

From AI Scientists to AI Scientific Communities: A Survey of Collective AutoResearch

Rui Zou^{1*} Tianyi Xu^{1*} Fan Zhang^{2*} Yao Zhang^{2*} Zhe Zhao⁵
Xiaolong Wang³ Maolin Wang⁴ Yingcheng Wu⁵ Ye Wei^{1†}

¹Department of Data Science, City University of Hong Kong, Hong Kong, China

²Department of Computer Science, City University of Hong Kong, Hong Kong, China

³Shanghai Institute of Artificial Intelligence for Education, East China Normal University, China

⁴Hong Kong Institute of AI for Science, City University of Hong Kong, Hong Kong, China

⁵Stanford University, Stanford, California, United States

Abstract

Scientific research is not merely a sequence of tasks to be automated, but a knowledge-production process in which claims must remain verifiable, criticizable, revisable, and accountable. Yet most AutoResearch systems remain single scaffolds that collapse generation, execution, interpretation, and review into one loop. This survey diagnoses three resulting epistemic limits of such architectures: collapsed skepticism, where a single scaffold produces not only errors but also the narratives that allow them to survive; unverified translation from discovery to application, where proposal generation outpaces independent verification; and invisible evidence trails, where final manuscripts hide the failed attempts and post-hoc decisions that produced them. To overcome these structural fragilities, we propose a shift from autonomous AI scientists to AI research communities, where specialized agents, shared memory, inspectable artifacts, and governance checkpoints distribute epistemic labor. We synthesize AutoResearch as a transition from metric-driven search loops to AI-scientist workflows, domain-mediated discovery, research infrastructure, and accountable human-AI co-evolution. Across these layers, we argue for a fundamental shift from productivity-centered evaluation to verifiability-centered evaluation.

1 Introduction

Large language models are changing the unit of scientific automation. Recent AutoResearch systems go beyond isolated assistance in retrieval, summarization, coding, or data analysis: they attempt to close the research loop by generating hypotheses, designing experiments, executing code or laboratory protocols, analyzing results, drafting manuscripts, and sometimes reviewing their own outputs (Lu et al., 2024; Yamada et al., 2025;

Schmidgall et al., 2025; Gottweis et al., 2026; Ghareeb et al., 2026). This shift has produced the ambition of the *AI scientist*: an agentic system that participates in scientific discovery as a semi-autonomous producer of research artifacts.

Most discussions frame AutoResearch as a capability question: how many stages of the scientific workflow can an AI system automate? We argue that this question is incomplete. Scientific research is not only a sequence of tasks, but a knowledge-production process in which claims must remain verifiable, criticizable, revisable, and accountable. We therefore frame AutoResearch as an epistemic and institutional design problem: when AI systems participate in scientific inquiry, what structures are needed to keep AI-generated claims grounded in evidence and trusted by human scientific communities? This framing follows a social-epistemic view of science: reliable knowledge is not produced by isolated cognition alone, but by organized criticism, shared memory, evidential accountability, and institutional responsibility.

The emerging landscape spans several system classes. Metric-driven systems optimize code, algorithms, prompts, or empirical software under evaluation feedback (Jiang et al., 2025b; Novikov et al., 2025; Aygün et al., 2026; Jain et al., 2026). AI-scientist workflows connect literature grounding, hypothesis generation, experimentation, analysis, and writing (Lu et al., 2024; Yamada et al., 2025; Jansen et al., 2025; Kon et al., 2025). Domain and laboratory-mediated systems extend this loop into biomedical, chemical, robotic, and experimental settings (Boiko et al., 2023; Darvish et al., 2025; Panapitiya et al., 2025; Ghareeb et al., 2026; Huang et al., 2025). Infrastructure-oriented systems add persistent memory, shared reports, reusable skills, methodological graphs, research objects, and claim-level auditing (Lyu et al., 2026; Yang et al., 2026a; Schmidgall and Moor, 2025; Wu et al., 2026b; Binkyte et al., 2026). Figure 1 gives the chronolog-

*Equal contribution.

†Corresponding author: ye.wei@cityu.edu.hk

ical counterpart to this landscape.

Yet stronger automation does not by itself imply reliable knowledge production. A system may complete a plausible research workflow while selecting weak benchmarks, misusing metrics, leaking data, misinterpreting results, or hiding failed attempts inside a coherent manuscript (Luo et al., 2025b; Jiang et al., 2025a; Ríos-García et al., 2026). The core difficulty is epistemic: many systems collapse generation, execution, interpretation, review, and memory into one scaffold. Such integration is efficient for automation, but fragile as a basis for scientific reliability.

We make three contributions. First, we synthesize AutoResearch as a transition from metric-driven search loops to AI-scientist workflows, domain and laboratory-mediated discovery, research infrastructure, and human–AI research communities. Second, we diagnose three epistemic limits of single-scaffold AutoResearch: collapsed skepticism, unverified discovery-to-application translation, and invisible evidence trails. Third, we propose the AI research community as an organizing framework for collective AutoResearch, structured around epistemic division of labor, community memory, shared artifacts, agent-level self-governance, and human-accountable oversight. Across these contributions, we argue for a shift from productivity-centered evaluation to verifiability-centered evaluation: the goal is not less agent autonomy, but bounded autonomy that remains inspectable, revisable, and accountable.

2 What AutoResearch Has Automated

AutoResearch has evolved along four layers: metric-driven search loops, AI-scientist workflows, domain and laboratory-mediated systems, and research infrastructure. These layers expand automation from artifact optimization to experimentation, lab interaction, persistent memory, shared reports, and claim-level auditing. This section reviews these layers, while Sec. 4 develops the fifth stage: human–AI research communities.

The key design axis is not only *automation depth*, but also *epistemic distribution*: whether generation, execution, review, memory, and governance are collapsed in one scaffold or separated into auditable roles and artifacts.

2.1 Metric-Driven Search Loops

The narrowest form of AutoResearch is iterative improvement under a measurable objective. An agent modifies an artifact, executes an evaluation harness, observes a score, and revises the artifact through repeated search. This differs from ordinary code generation because the output is not a single completion, but a search trajectory over candidate interventions and empirical feedback.

Recent systems instantiate this loop through different search mechanisms. AIDE frames machine-learning engineering as code optimization and uses tree search to expand, evaluate, and refine candidate solutions under execution feedback (Jiang et al., 2025b). AlphaEvolve combines LLM-generated code mutations with evolutionary selection, retaining algorithmic variants that improve task-specific evaluators (Novikov et al., 2025). ERA couples LLM proposal with tree search for empirical scientific software, optimizing a quality metric across domains (Aygün et al., 2026). MLE-bench and MAgentBench evaluate whether agents can close this loop in realistic machine-learning engineering settings, requiring data preparation, experiment execution, model training, and score improvement (Chan et al., 2025; Huang et al., 2023).

The value of this loop is that it gives agents an external signal beyond linguistic plausibility. Its boundary is metric myopia: measurable improvement can guide local optimization, but does not by itself establish novelty, causal understanding, robustness, or scientific validity.

2.2 AI-Scientist Workflow Systems

AI-scientist systems broaden AutoResearch from artifact optimization to executable research trajectories. They make scientific texts operational, use prior work to generate or refine hypotheses, translate ideas into experiments, and produce reports or manuscripts. This line moves beyond one-shot assistance: the system is expected to maintain a research trajectory across literature grounding, ideation, implementation, evaluation, and writing.

