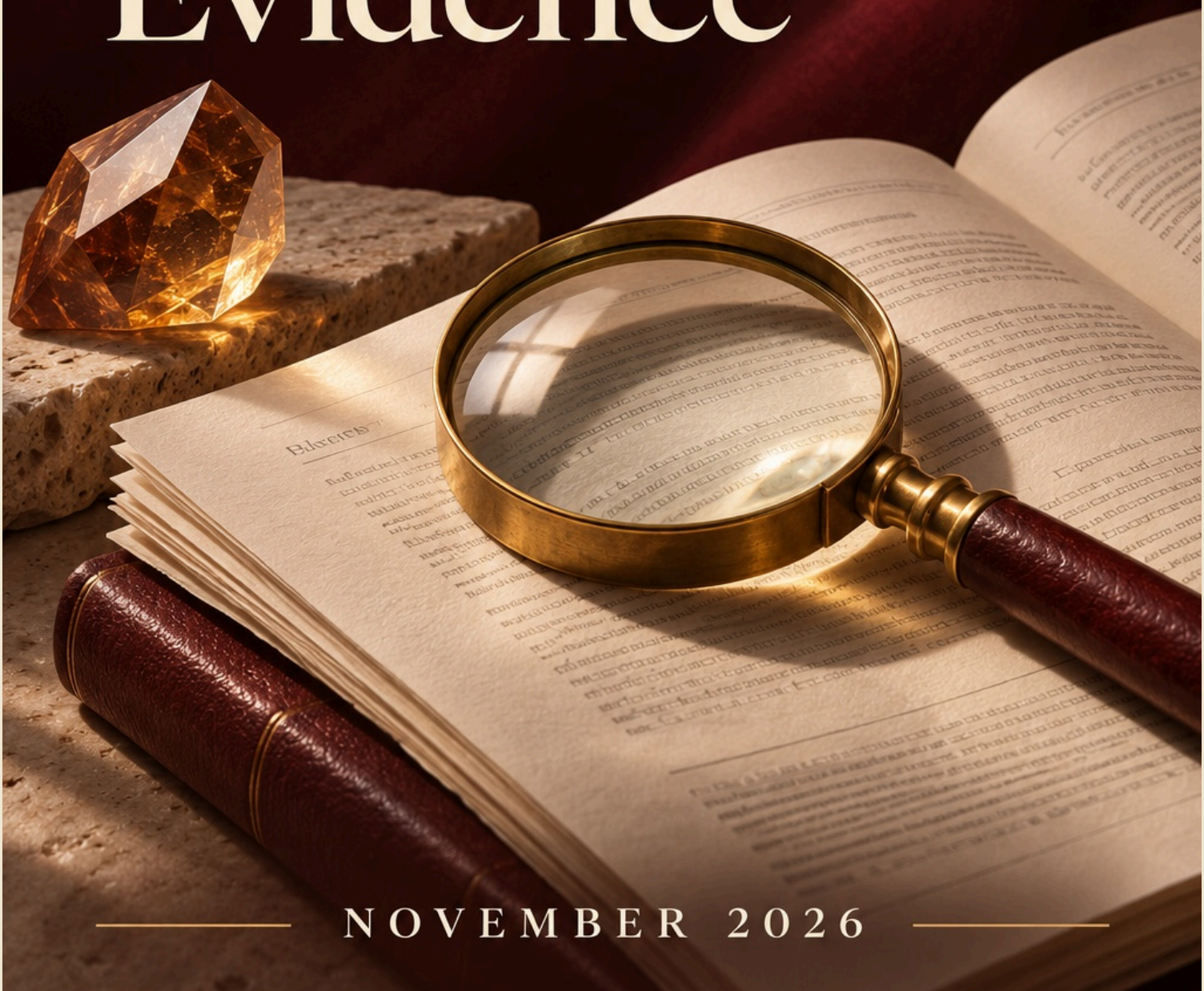
THE ARISING

AI Research Agents: From Search to Scientific Evidence



NOVEMBER 2026

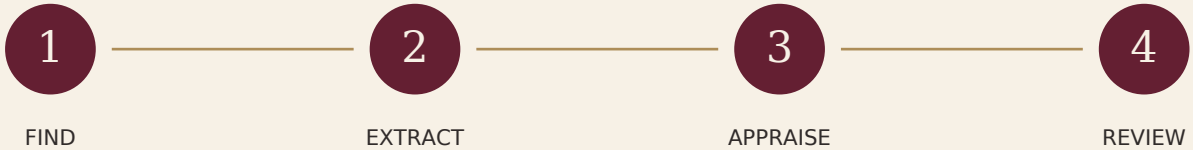
What should an agent be trusted to do?

Medical research arrives faster than any reader can comfortably absorb. An assistant that finds papers, connects ideas and prepares an explanation offers an appealing response. The hard question comes after retrieval: what has actually been learned, and what still needs to be checked? A useful research agent should make that distinction easier to see.

