agent systems inherit the weaknesses of the models inside them. Models can generate fluent but unsupported statements (Ji et al., 2023), which is especially costly when the output is a citation or an experimental claim. Without external feedback, models often fail to correct their own reasoning and may make it worse (J. Huang et al., 2024), which weakens designs that rely on agents reviewing each other's work when all agents share the same underlying model. Performance also degrades when relevant information sits in the middle of a long context (Liu et al., 2024), a concern for agents that accumulate long histories of tool outputs.

8.3 Epistemic and institutional risks

A different class of risk concerns how scientists use the output. Messeri and Crockett (2024) described several 'visions' of AI in research and argued that each can produce illusions of understanding, including the illusion of explanatory depth and the illusion of exploratory breadth, as well as a monoculture of methods and questions. The finding that LLM-generated ideas are judged more novel but slightly less feasible (Si et al., 2025) illustrates a practical version of the problem: an agent can produce proposals that appear exciting but are costly to test. Institutional risks include overload of peer review

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

by automatically generated manuscripts and the difficulty of assigning accountability when an agent made a decision; we raise these as implications of the evidence, not as findings of the cited studies.

Table 4

Failure Classes, Supporting Evidence, and Design Responses

Failure class	Evidence	Design response
Specification and role ambiguity	Cemri et al. (2025)	Written role contracts and structured task specifications (P4)
Inter-agent misalignment	Cemri et al. (2025)	Orchestrator mediation and structured hand-offs
Weak verification and termination	Cemri et al. (2025); J. Huang et al. (2024)	Independent verifiers and execution-based checks (P1, P2)
Unsupported claims	Ji et al. (2023)	Provenance ledger and citation verification (P3)
Long-context degradation	Liu et al. (2024)	Short hand-offs, retrieval, and summarised memory
Biased or costly evaluation	L. Zheng et al. (2023); Kapoor et al. (2024)	Mixed human and automated review; cost-aware reporting
Illusions of understanding	Messeri & Crockett (2024)	Human gates and clear accountability (P5)
Novel but infeasible ideas	Si et al. (2025)	Feasibility review before committing resources (P5)

Note. P1 to P5 refer to the design principles in Section 9.3.

In answer to RQ4, the failures that most limit reliability are weak specification, poor coordination, and unverified outputs, while the main risks to science are overconfidence in fluent results and homogenisation of research questions. Table 4 links each failure class to the supporting evidence and to a design response used in the reference architecture presented next.

9. Discussion

9.1 Principal findings

Four findings answer the research questions. First, role-specialised pipelines dominate the architectures in the corpus, but the most credible research results come from designs that also select among candidates through ranking, critique, or experiment. Second, agents cover literature synthesis and ideation far more often than experimentation, analysis, and writing, and only three systems have had their outputs confirmed in a laboratory. Third, benchmarks show that agents solve a minority of realistic research-engineering tasks, and no benchmark yet measures the correctness or novelty of a scientific conclusion. Fourth, the failure literature points to specification, coordination, and verification rather than raw model capability as the main sources of breakdown.

9.2 When multi-agent designs help

The evidence does not support a blanket claim that more agents are better. Reported gains are clearest where work can be divided into independent parts that are searched