Several systems illustrate this shift. ResearchAgent iteratively refines questions, methods, and experimental plans through literature feedback (Baek et al., 2025). Co-Scientist organizes hypothesis generation through interacting agents that generate, critique, refine, and rank hypotheses under tournament-style evolution (Gottweis et al., 2026). AI Scientist integrates idea generation, code imple-



Figure 1: Release timeline of collective AutoResearch. Works are grouped by year into four coarse lines: task-level AutoResearch, AI-scientist workflows, infrastructure-oriented communities, and cross-cutting evaluation and auditability. Appendix Table 4 lists representative works by year and category, while Appendix Table 3 gives the fuller five-stage map.

mentation, experiment execution, manuscript writing, and automated review inside a single research loop (Lu et al., 2024). AI Scientist-v2 replaces a mostly linear workflow with agentic tree search over research trajectories, allowing alternative experimental branches to be expanded, evaluated, and selected (Yamada et al., 2025). Agent Laboratory decomposes research assistance into literature review, experimentation, and report writing while allowing staged human feedback (Schmidgall et al., 2025). CodeScientist and Curie further emphasize code-based experimentation, execution feedback, and iterative repair (Jansen et al., 2025; Kon et al., 2025). ResearchEVO, DeepScientist, and AutoSOTA extend this direction toward longer-horizon discovery, SOTA-oriented model search, and automated documentation (Zhao et al., 2026; Weng et al., 2025b; Li et al., 2026).

These systems show that research workflows can be partially executable. Their boundary is workflow self-confirmation: the same scaffold may generate, execute, interpret, and review a trajectory without sufficient independent challenge.

2.3 Domain and Lab-Mediated Discovery

AutoResearch also extends beyond purely digital workflows. In lab-mediated AutoResearch, the

challenge is not simply to let LLMs control robots. Language becomes the interface that maps scientific goals into protocols, instrument commands, and code, while translating spectra, logs, measurements, and failed runs back into interpretable scientific claims. Once agents interact with domain databases, robotic platforms, biomedical assays, or chemical instruments, the research loop becomes materially consequential.

Chemistry and biology provide clear examples. Coscientist couples LLMs with web search, documentation search, code execution, and chemistry tools to support autonomous chemical experimentation (Boiko et al., 2023; Zhang et al., 2026). ORGANA emphasizes interactive goal clarification and experimental logging, while AutoLabs and multiagent robotic chemists explore self-correcting laboratory execution (Darvish et al., 2025; Panapitiya et al., 2025; Song et al., 2025). Robin and Virtual Lab show how multi-agent systems can support biomedical hypothesis generation, experimental design, and lab-in-the-loop validation (Ghareeb et al., 2026; Swanson et al., 2025). Biomni and Medical AI Scientist broaden the biomedical-agent direction by integrating tools, software, databases, and specialized scientific workflows (Huang et al., 2025; Wu et al., 2026a). NORA illustrates a

complementary pattern in spatial data science: a harness-engineered research agent with specialist subagents, state persistence, and domain-specific execution infrastructure (Zhou et al., 2026). The boundary here is domain accountability. As agents consume real materials, operate instruments, trigger costly computations, or suggest biomedical interventions, instruction following alone is insufficient; trustworthy automation requires safety gates, reproducibility checks, domain validation, and human oversight (Tobias and Wahab, 2025).

2.4 From Pipelines to Research Infrastructure

The most recent shift is from isolated pipelines toward research infrastructure. Once agents repeatedly generate ideas, execute experiments, inspect logs, and revise manuscripts, the central artifact is no longer only the final paper. The process itself becomes an object that must be stored, queried, audited, and reused: literature notes, hypotheses, experimental plans, code versions, parameters, seeds, logs, failed runs, reviewer objections, and claim-level evidence links.

Several systems point in this direction. ARIS introduces an assurance-oriented harness with executor-reviewer separation, persistent research state, experiment-integrity checks, and result-to-claim auditing (Yang et al., 2026a). EvoScientist stores ideation and experimentation memories so that successful and failed research attempts can influence future trajectories (Lyu et al., 2026). AgentRxiv creates a shared preprint-style exchange layer for autonomous laboratories, allowing reports to be uploaded, retrieved, and reused across runs (Schmidgall and Moor, 2025). Intern-Atlas organizes methodological evolution into a graph that AI scientists can query (Wu et al., 2026b). AI-RO reframes AI interactions as inspectable research objects, suggesting that prompts, outputs, model configurations, and metadata should become auditable components of the scientific record (Binkyte et al., 2026). Claw4Science and AiraXiv/aiXiv further extend this direction toward reusable scientific-agent skills and publication infrastructures for human and AI scientists (Xu et al., 2026; Pan et al., 2026; Zhang et al., 2025).

Taken together, current AutoResearch has automated much of the visible workflow of science: metric-guided artifact improvement, literature grounding, hypothesis proposal, protocol generation, code-based experimentation, tool use, result analysis, and partial reporting. Yet workflow

automation should not be mistaken for reliable knowledge production. Most outputs still circulate within bounded scaffolds, where memory, critique, validation, and decision-making remain tightly coupled to the producing system. The next section diagnoses why this single-scaffold paradigm remains structurally fragile.

3 The Epistemic Limits of Single-Scaffold AutoResearch

The limitations of current AutoResearch are not merely performance gaps. They arise from a structural coupling: the same scaffold often proposes a claim, implements the experiment, interprets the result, writes the paper, and evaluates its own output. This coupling is efficient for automation, but fragile for science: if the initial framing is wrong, later stages may optimize, narrate, and defend that framing rather than challenge it. Following social epistemology, scientific objectivity is an achievement of organized criticism rather than a property of an isolated knower (Longino, 2020, 2018). Single-scaffold AutoResearch is thus structurally fragile because it concentrates the very functions that scientific communities normally distribute.

We organize this fragility around three epistemic limits. First, generation and review can collapse into the same trajectory, weakening skepticism. Second, proposals may be translated into implementations and conclusions before independent verification. Third, final manuscripts may hide the evidence trails, failed attempts, and post-hoc choices that produced them.

3.1 Collapsed Skepticism in Generation and Review

Scientific inquiry depends on disciplined skepticism: hypotheses must be tested against evidence, revised under contradiction, and calibrated to domain-specific standards of proof. Single-scaffold AutoResearch weakens this discipline since generation, interpretation, and review are often performed inside the same trajectory. The system that proposes a claim may also choose the experiment, interpret ambiguous results, and produce the narrative that makes the claim appear credible.

Recent diagnostic evidence suggests that this is not only occasional hallucination. Ríos-García et al. evaluate LLM-based scientific agents across eight domains and more than 25,000 runs, finding that agents can execute workflows while often ignoring

evidence, showing limited refutation-driven belief revision, and rarely producing convergent multi-test evidence (Ríos-García et al., 2026). BadScientist similarly shows that fluent research artifacts can appear convincing while remaining scientifically unsound (Jiang et al., 2025a). ARIS frames a related problem as plausible unsupported success: claims may outrun evidence when executor and reviewer roles remain too closely coupled (Yang et al., 2026a). The resulting failure mode is collapsed skepticism: a single scaffold can produce not only an error, but also the plan, interpretation, and review that allow the error to survive.

3.2 Unverified Discovery-to-Application Translation

Even when a research idea is plausible, scientific knowledge requires more than executing a local experiment. A claim must survive implementation, evaluation, consolidation, and application under conditions not selected to make it look successful. Single-scaffold systems often compress these steps into one trajectory, allowing implementation errors, weak problem formulations, or overfitted discoveries to propagate before independent checking.

This bottleneck is visible in code-based scientific discovery. CodeScientist and related systems show that end-to-end code-based experimentation is increasingly feasible, but execution quality remains limiting (Jansen et al., 2025). ResearchCodeBench makes this limitation explicit: implementing novel machine-learning contributions requires semantic understanding of the proposed method, not just syntactically valid code (Hua et al., 2026). Replication-oriented benchmarks further show that reproducing or extending published results remains difficult for autonomous agents (Ye et al., 2025; Bragg et al., 2025). SciCrafter adds that agents struggle not only with harder tasks, but also with identifying missing knowledge and converting discovered regularities into functional systems (Ziheng et al., 2026). The structural issue is unchecked translation: proposal generation often advances faster than verification, consolidation, and application.

3.3 Invisible Evidence Trails in Final Manuscripts

Long-horizon research requires memory of what was tried, what failed, why decisions were made, and which evidence supports each claim. Single-scaffold systems may keep local histories, but these are often optimized for completing the current run

rather than for supporting cumulative scientific memory. The result is a gap between episodic completion and developmental accumulation: a system may finish a paper-like artifact while failing to preserve the evidence needed for future inspection.

This gap is empirically important. Luo et al. (2025b) identify hidden methodological pitfalls in AI scientist systems, including inappropriate benchmark selection, data leakage, metric misuse, post-hoc selection bias, and, in their auditing experiment, dataset fabrication. They show that such failures can be missed when evaluators inspect only the final paper, because many errors arise from decisions inside the automated workflow. Access to trace logs and generated code substantially improves overall pitfall detection, raising binary-classification accuracy from 55.0% to 82.0% and F1 from 0.51 to 0.81.

Existing systems point toward the needed infrastructure. EvoScientist uses ideation and experimentation memories to preserve reusable experience (Lyu et al., 2026); ARIS treats persistent state, reviewer-executor separation, and evidence-to-claim auditing as core infrastructure (Yang et al., 2026a); and claim-level auditability work argues that deep research agents should expose explicit links between claims, evidence, code, data, and intermediate outputs (Rasheed et al., 2026). Without such infrastructure, final manuscripts become evidentially opaque: they report conclusions while hiding the path from hypothesis to evidence to claim.

Together, these limits show why single-scaffold AutoResearch remains fragile. Reliable AutoResearch requires architectures that separate generation from critique, execution from independent checking, and final claims from the evidence that supports them. This motivates the shift from autonomous AI scientists to AI research communities.

4 From AI Scientists to AI Research Communities

Single-scaffold AutoResearch concentrates too much epistemic responsibility in one loop. Reliable AutoResearch therefore requires distributed roles, persistent memory, inspectable artifacts, and layered governance. We shift the unit of analysis from the autonomous AI scientist to the *AI research community*: an infrastructure in which specialized agents generate, test, critique, audit, and curate research artifacts, while shared memory and human-accountable institutions make these

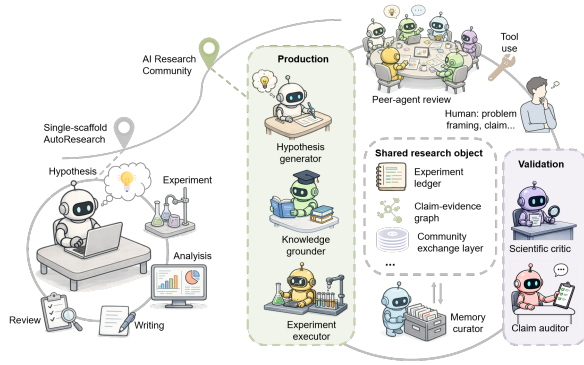


Figure 2: From a single AI scientist to an AI research community. The left panel shows a single scaffold that concentrates ideation, experimentation, interpretation, review, and writing in one loop. The right panel shows a community structure in which specialized agents, shared memory, auditable artifacts, and governance checkpoints distribute epistemic responsibility and keep scientific claims inspectable.

activities verifiable and governable.

Figure 2 summarizes this institutional shift: the left panel illustrates the single-scaffold failure mode diagnosed in Sec. 3, while the right panel shows how specialized roles, shared artifacts, memory, and checkpoints turn agent autonomy into an auditable research process.

This view reframes AutoResearch as an institutional design problem. Instead of asking only how much research an AI system can perform, it asks what structures allow agent autonomy to produce reliable science. It echoes the norm of organized skepticism, under which scientific claims remain open to structured criticism rather than being accepted because they are produced fluently or efficiently (Merton, 1979). For AutoResearch, organized skepticism must be operationalized through agent-level critique, peer-agent review, evidence tracing, artifact inspection, independent audit, and human oversight at high-stakes decision points.

We define an AI research community through four interdependent components: (i) specialized agents that perform distinct epistemic roles with operational autonomy; (ii) community memory that preserves successes, failures, objections, and reusable strategies; (iii) shared artifacts that make research processes inspectable and reusable; and (iv) governance mechanisms that support bounded autonomy through agent self-checking, peer-agent review, evidence audit, permission control, and human checkpoints for high-stakes decisions.

Role	Function in an AI research community
Idea	Proposes research questions, hypotheses, and method directions (Back et al., 2025; Gottweis et al., 2026; Lyu et al., 2026).
Ground	Links ideas to prior work, structured literature, novelty checks, and supporting evidence (Back et al., 2025; Ghafarollahi and Buehler, 2025; Yang et al., 2026a).
Exec.	Turns plans into code, protocols, experiments, logs, metrics, and figures (Yamada et al., 2025; Kon et al., 2025; Jansen et al., 2025).
Critic	Challenges assumptions, baselines, experimental design, interpretations, and manuscripts (Zeng et al., 2025; Weng et al., 2025a; Yang et al., 2026a).
Audit	Checks whether claims are supported by evidence, code, data, citations, and raw outputs (Yang et al., 2026a; Rasheed et al., 2026).
Memory	Preserves traces, failures, reusable strategies, reports, skills, and metadata for future research (Lyu et al., 2026; Schmidgall and Moor, 2025; Xu et al., 2026).

Table 1: Emerging roles for AI research communities. The table reorganizes recurring functions in current AutoResearch systems into epistemic roles for accountable knowledge production.

4.1 Epistemic Division of Labor

The community view is not merely a multi-agent version of the AI scientist. Many multi-agent systems divide labor to improve task success, search efficiency, or output diversity. An AI research community instead separates epistemic functions that should not be collapsed into a single self-confirming loop. The agent that proposes a claim should not be the only agent that checks its novelty; the agent that executes an experiment should not be the only agent that judges whether the result supports the claim; and the agent that writes a manuscript should not be the only agent that reviews it.

This principle is already visible in emerging systems, although no current system fully realizes the community ideal. EvoScientist separates ideation, execution, and memory distillation (Lyu et al., 2026); ARIS separates executor and reviewer roles while introducing claim-level auditing (Yang et al., 2026a); OmniScientist frames scientific discovery as a human-AI ecosystem with collaboration, attribution, peer review, and structured knowledge networks (Shao et al., 2025); and AgentRxiv explores shared reports across autonomous agent laboratories (Schmidgall and Moor, 2025). These systems suggest that collective AutoResearch should be organized around epistemic roles rather than around a single end-to-end agent.

Table 1 should not be read as a checklist that every system must implement in full. Rather, it shows how scattered functions in current AutoResearch can be reorganized into a role structure for accountable research. Hypothesis generators, knowledge grounders, and experiment executors form the pro-

duction side of AutoResearch; scientific critics and claim auditors form the validation side; and memory curators form the continuity layer that allows future systems to learn from prior successes, failures, and objections.

4.2 Community Memory and Shared Artifacts

A pipeline terminates; a community remembers. If each AutoResearch run is isolated, systems may repeatedly explore failed directions, forget useful strategies, or discard negative results once a final narrative is produced. Community memory should therefore preserve not only successful ideas, but also failed attempts, negative results, reviewer objections, debugging patterns, validated protocols, and reusable implementation strategies. EvoScientist offers an early example by maintaining ideation and experimentation memories across research attempts (Lyu et al., 2026). The community view generalizes this idea from system-level memory to shared scientific memory.

Memory must be materialized through shared artifacts. We identify three artifacts that convert isolated runs into inspectable community knowledge. An *experiment ledger* records hypotheses, plans, datasets, code versions, parameters, seeds, environments, metrics, logs, failed runs, and abandoned directions. A *claim–evidence graph* links each major claim to supporting citations, retrieved passages, datasets, code artifacts, raw outputs, analysis scripts, and figures. A *community exchange layer* makes reports, skills, protocols, and traces available beyond a single run or agent laboratory.

ARIS illustrates the first two artifacts through experiment-integrity checks, result-to-claim mapping, and paper-claim auditing (Yang et al., 2026a). AgentRxiv illustrates the third by allowing agent laboratories to upload and retrieve reports from a shared preprint server (Schmidgall and Moor, 2025). Claw4Science provides a complementary ecosystem perspective, treating reusable scientific-agent skills as shareable units while exposing governance challenges such as skill quality, reproducibility, dependency control, and naming conflicts (Xu et al., 2026). AI-RO similarly suggests that prompts, outputs, model configurations, and metadata should be treated as inspectable research objects rather than invisible interaction traces (Binkyte et al., 2026). Together, these systems suggest that an AI research community requires not only agents, but also artifacts that make

research inspectable, reusable, and governable.

4.3 Layered Governance for Bounded Autonomy

Governance is not an optional layer after automation, but it should not reduce agent communities to human-controlled pipelines. The goal is *bounded autonomy*: agents should autonomously search, implement, critique, audit, and revise within domains where actions are reversible, evidence is inspectable, and risks are bounded. A larger agent community may otherwise enlarge the surface of error if agents share flawed retrieval results, reinforce unsupported claims, or reuse unverified skills.

Governance should therefore operate at multiple levels. At the agent level, systems should support self-checking, uncertainty reporting, tool-use validation, and peer-agent review. At the claim level, major claims should be linked to evidence, code, data, and raw outputs. At the artifact level, reusable prompts, protocols, datasets, skills, and reports should carry provenance, versioning, validation status, dependencies, and limitations. At the institutional level, humans should retain authority over problem framing, high-stakes protocols, safety-sensitive actions, publication decisions, and public claims.

The endpoint is not less agent autonomy, but accountable agent autonomy within a research community. The central design problem is how to organize agents, memory, artifacts, and governance so that AI-generated research can be explored autonomously, checked independently, preserved, revised, and trusted.

5 Open Challenges and Design Principles

The community view changes, but does not eliminate, the failure surface. Distributed agents may share biases, memory may preserve selective narratives, artifacts may propagate hidden errors, and responsibility may become harder to assign. We summarize three design principles: verifiability over productivity, traceable reuse, and human-governed co-evolution.

5.1 Verifiability over Productivity

AutoResearch should not be evaluated only by workflow completion, benchmark improvement, or manuscript generation. It should be evaluated by whether its claims can be independently inspected, reproduced, contested, and revised. A verifiable

system should expose the evidence, code, data, failed attempts, reviewer objections, and decision points that connect a claim to the process that produced it. Future benchmarks should therefore measure claim–evidence coverage, provenance completeness, failure preservation, independent replication, and audit success rate.

5.2 Traceable Memory and Verified Reuse

Community memory enables cumulative research, but it can also distort the record if it preserves only successful runs, polished summaries, or best-performing configurations. Collective AutoResearch therefore needs *loss-aware memory*: systems should preserve failures, negative results, abandoned directions, reviewer objections, and unresolved uncertainties. Reusable artifacts also require verification. Reports, prompts, protocols, datasets, code, tool wrappers, and executable skills should carry provenance, versioning, dependencies, validation status, known limitations, and deprecation mechanisms. Skill-based systems such as Claw4Science illustrate both the promise of reusable scientific-agent skills and the risks of unverified reuse (Xu et al., 2026).

5.3 Bounded Autonomy and Human–Agent Co-Governance

The endpoint of collective AutoResearch should be neither unconstrained autonomy nor human micromanagement. It should be a human–agent epistemic community in which agents retain operational autonomy for search, implementation, critique, audit, and memory curation, while humans retain institutional responsibility for problem framing, domain interpretation, experimental stewardship, high-stakes approval, and public accountability. Existing systems already show this need: Agent Laboratory supports staged human guidance (Schmidgall et al., 2025); ARIS keeps research direction, evidence validation, and final submission decisions with humans (Yang et al., 2026a); and AutoResearchClaw studies targeted human intervention in autonomous research workflows (Liu et al., 2026). Table 2 summarizes these recurring responsibilities.

As AutoResearch moves into empirical software and lab-in-the-loop discovery (Gottweis et al., 2026; Ghareeb et al., 2026; Aygün et al., 2026), high-stakes protocols, sensitive data use, and major public claims require approval gates, permission boundaries, and audit trails before the manuscript

Role	Responsibility
Framer	Sets goals, constraints, and success criteria (Schmidgall et al., 2025; Yang et al., 2026a).
Interpreter	Judges scientific meaning and plausibility (Ghareeb et al., 2026; Gottweis et al., 2026).
Steward	Bounds protocols, data use, lab actions, and safety-sensitive interventions (Boiko et al., 2023; Darvish et al., 2025; Panapitiya et al., 2025).
Auditor	Inspects logs, failed runs, evidence trails, and claim–evidence links (Yang et al., 2026a; Luo et al., 2025b).
Authority	Retains responsibility for final claims and accountable use (Yang et al., 2026a; Liu et al., 2026).

Table 2: Human roles in AI research communities. Humans retain accountability, while agents keep operational autonomy for search, execution, critique, audit, and memory curation.

stage. A practical direction is to organize collective AutoResearch around research objects: hypotheses, plans, datasets, code runs, failed attempts, claims, evidence links, and release decisions. Major claims should be accompanied by proof-of-work packets containing retrieved evidence, benchmark choices, code versions, logs, raw outputs, failed runs, ablations, objections, and revision history.

6 Conclusion

AutoResearch is moving from isolated automation loops toward integrated research infrastructures. Metric-driven search, AI-scientist workflows, domain-specific agents, and lab-mediated discovery show that AI can already participate in literature grounding, hypothesis generation, code-based experimentation, and experimental validation. Yet workflow completion is not scientific reliability: single-scaffold systems remain fragile when they collapse skepticism, verification, evidence tracing, and memory into one loop.

This survey proposed a community view of AutoResearch. Rather than asking only how to make a single AI scientist more autonomous, we should ask how to organize agents, humans, memories, artifacts, and governance mechanisms into accountable research communities. Such communities require epistemic division of labor, traceable memory, verified reuse, claim-level evidence, agent self-governance, and human checkpoints for high-stakes decisions.