Here, an agent means a system that can choose and perform steps toward a goal using tools, rather than only produce a single answer. A search assistant, a code-writing research system and an instrument-connected laboratory agent may all fit that description. Their responsibilities and risks differ. The presence of an agent says little by itself about the quality of the work.

Consider a proposed review of a new health intervention. Finding an article is one task. Identifying the enrolled population, comparison group and measured outcome is another. Deciding whether the result warrants a recommendation requires further judgment. An editorial workflow should leave a visible boundary between these tasks instead of allowing a polished paragraph to stand in for all three.

The practical promise is greater reach with a shorter route back to the evidence. The test is whether a reader can inspect the source, understand the limitation and revise the conclusion. This issue examines three research examples, then proposes an accountable workflow. It makes no claim that such a workflow is already operating automatically in a particular publication.



A proposed editorial sequence. Passing a task to the next stage does not certify its evidence.

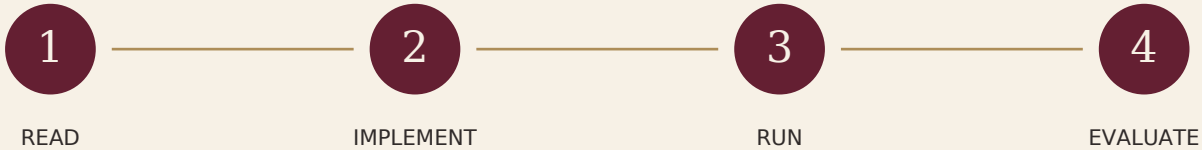
Replication is a demanding test

PaperBench asks agents to reproduce work from 20 ICML 2024 research papers, including building code and running experiments. Its author-developed rubrics divide this work into 8,316 gradable tasks. In the reported evaluation, the best tested agent achieved an average replication score of 21.0%. The authors used an automated judge and also studied its performance. [1]

These are historical results for the systems and conditions tested in that publication, not a ranking of today's products. The score measures progress against a rubric. It should not be described as the percentage of science that an AI can do, or the chance that an individual paper has been completely reproduced. Nor does this machine-learning benchmark establish medical usefulness.

For a research editor, the important design lesson is to ask for observable work. A convincing explanation of a paper is not a substitute for an executable analysis, and executable code is not sufficient if it answers the wrong question. An evaluation should specify which output counts, how it will be checked and which resources the system is allowed to use.

A local trial might begin with a small, fixed set of papers and questions whose answers have already been checked by a qualified reviewer. Save the initial outputs before correcting them. Record missed limitations, wrong denominators and unsupported claims. Otherwise, a beautifully repaired final article can conceal the actual amount of supervision needed. This is an editorial test proposal, not a claim of measured performance.



A benchmark can test specified work. Its score does not measure clinical effectiveness or universal scientific competence.

An idea still needs an experiment

Co-Scientist is a multi-agent system for generating and refining research hypotheses from scientific evidence and researcher-defined goals. The revised 2026 author record describes biomedical applications including drug repurposing, target discovery and antimicrobial resistance. It reports in vitro validation of acute myeloid leukemia drug candidates and combinations. [2]

That description places the work in a specific evidential category: AI-assisted hypothesis development with experimental follow-up. A result in cells can justify more investigation, but cannot establish that a treatment helps patients. The contribution of the system and the contribution of the researchers who designed and performed validation should remain distinguishable.

A useful hypothesis should expose itself to possible failure. What result would count against it? What competing explanation would produce the same observation? Which measurement would separate them? For an editorial reader, these questions are more informative than the apparent sophistication of a proposed mechanism. They can also reveal when an agent has mainly rewritten a familiar idea.

A practical hypothesis record could contain the proposed mechanism, its supporting observations, contradictory evidence, feasible tests and the reasons for choosing one test over another. Keep the original proposal alongside later revisions. If only the eventual success survives, the number of unsuccessful suggestions disappears and the system's usefulness becomes difficult to judge. This record is a proposed discipline for evaluating ideas, not a validated recipe for discovery.



Hypothesis quality, experimental support and patient benefit answer different questions. No automatic progression is implied.

When software reaches the laboratory

Coscientist, a separate system reported by Boiko and colleagues in 2023, combined language models with search, code and experimental tools. The paper demonstrated chemistry tasks including instrument control and selected reactions. It also evaluated optimization using existing reaction datasets. Those computational tests must be distinguished from newly performed physical experiments. [3]

This matters because an agent's description can make several different accomplishments sound alike. Selecting a promising condition from an existing dataset, generating a plausible protocol and obtaining a measured result from equipment are distinct events. A report should tell the reader which event occurred and what direct observation supports it.

An editorial assessment should ask who prepared the equipment, supplied the materials, defined the allowed actions and interpreted the measurements. These questions do not diminish an advance. They identify the setting in which the result can be understood. The word autonomous should be followed by a description of the boundary: autonomous at which step, using which tools, with which intervention points?