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

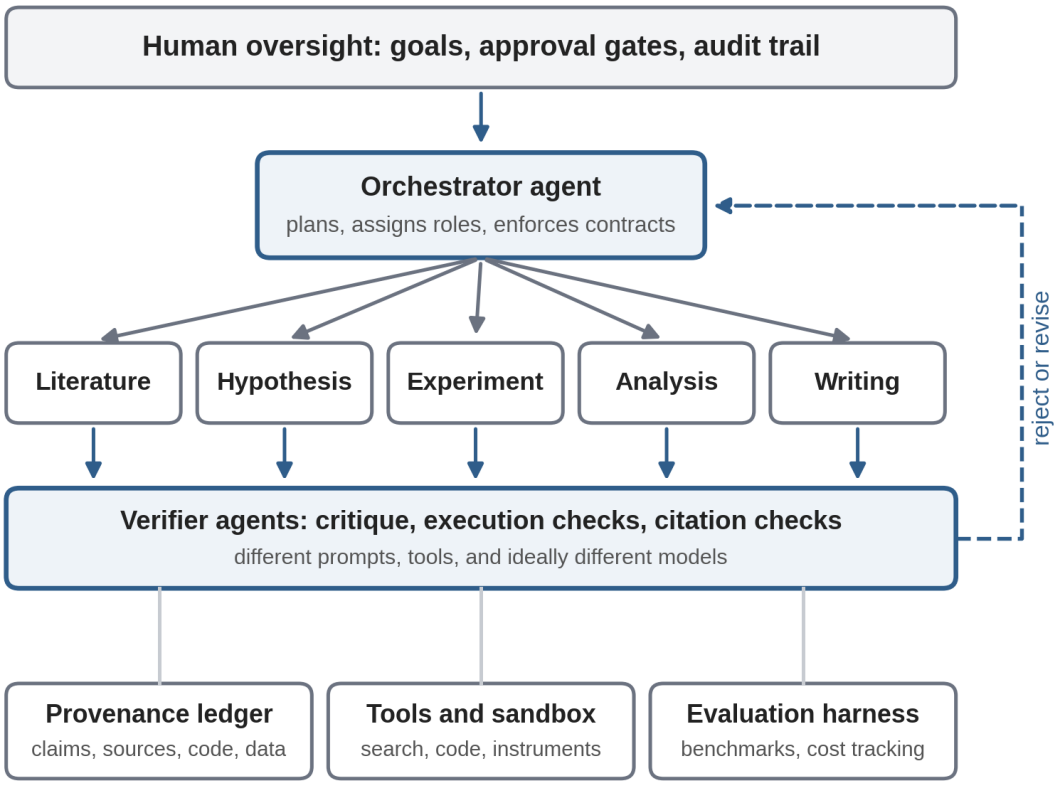
in parallel, as in open-ended web research (Anthropic, 2025), or where outputs can be checked mechanically by running code or a laboratory procedure (Boiko et al., 2023; Swanson et al., 2025). Gains are smaller or contested where agents mainly discuss among themselves (Cemri et al., 2025; Kapoor et al., 2024). A practical rule that follows, and which we offer as our interpretation, is to add an agent only when it contributes either parallel breadth or an independent check that a single agent could not supply.

9.3 A reference architecture for verification-centred automated research

Synthesising the evidence, we propose the reference architecture in Figure 3. It is not an implemented system; it organises the lessons of the reviewed sources into components and five design principles.

Figure 3

Reference Architecture for Verification-Centred Automated Research



Note. Proposed by the authors as a synthesis of the reviewed evidence. Top row: human gates; middle: generator agents under an orchestrator; below: verifier agents; bottom: shared infrastructure. The dashed path returns rejected outputs for revision.

P1. Separate generation from verification. Generator agents propose ideas, code, and text, while verifier agents with different prompts, tools, and ideally different underlying

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

models check them. Independent verification responds to the limits of self-correction (J. Huang et al., 2024) and to the correlated errors of identical agents.

P2. Prefer execution-based checks to opinion. Where a claim can be tested by running code, querying a database, or performing an assay, the test should override agent consensus, as the V3 systems in Table 2 did.

P3. Record provenance. Every claim, citation, dataset, and code version should be stored in a ledger that links the final text to its evidence. This reduces the risk of unsupported statements (Ji et al., 2023) and supports the reproducibility concerns raised by Kapoor et al. (2024) and Siegel et al. (2024).

P4. Write explicit role contracts. Each agent should have a written specification of its inputs, outputs, and stopping conditions, because specification failures were the largest class in the failure taxonomy (Cemri et al., 2025).

P5. Place humans at gates, not at the end. Human approval is most valuable before costly or irreversible actions, such as running a long experiment or sending a manuscript, which matches the finding that stage-wise human feedback improved output quality (Schmidgall et al., 2025) and the warning that scientists may over-trust fluent output (Messeri & Crockett, 2024).

9.4 Relation to earlier surveys

General surveys of LLM-MAS give broad taxonomies of profiles, communication, and capability acquisition (Guo et al., 2024; Li et al., 2024) and of applications (S. Chen et al., 2024), and a survey of collaboration mechanisms categorises actors, structures, and strategies (Tran et al., 2025). Surveys of scientific discovery describe the move from tools to autonomous scientists (T. Zheng et al., 2025) and the opportunities and challenges of generative AI for discovery (Reddy & Shojaee, 2025). The present review complements them in three ways: it restricts attention to research tasks, it codes each system by validation level rather than by capability claims alone, and it connects the failure and evaluation literature to a concrete design proposal.