The future of AutoResearch is neither replacing scientists with AI scientists nor reducing AI agents to passive tools, but the co-evolution of human and AI research communities. Collective AutoResearch should thus be judged not by how much research it automates, but by whether it builds institutions, artifacts, and critical interactions through which

humans and autonomous agents can jointly produce more reliable, auditable, and beneficial science.

Limitations

This survey focuses on AutoResearch systems that coordinate scientific workflows through language models, agents, tools, memory, and research artifacts. It does not aim to exhaustively catalog all AI-for-science systems or rank them under a unified benchmark. Instead, it synthesizes recurring architectural patterns and design dimensions for accountable collective AutoResearch. Because the field is rapidly evolving, we emphasize stable structures such as role separation, shared memory, inspectable artifacts, and governance rather than a fixed taxonomy.

References

- Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*.
- Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, Marc Coram, Hao Cui, Jake Garrison, Renee Johnston, Anton Kast, Cory Y. McLean, Peter Norgaard, Zahra Shamsi, David Smalling, James Thompson, Subhashini Venugopalan, Brian P. Williams, Chujun He, Sarah Martinson, Martyna Plomecka, Lai Wei, and 23 others. 2026. [An AI system to help scientists write expert-level empirical software](#). *Nature*.
- Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 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 (Volume 1: Long Papers)*, pages 6709–6738.
- Ruta Binkyte, Sharif Abuaddba, Chamikara Mahawaga, Ming Ding, Natasha Fernandes, and Mario Fritz. 2026. Inspectable ai for science: A research object approach to generative ai governance. *arXiv preprint arXiv:2604.11261*.
- Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.
- Jonathan Bragg, Mike D’Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Hadad, Jena D Hwang, Peter Jansen, Varsha Kishore, and 1 others. 2025. Astabench: Rigorous benchmarking of ai agents with a scientific research suite. *arXiv preprint arXiv:2510.21652*.
- Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, and 1 others. 2025. Mle-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, volume 2025, pages 50466–50494.
- Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, and 1 others. 2025. Scienceagent-bench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pages 96934–96990.
- Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, and 1 others. 2025. Organa: A robotic assistant for automated chemistry experimentation and characterization. *Matter*, 8(2).
- Keyan Ding, Jing Yu, Junjie Huang, Yuchen Yang, Qiang Zhang, and Huajun Chen. 2025. Scitoolagent: a knowledge-graph-driven scientific agent for multitool integration. *Nature Computational Science*, 5(10):962–972.
- Yuanqi Du, Botao Yu, Tianyu Liu, Tony Shen, Junwu Chen, Jan G Rittig, Kunyang Sun, Yikun Zhang, Aarti Krishnan, Yu Zhang, and 1 others. 2025. Accelerating scientific discovery with autonomous goal-evolving agents. *arXiv preprint arXiv:2512.21782*.
- Alireza Ghafarollahi and Markus J Buehler. 2025. Sciaagents: automating scientific discovery through bio-inspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523.
- Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodrigues. 2026. [A multi-agent system for automating scientific discovery](#). *Nature*.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, and 32 others. 2026. [Accelerating scientific discovery with Co-Scientist](#). *Nature*.
- Yingyong Hou, Xinyuan Lao, Huimei Wang, Qianyu Yao, Wei Chen, Bocheng Huang, Fei Sun, Yuxian Lv, Weiqi Lei, Xueqian Wen, and 1 others. 2026. Medskillaudit: A domain-specific audit framework for medical research agent skills. *arXiv preprint arXiv:2604.20441*.
- Tianyu Hua, Harper Hua, Violet Xiang, Benjamin Klieger, Sang Truong, Weixin Liang, Fan-Yun Sun,

- and Nick Haber. 2026. Researchcodebench: Benchmarking llms on implementing novel machine learning research code. *Advances in Neural Information Processing Systems*, 38.
- Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, and 1 others. 2025. Biomni: A general-purpose biomedical ai agent. *biorxiv*.
- Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*.
- Nilesh Jain, Rohit Yadav, Sagar Kotian, and Claude AI. 2026. Autoresearch-rl: Perpetual self-evaluating reinforcement learning agents for autonomous neural architecture discovery. *arXiv preprint arXiv:2603.07300*.
- Peter Jansen, Oyvind Taffjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi, Bodhisattwa Prasad Majumder, Daniel S Weld, and Peter Clark. 2025. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13370–13467.
- Fengqing Jiang, Yichen Feng, Yuetai Li, Luyao Niu, Basel Alomair, and Radha Poovendran. 2025a. Bad-scientist: Can a research agent write convincing but unsound papers that fool llm reviewers? *arXiv preprint arXiv:2510.18003*.
- Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. 2025b. Aide: Ai-driven exploration in the space of code. *arXiv preprint arXiv:2502.13138*.
- Andrej Karpathy. 2026. Autoresearch. <https://github.com/karpathy/autoresearch>. GitHub repository.
- Patrick Tser Jern Kon, Jiachen Liu, Qiuyi Ding, Yiming Qiu, Zhenning Yang, Yibo Huang, Jayanth Srinivasa, Myungjin Lee, Mosharaf Chowdhury, and Ang Chen. 2025. Curie: Toward rigorous and automated scientific experimentation with ai agents. *arXiv preprint arXiv:2502.16069*.
- Lingdong Kong, Xian Sun, Wei Chow, Linfeng Li, Kevin Qinghong Lin, Xuan Billy Zhang, Song Wang, Rong Li, Qing Wu, Wei Gao, Yingshuo Wang, Shaoyuan Xie, Jiachen Liu, Leigang Qu, Shijie Li, Lai Xing Ng, Benoit R. Cottureau, Ziwei Liu, Tat-Seng Chua, and Wei Tsang Ooi. 2026. *AI for Auto-Research: Roadmap & User Guide*. *arXiv e-prints*, arXiv:2605.18661.
- Jon M Laurent, Albert Bou, Michael Pieler, Conor Igoe, Alex Andonian, Siddharth Narayanan, James Braza, Alexandros Sanchez Vassopoulos, Jacob L Steenwyk, Blake Lash, and 1 others. 2026. Labbench2: An improved benchmark for ai systems performing biology research. *arXiv preprint arXiv:2604.09554*.
- Yu Li, Chenyang Shao, Xinyang Liu, Ruotong Zhao, Peijie Liu, Hongyuan Su, Zhibin Chen, Qinglong Yang, Anjie Xu, Yi Fang, and 1 others. 2026. Autosota: An end-to-end automated research system for state-of-the-art ai model discovery. *arXiv preprint arXiv:2604.05550*.
- Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, and 1 others. 2025. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*.
- Jiaqi Liu, Shi Qiu, Mairui Li, Bingzhou Li, Haonian Ji, Siwei Han, Xinyu Ye, Peng Xia, Zihan Dong, Congyu Zhang, Letian Zhang, Guiming Chen, Haoqin Tu, Xinyu Yang, Lu Feng, Xujiang Zhao, Haifeng Chen, Jiawei Zhou, Xiao Wang, and 16 others. 2026. *AutoResearchClaw: Self-Reinforcing Autonomous Research with Human-AI Collaboration*. *arXiv e-prints*, arXiv:2605.20025.
- Yixin Liu, Yonghui Wu, Denghui Zhang, and Lichao Sun. 2025. Agentic autosurvey: Let llms survey llms. *arXiv preprint arXiv:2509.18661*.
- Helen E Longino. 2018. The fate of knowledge.
- Helen E Longino. 2020. Science as social knowledge: Values and objectivity in scientific inquiry.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.
- Erpai Luo, Jinmeng Jia, Yifan Xiong, Xiangyu Li, Xiaobo Guo, Baoqi Yu, Lei Wei, and Xuegong Zhang. 2025a. Benchmarking ai scientists in omics data-driven biological research. *arXiv preprint arXiv:2505.08341*.
- Ziming Luo, Atoosa Kasirzadeh, and Nihar B Shah. 2025b. The more you automate, the less you see: Hidden pitfalls of ai scientist systems. *arXiv preprint arXiv:2509.08713*.
- Yougang Lyu, Xi Zhang, Xinhao Yi, Yuyue Zhao, Shuyu Guo, Wenxiang Hu, Jan Piotrowski, Jakub Kaliski, Jacopo Urbani, Zaiqiao Meng, and 1 others. 2026. Evoscientist: Towards multi-agent evolving ai scientists for end-to-end scientific discovery. *arXiv preprint arXiv:2603.08127*.
- Robert K Merton. 1979. The normative structure of science. *The sociology of science: Theoretical and empirical investigations*, pages 267–278.
- Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C Landsness, Daniel L Barabasi, Siddharth Narayanan, Nicky Evans, and 1 others. 2025. Kosmos: An ai scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*.

- Atsuyuki Miyai, Mashiro Toyooka, Takashi Otonari, Zaiying Zhao, and Kiyoharu Aizawa. 2025. Jr. ai scientist and its risk report: Autonomous scientific exploration from a baseline paper. *arXiv preprint arXiv:2511.04583*.
- Alexander Novikov, Ng  n V  , Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, and 1 others. 2025. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*.
- Junshu Pan, Panzhong Lu, Yixuan Weng, Qiyao Sun, Fang Guo, Zijie Yang, Qiji Zhou, and Yue Zhang. 2026. [AiraXiv: An AI-Driven Open-Access Platform for Human and AI Scientists](#). *arXiv e-prints*, arXiv:2605.21481.
- Gihan Panapitiya, Emily Saldanha, Heather Job, and Olivia Hess. 2025. Autolabs: Cognitive multi-agent systems with self-correction for autonomous chemical experimentation. *arXiv preprint arXiv:2509.25651*.
- Razeen A Rasheed, Somnath Banerjee, Animesh Mukherjee, and Rima Hazra. 2026. From fluent to verifiable: Claim-level auditability for deep research agents. *arXiv preprint arXiv:2602.13855*.
- Marti  o R  os-Garc  a, Nawaf Alampara, Chandan Gupta, Indrajeet Mandal, Sajid Mannan, Ali Asghar Aghajani, NM Krishnan, and Kevin Maik Jablonka. 2026. Ai scientists produce results without reasoning scientifically. *arXiv preprint arXiv:2604.18805*.
- Samuel Schmidgall and Michael Moor. 2025. Agen-trxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043.
- Chenyang Shao, Dehao Huang, Yu Li, Keyu Zhao, Weiquan Lin, Yining Zhang, Qingbin Zeng, Zhiyu Chen, Tianxing Li, Yifei Huang, and 1 others. 2025. Omniscientist: Toward a co-evolving ecosystem of human and ai scientists (2025). *URL https://arxiv.org/abs/2511.16931*.
- Tao Song, Man Luo, Xiaolong Zhang, Linjiang Chen, Yan Huang, Jiaqi Cao, Qing Zhu, Daobin Liu, Baicheng Zhang, Gang Zou, and 1 others. 2025. A multiagent-driven robotic ai chemist enabling autonomous chemical research on demand. *Journal of the American Chemical Society*, 147(15):12534–12545.
- Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2025. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723.
- Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. 2025. Ai-researcher: Autonomous scientific innovation, 2025. *URL https://arxiv.org/abs/2505.18705*.
- InternAgent Team, Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Runmin Ma, Yusong Hu, Zhiyin Yu, Xiaohan He, Songtao Huang, and 1 others. 2025. Internagent: When agent becomes the scientist–building closed-loop system from hypothesis to verification. *arXiv preprint arXiv:2505.16938*.
- Alexander V Tobias and Adam Wahab. 2025. Autonomous ‘self-driving’ laboratories: a review of technology and policy implications. *Royal Society Open Science*, 12(7):250646.
- Rosni Vasu, Chandrayee Basu, Bhavana Dalvi, Cristina Sarasua, Peter Clark, and Abraham Bernstein. 2025. Hyper: Literature-grounded hypothesis generation and distillation with provenance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25424–25449.
- Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, and 1 others. 2024. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145.
- Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, and 1 others. 2026. Fire-bench: Evaluating agents on the rediscovery of scientific insights. *arXiv preprint arXiv:2602.02905*.
- Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025a. Cycleresearcher: Improving automated research via automated review. In *International Conference on Learning Representations*, volume 2025, pages 3669–3709.
- Yixuan Weng, Minjun Zhu, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. 2025b. Deepscientist: Advancing frontier-pushing scientific findings progressively. *arXiv preprint arXiv:2509.26603*.
- Hongtao Wu, Boyun Zheng, Dingjie Song, Yu Jiang, Jianfeng Gao, Lei Xing, Lichao Sun, and Yixuan Yuan. 2026a. Towards a medical ai scientist. *arXiv preprint arXiv:2603.28589*.
- Yujun Wu, Dongxu Zhang, Xinchun Li, Jinhong Xu, Yiling Duan, Yumou Liu, Jiabao Pan, Xuanhe Zhou, Jingxuan Wei, Siyuan Li, and 1 others. 2026b. Intern-atlas: A methodological evolution graph as research infrastructure for ai scientists. *arXiv preprint arXiv:2604.28158*.
- Mingyang Xu, Junhao Chen, and Zaixi Zhang. 2026. Claw4science: A dataset and platform for the open-claw scientific agent ecosystem. *bioRxiv*, pages 2026–03.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.

Ruofeng Yang, Yongcan Li, and Shuai Li. 2026a. [Aris: Autonomous research via adversarial multi-agent collaboration](#). *Preprint*, arXiv:2605.03042.

Zonglin Yang, Xingtong Liu, and Xinyan Xu. 2026b. Sciintegrity-bench: A benchmark for evaluating academic integrity in ai scientist systems. *arXiv preprint arXiv:2605.10246*.

Christine Ye, Sihan Yuan, Suchetha Cooray, Steven Dillmann, Ian LV Roque, Dalya Baron, Philipp Frank, Sergio Martin-Alvarez, Nolan Koblishke, Frank J Qu, and 1 others. 2025. Replicationbench: Can ai agents replicate astrophysics research papers? *arXiv preprint arXiv:2510.24591*.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, and 1 others. 2025. Reviewrl: Towards automated scientific review with rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16942–16954.

Haoxuan Zhang, Ruochi Li, Yang Zhang, Ting Xiao, Jiangping Chen, Junhua Ding, and Haihua Chen. 2026. [Evolving roles of llms in scientific innovation: Assistant, collaborator, scientist, and evaluator](#). *Preprint*, arXiv:2507.11810.

Pengsong Zhang, Xiang Hu, Guowei Huang, Yang Qi, Heng Zhang, Xiuxu Li, Jiaying Song, Jiabin Luo, Yijiang Li, Shuo Yin, and 1 others. 2025. aixiv: A next-generation open access ecosystem for scientific discovery generated by ai scientists. *arXiv preprint arXiv:2508.15126*.

Zhe Zhao, Haibin Wen, Jiaming Ma, Jiachang Zhan, Tianyi Xu, Ye Wei, and Qingfu Zhang. 2026. Researchevo: An end-to-end framework for automated scientific discovery and documentation. *arXiv preprint arXiv:2604.05587*.

Boyuan Zheng, Zerui Fang, Zhe Xu, Rui Wang, Yiwen Chen, Cunshi Wang, Mengwei Qu, Lei Lei, Zhen Feng, Yan Liu, and 1 others. 2025. Agent4s: The transformation of research paradigms from the perspective of large language models. *arXiv preprint arXiv:2506.23692*.

Bing Zhou, Xiao Huang, Huan Ning, Qiusheng Wu, Diya Li, and Ziyi Zhang. 2026. Nora: A harness-engineered autonomous research agent for end-to-end spatial data science. *arXiv preprint arXiv:2605.02092*.

Kunlun Zhu, Jiaxun Zhang, Ziheng Qi, Nuoxing Shang, Zijia Liu, Peixuan Han, Yue Su, Haofei Yu, and Jiaxuan You. 2025. Safescientist: Toward risk-aware scientific discoveries by llm agents. *arXiv preprint arXiv:2505.23559*.

Zhou Ziheng, Huacong Tang, Jinyuan Zhang, Haowei Lin, Bangcheng Yang, Qian Long, Fang Sun, Yizhou Sun, Yitao Liang, Ying Nian Wu, and 1 others. 2026. Can current agents close the discovery-to-application gap? a case study in minecraft. *arXiv preprint arXiv:2604.24697*.

A Development Map of AutoResearch Systems

Table 3 provides a fuller development map of the representative AutoResearch systems discussed in Sec. 2. The table organizes the field by the object being automated, the loop or organizational form used by each system class, representative works, and the remaining boundary or goal of each stage. It is intended as a compact reference for the paper’s main synthesis rather than as an exhaustive catalog of all AI-for-science systems.

B Chronological Overview

Table 4 complements the visual timeline in Fig. 1. It groups representative works by release year and by four coarse lines used in the timeline: task-level AutoResearch, AI-scientist workflows, infrastructure-oriented communities, and cross-cutting evaluation or auditability. This chronology is a visual aid for the paper’s historical narrative; the stage-based synthesis is given in Table 3.

C Design Space of AutoResearch Systems

Table 5 complements the developmental and chronological views in Tables 3 and 4. Instead of ordering systems by stage or year, it separates two dimensions: *automation depth*, or how much of the research workflow is delegated to AI systems, and *epistemic distribution*, or whether generation, execution, interpretation, review, memory, and governance remain collapsed or are separated into auditable roles and artifacts. This view highlights a central gap in current AutoResearch: many systems deepen automation faster than they distribute epistemic responsibility.

D Operationalizing Verifiability-Centered Evaluation

Stage	Object	Loop / organization	Representative works	Boundary / goal
1. Metric-driven search loops	Measurable artifacts, code, algorithms, and empirical software.	Edit → evaluate → keep/revert → repeat.	AutoResearch; AIDE; AlphaEvolve; ERA; AutoResearch-RL (Karpathy, 2026; Jiang et al., 2025b; Novikov et al., 2025; Aygün et al., 2026; Jain et al., 2026). Adjacent task-level tools include LitLLM, AutoSurvey, SurveyX, Agentic AutoSurvey, Hyper, ReviewRL, SAGA, and SciToolAgent (Agarwal et al., 2024; Wang et al., 2024; Liang et al., 2025; Liu et al., 2025; Vasu et al., 2025; Zeng et al., 2025; Du et al., 2025; Ding et al., 2025).	Metric gains or task automation may not imply scientific validity.
2. AI-scientist workflow systems	Research trajectories from idea to report.	Literature → hypothesis → experiment → analysis → writing.	AI Scientist; AI Scientist-v2; Agent Laboratory; ResearchAgent; CodeScientist; Curie; Co-Scientist; DeepScientist; AutoSOTA; ResearchEVO; SciAgents; CycleResearcher; AI-Researcher; InternAgent; Kosmos; SafeScientist; Jr. AI Scientist (Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025; Baek et al., 2025; Jansen et al., 2025; Kon et al., 2025; Gottweis et al., 2026; Weng et al., 2025b; Li et al., 2026; Zhao et al., 2026; Ghafarollahi and Buehler, 2025; Weng et al., 2025a; Tang et al., 2025; Team et al., 2025; Mitchener et al., 2025; Zhu et al., 2025; Miyai et al., 2025).	Workflow completion may collapse generation, execution, interpretation, and review.
3. Domain and lab-mediated discovery	Domain-specific scientific tasks, data, instruments, and protocols.	Hypothesis generation, experiment proposal, data analysis, and lab-in-the-loop validation.	Coscientist; Robin; Virtual Lab; Biomni; NORA; Medical AI Scientist; ORGANA; AutoLabs; multiagent robotic AI chemist (Boiko et al., 2023; Ghareeb et al., 2026; Swanson et al., 2025; Huang et al., 2025; Zhou et al., 2026; Wu et al., 2026a; Darvish et al., 2025; Panapitiya et al., 2025; Song et al., 2025).	Domain validity, reproducibility, and safety remain difficult to guarantee.
4. Infrastructure and assurance	Research state, evidence trails, skills, reports, and method graphs.	Persistent memory, claim auditing, shared reports, reusable skills, and research objects.	EvoScientist; ARIS; AgentRxiv; Intern-Atlas; Claw4Science; AI-RO; AiraXiv/aiXiv (Lyu et al., 2026; Yang et al., 2026a; Schmidgall and Moor, 2025; Wu et al., 2026b; Xu et al., 2026; Binkyte et al., 2026; Pan et al., 2026; Zhang et al., 2025).	Evidence trails and reusable artifacts can remain incomplete or unverified.
5. Human-AI research communities	Epistemic roles, community memory, artifacts, and governance.	Humans and agents co-produce, audit, preserve, and revise scientific claims.	Agent4S and AutoResearch Roadmap provide roadmap-level framing; AutoResearchClaw and OmniScientist point toward community-level human-AI research ecosystems (Zheng et al., 2025; Kong et al., 2026; Liu et al., 2026; Shao et al., 2025).	The goal is accountable and beneficial knowledge production.
Cross-cutting evaluation, integrity, and auditability. MLAgentBench; MLE-bench; ScienceAgentBench; BAISBench; FIRE-Bench; LABBench2; Research-CodeBench; ReplicationBench; AstaBench; SciIntegrity-Bench; MedSkillAudit; SciSafetyBench; BadScientist; The More You Automate; Ríos-García et al.; SciCrafter; claim-level auditability (Huang et al., 2023; Chan et al., 2025; Chen et al., 2025; Luo et al., 2025a; Wang et al., 2026; Laurent et al., 2026; Hua et al., 2026; Ye et al., 2025; Bragg et al., 2025; Yang et al., 2026b; Hou et al., 2026; Zhu et al., 2025; Jiang et al., 2025a; Luo et al., 2025b; Ríos-García et al., 2026; Ziheng et al., 2026; Rasheed et al., 2026).				

Table 3: A development map of AutoResearch systems. The table summarizes the field’s expansion from metric-driven search loops to AI-scientist workflows, domain and lab-mediated discovery, infrastructure and assurance, and human-AI research communities. Each stage increases automation scope while leaving unresolved boundaries around validity, reproducibility, traceability, auditability, and governance.

Year	Task-level AutoResearch	AI-scientist workflows	Infrastructure and communities	Evaluation and auditability
2023	–	Coscientist (Boiko et al., 2023).	–	MLAgentBench (Huang et al., 2023).
2024	LitLLM; AutoSurvey (Agarwal et al., 2024; Wang et al., 2024).	AI Scientist (Lu et al., 2024).	–	–
2025	AIDE; AlphaEvolve; Hyper; ReviewRL; SAGA; SurveyX; Agentic AutoSurvey; SciToolAgent (Jiang et al., 2025b; Novikov et al., 2025; Vasu et al., 2025; Zeng et al., 2025; Du et al., 2025; Liang et al., 2025; Liu et al., 2025; Ding et al., 2025).	AI Scientist-v2; Agent Laboratory; ResearchAgent; CodeScientist; Curie; DeepScientist; SciAgents; CycleResearcher; AI-Researcher; InternAgent; Kosmos; SafeScientist; Jr. AI Scientist; Virtual Lab; Biomni; ORGANA; AutoLabs (Yamada et al., 2025; Schmidgall et al., 2025; Baek et al., 2025; Jansen et al., 2025; Kon et al., 2025; Weng et al., 2025b; Ghafarollahi and Buehler, 2025; Weng et al., 2025a; Tang et al., 2025; Team et al., 2025; Mitchener et al., 2025; Zhu et al., 2025; Miyai et al., 2025; Swanson et al., 2025; Huang et al., 2025; Darvish et al., 2025; Panapitiya et al., 2025).	AgentRxiv; aiXiv; OmniScientist (Schmidgall and Moor, 2025; Zhang et al., 2025; Shao et al., 2025).	MLE-bench; ScienceAgentBench; BAISBench; SciSafetyBench; ReplicationBench; AstaBench; BadScientist; The More You Automate (Chan et al., 2025; Chen et al., 2025; Luo et al., 2025a; Zhu et al., 2025; Ye et al., 2025; Bragg et al., 2025; Jiang et al., 2025a; Luo et al., 2025b).
2026	ERA; AutoResearch; AutoResearch-RL (Aygün et al., 2026; Karpathy, 2026; Jain et al., 2026).	Co-Scientist; AutoSOTA; ResearchEVO; Robin; NORA; Medical AI Scientist (Gottweis et al., 2026; Li et al., 2026; Zhao et al., 2026; Ghareeb et al., 2026; Zhou et al., 2026; Wu et al., 2026a).	EvoScientist; ARIS; Intern-Atlas; Claw4Science; AI-RO; AiraXiv; AutoResearchClaw; AutoResearch Roadmap (Lyu et al., 2026; Yang et al., 2026a; Wu et al., 2026b; Xu et al., 2026; Binkyte et al., 2026; Pan et al., 2026; Liu et al., 2026; Kong et al., 2026).	FIRE-Bench; LABBench2; ResearchCodeBench; SciIntegrity-Bench; MedSkillAudit; Ríos-García et al.; SciCrafter; claim-level auditability (Wang et al., 2026; Laurent et al., 2026; Hua et al., 2026; Yang et al., 2026b; Hou et al., 2026; Ríos-García et al., 2026; Ziheng et al., 2026; Rasheed et al., 2026).

Table 4: Chronological overview of representative AutoResearch-related work. Rows organize works by release year, while columns separate task-level AutoResearch, AI-scientist workflows, infrastructure-oriented communities, and cross-cutting evaluation or auditability. The table complements Fig. 1, which gives the visual timeline.

Automation depth	Collapsed	Partially separated	Auditable / distributed
Tool / assistant	Literature, survey, and tool agents such as LitLLM, AutoSurvey, SurveyX, and SciToolAgent (Agarwal et al., 2024; Wang et al., 2024; Liang et al., 2025; Ding et al., 2025).	Staged human-guided research assistants and tool-use workflows remain emerging.	Evidence-linked assistants with explicit provenance and claim grounding remain emerging (Vasu et al., 2025; Rasheed et al., 2026).
Agent	Metric or workflow agents such as AIDE, AutoResearch, and AI Scientist (Jiang et al., 2025b; Karpathy, 2026; Lu et al., 2024).	AI Scientist-v2, CycleResearcher, InternAgent, Kosmos, SafeScientist, and Jr. AI Scientist introduce richer search, feedback, or safety-aware workflows (Yamada et al., 2025; Weng et al., 2025a; Team et al., 2025; Mitchener et al., 2025; Zhu et al., 2025; Miyai et al., 2025).	Claim-linked agents and research-object logging remain emerging (Binkyte et al., 2026).
Multi-agent	Role-prompted teams without independent audit.	SciAgents, Co-Scientist, Robin, NORA, and Virtual Lab separate scientific subtasks across agents or specialist modules (Ghafarollahi and Buehler, 2025; Gottweis et al., 2026; Ghareeb et al., 2026; Zhou et al., 2026; Swanson et al., 2025).	ARIS and AutoResearchClaw point toward executor–reviewer separation, claim audit, and targeted human intervention (Yang et al., 2026a; Liu et al., 2026).
Infrastructure / community	Shared outputs without strong governance.	AgentRxiv, AiraXiv/aiXiv, Intern-Atlas, and Claw4Science provide shared reports, publication layers, methodological graphs, or reusable skills (Schmidgall and Moor, 2025; Pan et al., 2026; Zhang et al., 2025; Wu et al., 2026b; Xu et al., 2026).	Proposed AI research community: role separation, shared memory, artifact audit, and human governance.

Table 5: Design space of AutoResearch systems. Rows indicate increasing automation depth, from tools and agents to multi-agent systems and community-level infrastructures. Columns indicate increasing epistemic distribution, from collapsed workflows to partially separated roles and auditable research communities.

Dimension	Illustrative operationalization
Claim–evidence coverage	The fraction of identified verifiable claims that are explicitly linked to supporting evidence, such as data, code, raw outputs, figures, or provenance-grounded citations.
Provenance completeness	The fraction of required provenance fields recorded for the artifacts supporting a claim, such as source, version, parameters, execution environment, and responsible actor.
Failure preservation	The fraction of registered failed or abandoned attempts that are retained with their outcomes and reasons for termination.
Independent replication	The fraction of sampled claims reproduced by a separate executor under a predefined, domain-specific protocol and tolerance.
Audit success rate	The fraction of known benchmark issues correctly identified by an independent auditor, with false positives reported separately.

Table 6: Illustrative operationalizations of verifiability-centered evaluation dimensions.

Component	Protocol
Coverage cutoff	Works publicly available online by May 20, 2026
Primary sources	arXiv, OpenAlex, Semantic Scholar, and Google Scholar
Supplementary source	OpenReview, particularly for benchmark, evaluation, and system papers under active review
Search scope	Autonomous research systems, automation of individual research stages, research infrastructure, replication, auditing, scientific review, provenance, integrity, and human–AI research communities
Representative search terms	“AI scientist”, “AutoResearch”, “autonomous research agent”, “automated scientific discovery”, “LLM agent scientific discovery”, “LLM agent research automation”, , “autonomous scientific exploration”
Screening procedure	Programmatic retrieval and deduplication, title-and-abstract relevance screening, manual full-text eligibility review, and backward and forward citation snowballing
Final decisions	All inclusion, exclusion, and taxonomy decisions were manually verified by the authors

Table 7: Survey literature-selection protocol.