Laboratory work also changes the stakes of an error. A wrong summary may mislead a reader; a wrong instruction may damage equipment or expose people to harm. Institutions need appropriate technical, safety and ethical review before connecting a system to physical capabilities. This article offers no experimental protocol. Its subject is how to read claims about such systems without confusing a bounded demonstration with unrestricted scientific independence.



Planning and physical execution require separate checks. Instrument access is not permission to run any experiment.

Make every claim traceable

A proposed research-to-publication workflow begins with a question narrow enough to check. Preserve the search terms, selection rules and source identity. Separate duplicate reports of the same study from genuinely independent evidence. The output of this first stage should be a set of candidate sources, not a declaration that their conclusions are established.

Next, extract the claim with its context: population or model, intervention or exposure, comparator, duration, endpoint and uncertainty. Include the source passage or table that supports the wording. An abstract can identify a study and its headline result, but it may not provide the details needed for a full appraisal. Mark the reading scope so later writers know what remains unexamined.

Before drafting, decide whether each claim should be retained, narrowed, held or removed. A mechanistic interpretation belongs beside the observation rather than silently replacing it. A statistically detectable change in a surrogate measure should not become a promise about how long or how well someone will live. If the necessary evidence is missing, leaving the question open is an editorial result.

Finally, give the claim a stable identifier and connect it to every place it is used. If a correction changes the denominator or interpretation, the linked article and summary can be revisited together. Keep approval separate from source verification. A checked citation is not a publication decision, and a second pass by the same system is not an independent review. These are proposed controls, not a claim that any named review process has occurred.



A claim record links evidence, interpretation and use. Correction should follow those links wherever the claim appears.

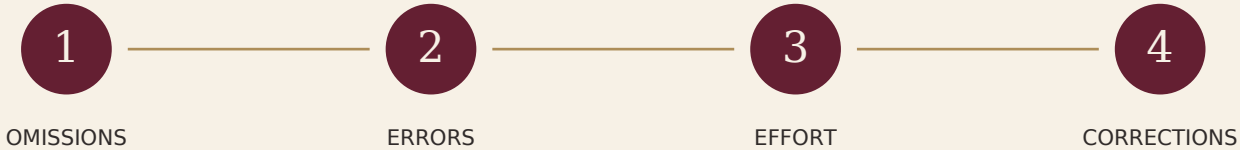
Measure the work that remains

Time saved is an attractive measure, but it needs a complete denominator. Count the time spent defining the task, locating inaccessible sources, checking outputs, correcting mistakes and resolving disagreement. A fast first draft can still create a slow editorial job. Conversely, a system that reliably identifies what it cannot answer may be more useful than one that always delivers a confident conclusion.

A sensible pilot should include difficult cases, not only representative successes. Try a corrected paper, conflicting endpoints, a negative result and a promising preclinical finding. Ask reviewers to inspect numerical accuracy and the strength of the language separately. A sentence may reproduce a number correctly while implying a clinical benefit that the study never measured.

Use the pilot to decide which tasks to delegate next. Retrieval may be ready for broad assistance while appraisal still needs close review. Summaries for public education may require different checks from internal research notes. Any expansion of the system's responsibilities should follow evidence about that specific responsibility, with a person able to stop or reverse the process.

The most valuable research assistant may be the one that leaves us with better questions and a clearer account of our uncertainty. Wider attention is useful only if it strengthens the connection between a claim and the world it describes. For a reader deciding what to believe about health, the essential promise is modest and demanding: make evidence easier to inspect, and make mistakes easier to correct.



Proposed evaluation dimensions: missed evidence, factual errors, reviewer effort and correction speed. No performance values are implied.

Primary sources

[1] Starace G, Jaffe O, Sherburn D, et al. PaperBench: Evaluating AI’s Ability to Replicate AI Research. PMLR. 2025;267:56843-56873. https://proceedings.mlr.press/v267/starace25a.html

[2] Gottweis J, Weng WH, Daryin A, et al. Accelerating scientific discovery with Co-Scientist. arXiv:2502.18864v2. 29 June 2026. Related Nature DOI: 10.1038/s41586-026-10644-y. https://arxiv.org/abs/2502.18864v2

[3] Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624:570-578. DOI: 10.1038/s41586-023-06792-0. https://pmc.ncbi.nlm.nih.gov/articles/PMC10733136/

What these studies support

PaperBench: publication abstract and benchmark description consulted. Co-Scientist: revised author record and abstract consulted; the linked journal full text was not accessible. Coscientist: accessible full-text sections on architecture, experiments and optimization consulted. This is a bounded reading, not an independent reproduction or exhaustive review of supplements.

The examples concern computational tasks, hypotheses and laboratory research. None establishes improved patient outcomes from using a research agent. Editorial workflow diagrams are proposals, not measured system performance.