10. Limitations and Future Directions

10.1 Limitations

Several limitations qualify these conclusions. The corpus is a purposive sample of 50 sources, so the counts describe this corpus and not the field. A single coder assigned all codes, and the coding of architecture and validation level involved judgement; other coders might classify some systems differently. The search window ended in December 2025 and the field changes quickly, so newer systems and benchmark results are not covered. Most performance figures are the authors' own reports, and several sources are preprints that have not completed peer review. Only English-language sources were included. Finally, because the review did not re-run any system, it cannot verify the results that the original authors reported.

10.2 Future directions

Six directions follow from the evidence. First, benchmarks are needed that score the validity and novelty of scientific conclusions, for example through prospective experimental tests. Second, evaluations should report cost per validated result alongside accuracy (Kapoor et al., 2024). Third, teams of heterogeneous models could reduce correlated errors, an idea that so far has had little direct testing in research settings.

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Fourth, provenance and reporting standards for agent-generated research would let reviewers see which steps were automated and which were checked. Fifth, controlled studies of human–agent workflows should establish where human gates add the most value. Sixth, evidence from the natural and engineering sciences should be complemented by studies in domains where experiments are slower or less well defined, such as the social sciences and medicine.

11. Conclusion

LLM-based multi-agent systems have moved within three years from general conversation frameworks to specialised research systems that read, propose, and in a few domains test scientific ideas. The evidence reviewed here shows genuine progress in literature synthesis and hypothesis generation, a handful of results confirmed in the laboratory, and persistent weaknesses in verification, coordination, and evaluation. Multi-agent organisation is a tool rather than a guarantee: it helps when it adds parallel breadth or an independent check, and it fails when roles are vague and outputs go unverified. For researchers who adopt these systems, the main lesson is to design for verification and provenance first and autonomy second. For the field, the most pressing need is evaluation that measures whether agent-generated findings are correct and worth having. Progress on those two fronts will decide whether automated research becomes a dependable instrument or remains an impressive demonstration.

References

Anthropic. (2025, June 13). *How we built our multi-agent research system* [Engineering blog post]. https://www.anthropic.com/engineering/built-multi-agent-research-system

Baek, J., Jauhar, S. K., Cucerzan, S., & Hwang, S. J. (2025). ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature, 624*(7992), 570–578. https://doi.org/10.1038/s41586-023-06792-0

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E., & Stoica, I. (2025). *Why do multi-agent LLM systems fail?* arXiv. https://doi.org/10.48550/arXiv.2503.13657

Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., Weng, L., & Mądry, A. (2025). MLE-bench: Evaluating machine learning agents on machine learning engineering. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Chen, S., Liu, Y., Han, W., Zhang, W., & Liu, T. (2024). *A survey on LLM-based multi-agent system: Recent advances and new frontiers in application*. arXiv. https://arxiv.org/abs/2412.17481

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Chan, C.-M., Yu, H., Lu, Y., Hung, Y.-H., Qian, C., Qin, Y., Cong, X., Xie, R., Liu, Z., Sun, M., & Zhou, J. (2024). AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B., Yu, B., Li, Y., Liao, Z., Wei, C., Lu, Z., Dey, V., Xue, M., Baker, F. N., Burns, B., Adu-Ampratwum, D., Huang, X., Ning, X., Gao, S., Su, Y., & Sun, H. (2025). ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

D'Arcy, M., Hope, T., Birnbaum, L., & Downey, D. (2024). *MARG: Multi-agent review generation for scientific papers*. arXiv. https://arxiv.org/abs/2401.04259

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235). PMLR.

Ghafarollahi, A., & Buehler, M. J. (2025). SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials, 37*(22), Article 2413523. https://doi.org/10.1002/adma.202413523

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., . . . Natarajan, V. (2025). *Towards an AI co-scientist*. arXiv. https://arxiv.org/abs/2502.18864

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (pp. 8048–8057). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2024/890

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., & Schmidhuber, J. (2024). MetaGPT: Meta programming for a multi-agent collaborative framework. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Hu, S., Lu, C., & Clune, J. (2025). Automated design of agentic systems. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Huang, Q., Vora, J., Liang, P., & Leskovec, J. (2024). MLAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235). PMLR.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., &

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. R. (2024). SWE-bench: Can language models resolve real-world GitHub issues? In *Proceedings of the Twelfth International Conference on Learning Representations*.

Jin, Y., Zhao, Q., Wang, Y., Chen, H., Zhu, K., Xiao, Y., & Wang, J. (2024). AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). *AI agents that matter*. arXiv. https://arxiv.org/abs/2407.01502

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023). CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems* (Vol. 36).

Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. *Vicinagearth, 1*, Article 9. https://doi.org/10.1007/s44336-024-00009-2

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., & Tu, Z. (2024). Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., McFarland, D. A., & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI, 1*(8). https://doi.org/10.1056/AIoa2400196

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 12*, 157–173. https://doi.org/10.1162/tacl_a_00638

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://arxiv.org/abs/2408.06292

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

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. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (Article 2). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., & Sun, M. (2024). ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1). Association for

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Computational Linguistics.

Reddy, C. K., & Shojaee, P. (2025). *Towards scientific discovery with generative AI: Progress, opportunities, and challenges*. arXiv. https://arxiv.org/abs/2412.11427

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z., & Barsoum, E. (2025). *Agent Laboratory: Using LLM agents as research assistants*. arXiv. https://arxiv.org/abs/2501.04227

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems* (Vol. 36).

Si, C., Yang, D., & Hashimoto, T. (2025). Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Siegel, Z. S., Kapoor, S., Nadgir, N., Stroebl, B., & Narayanan, A. (2024). CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*.

Skarlinski, M. D., Cox, S., Laurent, J. M., Braza, J. D., Hinks, M., Hammerling, M. J., Ponnapati, M., Rodriques, S. G., & White, A. D. (2024). *Language agents achieve superhuman synthesis of scientific knowledge*. arXiv. https://arxiv.org/abs/2409.13740

Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S., Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson, W., Heidecke, J., Glaese, A., & Patwardhan, T. (2025). PaperBench: Evaluating AI's ability to replicate AI research. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR.

Su, H., Chen, R., Tang, S., Yin, Z., Zheng, X., Li, J., Qi, B., Wu, Q., Li, H., Ouyang, W., Torr, P., Zhou, B., & Dong, N. (2025). Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E., & Zou, J. (2025). The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*. https://doi.org/10.1038/s41586-025-09442-9

Tran, K.-T., Dao, D., Nguyen, M.-D., Pham, Q.-V., O'Sullivan, B., & Nguyen, H. D. (2025). *Multi-agent collaboration mechanisms: A survey of LLMs*. arXiv. https://arxiv.org/abs/2501.06322

Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., & Zou, J. (2024). *Mixture-of-Agents enhances large language model capabilities*. arXiv. https://arxiv.org/abs/2406.04692

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), Article 186345. https://doi.org/10.1007/s11704-024-40231-1

Wang, Q., Downey, D., Ji, H., & Hope, T. (2024). SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of*

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

the Association for Computational Linguistics (Vol. 1). Association for Computational Linguistics.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2024). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Proceedings of the First Conference on Language Modeling*.

Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., & Ha, D. (2025). *The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search*. arXiv. https://arxiv.org/abs/2504.08066

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the Eleventh International Conference on Learning Representations*.

Zhang, J., Xiang, J., Yu, Z., Teng, F., Chen, X., Chen, J., Zhuge, M., Cheng, X., Hong, S., Wang, J., Zheng, B., Liu, B., Luo, Y., & Wu, C. (2025). AFlow: Automating agentic workflow generation. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems* (Vol. 36).

Zheng, T., Deng, Z., Tsang, H. T., Wang, W., Bai, J., Wang, Z., & Song, Y. (2025). *From automation to autonomy: A survey on large language models in scientific discovery*. arXiv. https://arxiv.org/abs/2505.13259

Zhuge, M., Zhao, C., Ashley, D. R., Wang, W., Khizbullin, D., Xiong, Y., Liu, Z., Chang, E., Krishnamoorthi, R., Tian, Y., Shi, Y., Chandra, V., & Schmidhuber, J. (2025). Agent-as-a-judge: Evaluate agents with agents. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR.