Science at the Frontier: How Scientists Use ChatGPT*

Chris Liao† Carl Shan‡ Aaron Chatterji§

November 6, 2026

Abstract

We study how academic scientists use AI, analyzing over half a million scientific messages from 1,819 scientists enrolled in the ChatGPT for Academic Researchers plan. We document four key findings. First, AI use is concentrated in a few scientific activities: just two—reading and writing—account for over half of the average scientist's scientific messages. Second, AI plays a role in shaping which ideas scientists consider and how they develop them into research questions. When exploring research directions, scientists often ask AI to propose what to investigate; when developing specific research questions, scientists typically supply the question or idea and ask AI to refine or evaluate it. Third, scientists remain involved in research execution, following up on AI output after about one-third of requests, on average. Within the same scientific task, scientists follow-up more often for requests requiring greater scientific expertise. Finally, 98% of survey respondents expect AI to increase the rate of scientific discovery in their field, most often citing faster completion of research tasks.

*We thank Caroline Chin, Nathaniel Craig, Sarah Friar, Paul Hamilton, David Holtz, Drew Johnston, Annie Lee, Eric Lopez-Ziemann, Sarah Metz, Elliott Mokski, Kyle Myers, Christopher Ong, Nidhi Patel, Alex Martin Richmond, Jen Robinson, Harrison Satcher, Cassandra Duchan Solis, Ker Than, Gawesha Weeratunga, Zak Weiler, Feitong Yang, and Meng Jia Yang for their feedback, suggestions, and support. We are grateful to Gabriel Bo, Emile Contal, Yilin He, Belinda Truong, and Peter Zhao for their help with the classifications.

†University of California, Berkeley and OpenAI. Email: chrisliao@berkeley.edu. Chris Liao is a paid contractor at OpenAI.

‡OpenAI. Email: cshan@openai.com.

§Duke University and OpenAI. Email: ronnie@duke.edu.

1 Introduction

Scientific progress is an important source of economic growth and improved living standards (Romer, 1990). Recent advances in artificial intelligence (AI) have raised hopes that scientific progress can be accelerated (Altman, 2024; Amodei, 2024), motivated in part by the potential for AI technologies to assist scientists with knowledge-work tasks like developing new ideas or executing research tasks (Cockburn et al., 2019; Wang et al., 2023). How fully this potential is realized depends on how scientists are able to use AI across scientific work. Existing evidence suggests that the benefits of AI assistance vary widely across knowledge-work tasks (Patwardhan et al., 2026; Dell'Acqua et al., 2026). This matters because productivity gains in only some scientific tasks may have limited effects on overall scientific progress if other essential tasks remain bottlenecks (Agrawal et al., 2024; Anandkumar, 2026).

In this paper, we study how scientists use AI across scientific work. First, to understand where scientists find value in AI use, we examine how extensively AI is used across the scientific workflow. Second, because new ideas drive scientific progress and long-run economic growth (Romer, 1990), we examine how scientists divide ideation—exploring ideas and developing research questions—between themselves and AI. Third, while AI can help with turning ideas into discoveries by carrying out research tasks, scientists may still need to spend time checking and improving its outputs. We therefore measure how often scientists request clarification, checking, or revision of AI output. Since scientific research tasks differ across fields and scientists' experience and responsibilities may shape how they use AI, we compare AI use across 7 scientific fields—biology, chemistry and materials sciences, computer science, earth and planetary sciences, engineering, mathematics, and physics—and 3 age groups: scientists aged 18–29, 30–39, and 40 and older. Finally, to understand the broader implications of AI for scientific work, we survey scientists about their expectations for how and how much AI will accelerate scientific discovery, their self-reported productivity gains from AI use, and the bottlenecks that continue to constrain their research productivity.

Our analysis relies on two novel datasets. First, we assemble a dataset of AI usage by 1,819 scientists enrolled in the ChatGPT for Academic Researchers plan. Over a 34-day span in 2026, these users sent 580,186 messages about scientific work. To categorize what scientific work scientists use AI for, we map messages to one of 10 scientific workflows, and within each workflow, a specific scientific task,[1] using a taxonomy based on SciNet (Althoff, 2026; Alamos and Althoff, 2026). We supplement this taxonomy with additional classifications characterizing how scientists engage with AI on ideation and the execution of scientific work. Our second dataset consists of responses from a survey of scientists on the ChatGPT for Academic Researchers plan conducted by OpenAI. We survey scientists to learn about aspects of AI's role in their research that usage data alone cannot speak to. To protect scientists' privacy and intellectual property, our analysis of scientific messages relied exclusively on outputs from privacy-preserving classifiers that categorized AI use along predefined di-

[1] In total, our taxonomy consists of 40 scientific tasks across all 10 workflows. The task taxonomy also includes an "other" category.

mensions. At no point did any researchers observe or have access to the underlying scientist messages.[2]

Analyzing these two datasets yields four main findings. First, 85% of the average scientist's scientific messages to ChatGPT and Codex fall in four workflows: reading & knowledge acquisition, scientific analysis, software development, and writing. Reading & knowledge acquisition and writing account for over half. Other aspects of scientific work like ideation, data gathering, peer review or administrative work are less represented. Second, however, some scientists are using AI to explore and develop research ideas. In our sample, 42.7% of scientists average more than two messages about ideation per week. Scientists often ask AI to suggest possible directions when exploring new research ideas. When developing ideas into specific research questions, they tend to provide the initial idea and use AI to help improve it. Third, scientists remain actively involved when delegating research task execution to AI, following up to clarify, check or revise AI output 32% of the time. Scientists are most involved in workflows where AI capabilities have received substantial attention: writing, scientific analysis and software development. Within tasks, scientists are more involved when the task requires greater scientific expertise. Finally, 98% of survey respondents expect AI to accelerate scientific discovery in their field. 21.6% expect AI to shorten a 10-year discovery timeline to three years or less. Faster research execution is the most commonly cited mechanism for accelerated scientific discovery, and scientists report larger productivity gains in workflows where they use AI more. However, insufficient personnel, funding and time remain bottlenecks to increased research productivity.

We make two contributions to the literature. Our first contribution is to the empirical literature on scientists' AI use. Codreanu et al. (2026) examine AI use across research tasks and complementarities between a general-purpose model, Gemini, and specialized AI models. Hermle et al. (2026) study Claude use across fields and tasks and examine whether model capabilities explain differences in AI use intensity. They estimate potential increases in the rate of scientific progress under different scenarios for AI automation by calibrating an ideas production function. Our study focuses on how scientists currently use AI, by examining not only which tasks scientists use AI for, but also how they divide the work of research ideation and execution between themselves and AI. Moreover, our analysis uses ChatGPT usage data from known academic scientists, whereas both studies analyze AI interactions classified as scientific work. Like ours, both studies combine data on AI interactions with surveys of scientists. Other work studying how scientists use AI consists of surveys of scientists' reported AI use (Andersen et al., 2025; Hrycyshyn and Eassom, 2025; Chugunova et al., 2026).

Our second contribution is to the literature on AI's implications for scientific productivity. Existing large-scale observational studies relate inferred AI use to scientific output and the direction of research (Kusumegi et al., 2025; Hao et al., 2026). Hill and Stein (2026) ana-

[2]By default, content from the ChatGPT for Academic Researchers plan is not used to train OpenAI models (OpenAI, 2026b).

lyze the causal effect of AlphaFold on experimental structural biology and downstream basic and applied research, tracing its effects across multiple stages of a specific scientific area. We complement these studies with survey evidence on scientists' self-reported productivity gains from AI across scientific workflows and fields. Economic theory shows how bottlenecks can limit the extent to which improvements in individual tasks translate into broader scientific progress (Aghion et al., 2019; Jones, 2026). We provide empirical evidence on these productivity gains and bottlenecks: where scientists report gains from AI, how those gains relate to AI use intensity, and which bottlenecks continue to limit research productivity.

2 Data and measurement

2.1 Usage data

2.1.1 Sample construction and summary statistics

ChatGPT for Academic Researchers, an OpenAI program launched on July 29, 2026, provided researchers at selected academic institutions with free access to OpenAI frontier models and higher usage limits. The program targeted research faculty and postdoctoral researchers actively researching topics in biological sciences, chemistry and materials, computer science, Earth and planetary sciences, engineering, mathematics, and physics. Seats were allocated by lottery among eligible applicants (OpenAI, 2026a). Each approved applicant received a ChatGPT workspace and could invite up to four colleagues from their institution to join it. We refer to the applicant as the workspace owner and the invited colleagues as collaborators. Appendix A.1 describes the program's published eligibility requirements.

We study messages sent by participants in the program through ChatGPT's Chat and Work modes and through Codex, including the Codex command-line interface, from August 20, 2026 through September 22, 2026. Our unit of observation, a message, is a user message in ChatGPT's Chat or Work mode or a native turn in Codex. The cohort of scientists we study consists of the 1,819 scientists who sent at least one scientific message during this window by September 2, ensuring all scientists are observed for at least 21 calendar dates, including their first date of use. The cohort sent 1,262,893 messages, of which 580,186 were classified as scientific or scholarly messages.[3] Appendix A.2 describes sample construction.

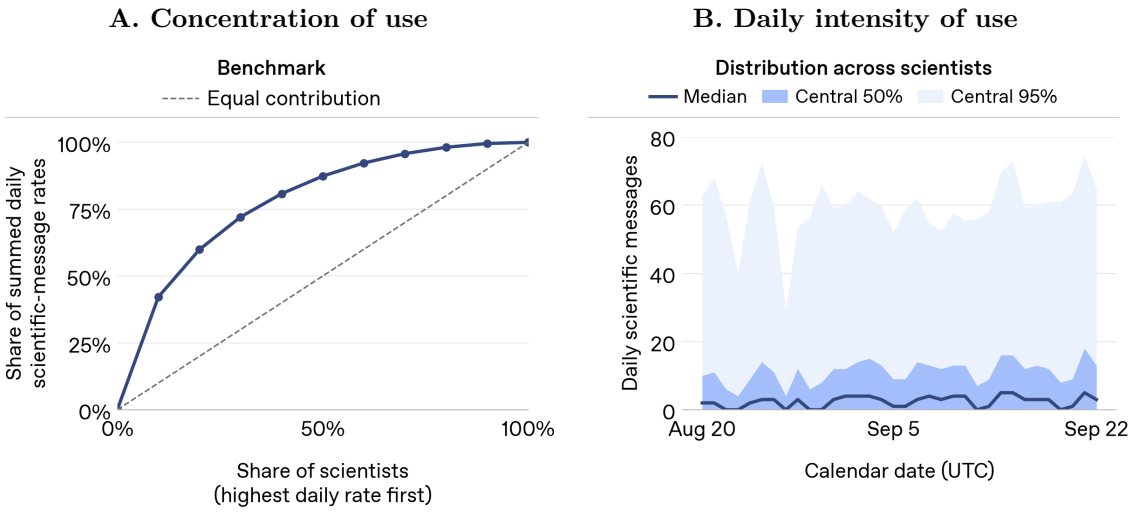
At no point in this or subsequent work did researchers view the underlying scientist messages. By default, content from the ChatGPT for Academic Researchers plan is not used to train OpenAI models.[4]

Panel A of Figure 1 plots the inverse Lorenz curve of scientists' average daily scientific-message rates, showing the cumulative share of the summed daily rates accounted for by

[3]Scientific tasks account for 45.9% of all recorded messages and 56.6% of the 1,025,859 classified messages; 237,034 messages are unclassified. Figure A1 shows the distribution across request-status classifications.

[4]See OpenAI (2026b) for the plan's default data-use policy.

Figure 1: Concentration and intensity of scientific AI use



Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. Panel A shows cumulative shares of scientific AI use. Points mark successive deciles of scientists, ranked from highest to lowest average daily scientific messages. Each scientist's daily average divides their scientific messages by the number of calendar days from their first observed scientific message through September 22. Cumulative shares use the sum of these daily averages. Panel B shows the distribution of scientific messages sent on each date among scientists observable that day, beginning on the later of August 20 and their earliest recorded program-membership date. The line shows the median; the dark and light shaded bands show the 25th–75th and 2.5th–97.5th percentiles, respectively.

scientists ranked from most to least active.[5] The distribution of scientific messages per day is right-skewed; the most active 10% of scientists account for 42% of scientific-message activity, after adjusting for differences in how long each scientist's use is observed. Scientific AI use varies substantially even within this cohort that applied for expanded AI access.[6]

Panel B of Figure 1 shows the distribution of scientific messages across scientists over time. Across all eligible scientist-days from August 20 through September 22, 2026, the median number of scientific messages was 2.[7] The 75th and 97.5th percentiles were 12 and 61 scientific messages, respectively, whereas the 25th percentile was 0. The distribution of usage intensity also remains broadly similar over the sample period. To ensure that our analyses are not skewed by heavy users, in all subsequent analyses, we give scientists equal weight when averaging their workflow shares.

Appendix Table A1 reports, for each scientific field and age group, the mean, median, and interquartile range of scientific messages per day. The median scientist in biology averages 4.9 scientific messages per day, compared with 8.3 in earth and planetary sciences.[8] Younger

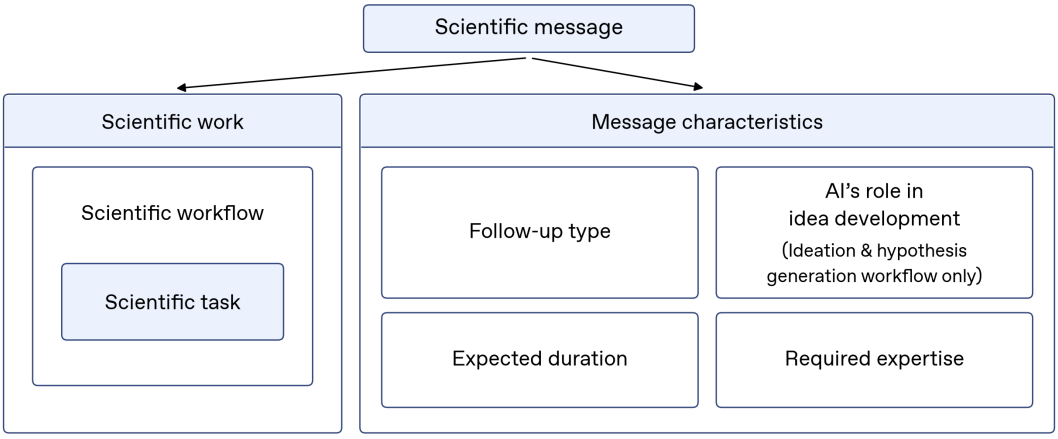
[5]We measure the concentration of AI use using scientific messages per day because scientists join the plan on different dates after our sample period starts.

[6]Usage among scientists does appear to be less concentrated among heavy users than among the broader public. Data from the survey research company YouGov found that within a self-selected sample of American ChatGPT users, the top decile accounted for 63% of prompts (Das et al., 2026).

[7]Figure 1B shows the daily distributions. The percentiles reported here pool all eligible scientist-days from August 20 through September 22, 2026, including days with no scientific messages.

[8]A few days of heavy AI use can raise a scientist's average number of scientific messages per day, so the

Figure 2: Classification of scientific messages



Notes: Scientific work is classified by scientific workflow, a broad category of scientific work, and scientific task, a specific action within a workflow, as defined in Section 2.1.2. Follow-up type describes the relationship to prior AI output. AI's role in idea development describes who is asked to supply or shape the scientific question or target. Expected duration is the active time a competent, experienced person would need to complete the requested task without AI. Required expertise is the minimum preparation needed to complete the request.

scientists also send more scientific messages per day: the median scientist aged 18–29 averages 7.1, compared with 6.9 among those aged 30–39 and 4.6 among those aged 40 and older.

These patterns raise two questions for further research. First, how much of the variation in AI use across scientists reflects differences in what AI can contribute to scientists' research, and how much reflects other barriers, such as limited awareness about AI capabilities? Second, if less intensive users could realize productivity gains by increasing AI use, how much could more intensive use of AI capabilities accelerate scientific progress?

2.1.2 Task taxonomy and message characteristics

We classify each scientific message along two dimensions: first, by what scientific work is being done, and second, by four characteristics describing how the scientist engages with AI. Figure 2 visualizes the classification scheme we use. We use GPT-5.6-Terra to classify messages.

Our taxonomy of scientific work adapts SciNet's taxonomy to define 10 scientific workflows and 40 scientific tasks. Workflows are defined as broad categories of scientific work and scientific tasks are defined as specific actions within a workflow (Althoff, 2026; Alamos and Althoff, 2026).[9] Table 1 defines each scientific workflow and provides example messages corresponding to each. For example, Reading & knowledge acquisition is defined as "find, explain, compare, or synthesize existing scholarly knowledge" and one task within this work-

median of these averages in Table A1 typically exceeds the median scientific message count across scientists on a single day in Figure 1B.

[9]Appendix A.3.1 describes how our task taxonomy adapts SciNet's universal task taxonomy, including what changes we make.

Table 1: Taxonomy of scientific workflows

Workflow	Definition	Illustrative example messages
Reading & knowledge acquisition	Find, explain, compare, or synthesize existing scholarly knowledge.	"What did this paper do?"
Ideation & hypothesis generation	Formulate research questions, hypotheses, conjectures, models, approaches, or directions.	"What route could prove this?"
Administration	Coordinate research funding, personnel, logistics, equipment, ethics, or institutional approvals.	Prepare its budget or ethics approval (a grant).
Data gathering	Design collection protocols or instruments, or acquire, collect, curate, or document original research evidence.	Acquire observations, records, measurements, or datasets as research evidence
Research software & code development	Implement, debug, test, optimize, or maintain research code and computational infrastructure.	Implement, modify, debug, test, configure, or optimize it (research software).
Scientific analysis	Consists of two workflows 1. Empirical analysis: clean, estimate, model, simulate, visualize, or interpret empirical research data. 2. Theoretical and formal analysis: prove, disprove, derive, bound, or validate mathematical or formal claims.	"Run this regression and report the estimate." / "Prove this theorem."
Writing & communication	Draft, revise, or present research manuscripts, grant narratives, or scientific visuals.	"Shorten paragraph two of the manuscript I uploaded"
Peer review & service	Perform an established referee, editorial, funding-panel, or formal academic-service role.	"I'm reviewing this manuscript for a journal. Assess its scientific rigor and contribution."
End-to-end autonomous research	Use only when AI or an AI-agent system is asked to independently coordinate one integrated scientific-discovery run that includes all three existing-workflow components: 1. Reading & knowledge acquisition 2. Ideation & hypothesis generation 3. One downstream branch: either Data gathering plus Empirical analysis, or Theoretical and formal analysis.	"Run an AutoResearchClaw-style topic-to-paper workflow on this topic in fully autonomous mode: search literature, generate and screen hypotheses, collect experimental results, analyze them, and produce the report."
Mentorship & teaching	Teach or supervise identified students, trainees, assistants, or postdoctoral researchers.	Course materials or guidance for an identified trainee

Notes: The table lists each scientific workflow, its definition, and illustrative example messages. Workflows are ordered in the approximate order we believe they first occur in the scientific research process. Scientific analysis combines Empirical analysis and Theoretical and formal analysis. The definition of End-to-end autonomous research distinguishes between these components of Scientific analysis. The definitions are those used in the classification prompt.

flow is "Explain scientific concepts". Appendix Table A2 provides the full task definitions and Appendix A.3.2 describes the classification procedure in more detail.

We also characterize how scientists engage with AI. Table 2 defines each message characteristic and its possible values, and provides example messages when possible. **Follow-up type** distinguishes initial requests from responses to prior AI output. For responses to prior

Table 2: Characteristics of scientific messages

Characteristic	Definition and possible values	Example message-value pairs
Follow-up type	An initial request is a scientific message that starts work without relying on prior AI output. Other messages may request clarification, checking, revision, or extension/application, or express a reaction. **Clarification** seeks explanation or understanding of prior AI output; **checking** questions or checks that output without identifying a specific correction; **revision** requests a change, fix, or replacement; **extension/application** builds on prior AI output to request a continuation, application, or new deliverable; and **reaction** evaluates that output without requesting further work. Insufficient evidence means dependence on prior AI output is apparent or plausible but the relationship cannot be established.	"Why does that step work?" → clarification. "Does that step actually work?" → checking. "Replace that incorrect step" → revision. "Apply that result to my model" → extension/application. "Thanks, that solved it" → reaction.
AI's role in idea development	**Scientist-supplied:** the scientist supplies a concrete scientific question, problem, claim, or target and does not ask AI to materially change it. **AI-refined:** AI is asked to materially narrow, broaden, redefine, or reframe a supplied question or target. **AI-formulated:** AI is asked to choose or formulate the question or target from a topic, evidence, goal, or broad area. We analyze this characteristic within Ideation & hypothesis generation messages only. Insufficient evidence means the message does not establish who determines the question or target.	"What are your thoughts on this aspect of the proof?" → scientist-supplied. "Reframe this research question: . . . " → AI-refined. "Suggest a research question about . . . " → AI-formulated.
Expected duration	Estimated active human minutes a competent, experienced person would need to complete the work without AI, excluding passive waiting, following Johnston et al. (2026). We coarsen responses into three groups: under one hour, one to under four hours, and four hours or more. Messages without a substantive request are assigned zero minutes.	
Required expertise	Minimum preparation required by the work: general/foundational; standard technical/professional; advanced research; frontier/subfield specialist. Insufficient evidence means the minimum preparation cannot be inferred from the available message context.	

Notes: The Characteristic column lists the four characteristics of engagement with AI: follow-up type, AI's role in idea development, expected duration, and required expertise. The Definition and possible values column lists each characteristic's definition and possible values. The Example message-value pairs column lists illustrative messages and their assigned values.

AI output, it captures whether scientists request clarification, checking, revision, or extension/application of prior AI output, or express a reaction to it. Within ideation & hypothesis generation, **AI's role in idea development** identifies who is asked to define the scientific question or target: the scientist supplies it, asks AI to refine it, or asks AI to formulate it from broader inputs. Insufficient evidence means the message does not establish who determines the question or target. **Expected duration** measures the active time a competent, experienced person would need to complete the requested task without AI. **Required expertise**

measures the minimum knowledge and training needed to complete the task, ranging from general or foundational preparation to standard technical or professional training, advanced research expertise, and frontier expertise in a subfield.

2.2 Survey data

We surveyed scientists on the ChatGPT for Academic Researchers plan from September 30, 2026 through October 7, 2026 about their experience using AI for scientific work and their expectations for the pace of scientific discovery. Out of 270 valid completed responses, 251 respondents consented to having their results included in this analysis. Appendix B.1 lists all referenced survey questions and their response options.

We elicited expectations about scientific discovery by asking respondents to consider an important discovery in their field that would take ten years without AI. Respondents reported how much sooner, if at all, they expected the discovery to occur with AI (Q1). Those expecting acceleration were then invited to explain how AI would bring it about in an open-ended response (Q2). We received 224 written responses. The mechanism analysis contains 367 distinct claims, with each claim classified into one of eight mechanism categories or a separate "Not specified" category. These nine categories are organized into four groups: faster research execution, improved search for ideas, expanded research directions, and other or unspecified. Appendix B.2 describes the categories and classification procedure.

We complement these questions about overall scientific progress with questions about scientists' current AI use, perceived productivity gains across scientific workflows, and constraints scientists face in their research. Respondents first identified the scientific workflows for which they had used AI during the preceding 30 days (Q3). For each selected workflow, they reported the share of total workflow time spent actively using AI (Q4) and estimated how much more or less time they would have needed without AI to complete the same work at the same quality (Q5). Follow-up questions were shown only for selected workflows, so we code confirmed nonuse as 0% AI use and no productivity gain from AI use. Respondents were also asked about their current research bottlenecks (Q6) and difficulties they encountered obtaining value from ChatGPT for Academic Researchers (Q7).

Finally, respondents identified their research field or fields (Q8), with multiple selections permitted.[10]

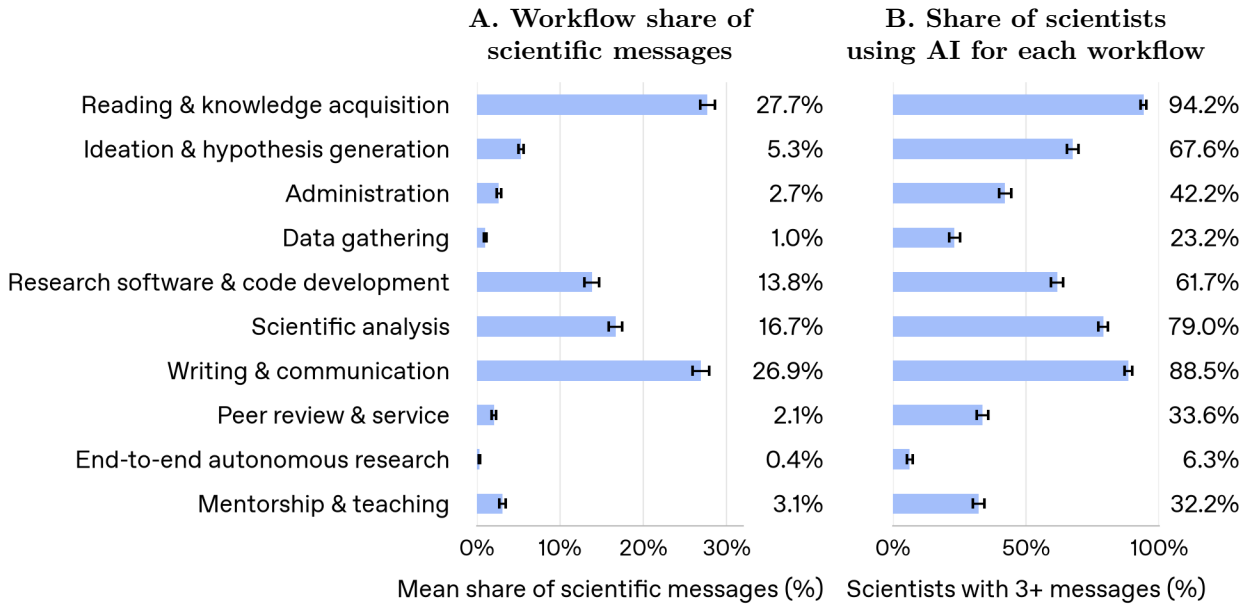
3 AI use in scientific research

3.1 AI use across scientific work

In this section, we examine where scientists use AI across scientific work to understand where they find value in it. We quantify how much they use AI across the workflows and tasks

[10]These field categories differ from those used in the usage-data analysis.

Figure 3: AI use across scientific workflows



Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. Panel A shows the workflow share of scientific messages, with pointwise 95% scientist-level t confidence intervals. The workflow share of scientific messages is the mean of each scientist's share of all scientific messages devoted to that workflow. Panel B shows the share of scientists using AI for each workflow, with 95% Wilson confidence intervals. The share of scientists using AI for each workflow is the percentage who sent at least 3 scientific messages in that workflow during August 20–September 22. Workflows are ordered in the approximate order we believe they first occur in the scientific research process. Other scientific work is omitted from the plots but remains in the denominators.

described in Section 2.1.2 and examine how use varies across scientific fields and age groups.

Panel A of Figure 3 shows the average distribution of scientists' scientific messages across the 10 scientific workflows.[11] We list workflows in the approximate order we believe they first occur in the scientific process. Usage is concentrated in four scientific workflows: reading & knowledge acquisition, writing & communication, scientific analysis, and research software & code development together account for 85% of the average scientist's scientific messages. Reading & knowledge acquisition and writing & communication alone account for over half. These workflows help scientists realize the benefits of the nonrival nature of ideas: reading & knowledge acquisition enables them to learn from existing scientific knowledge, while writing & communication enables them to share their own findings for others to build on (Arrow, 1962; Romer, 1990). This also raises two questions for further research about the implications of AI use on scientists' expertise and the quality of their output. First, how does AI-assisted reading & knowledge acquisition affect scientists' ability to understand and apply scientific knowledge? Second, what is the effect of AI-assisted writing & communication on the quality of written or presented research?

[11]Other scientific work, which falls outside the 10 workflows, is omitted from the workflow plots and accounts for 0.37% of the average scientist's scientific messages. During August 20–September 22, 9.84% of scientists sent at least 3 scientific messages in this category.

One surprising difference between our findings and those of Hermle et al. (2026) is how much scientists use AI for reading & knowledge acquisition. We find that reading & knowledge acquisition accounts for 27.7% of the average scientist's scientific messages, compared with 7.6% of research sessions on Claude and 3.7% on Claude Code.[12] Lower relative usage of Claude for reading & knowledge acquisition is offset by higher usage of Claude for data analysis. This contrast may reflect researchers using ChatGPT more for knowledge acquisition and Claude more for computational work or sample-driven usage differences as our sample consists of verified academic scientists.

Panel B of Figure 3 measures the extent to which use of AI for each workflow has diffused across scientists, by showing the share of scientists who sent at least 3 scientific messages in each workflow.[13] The four workflows where usage is concentrated are also among the most widely used by scientists. Panel B also shows that AI use for ideation & hypothesis generation is widespread despite accounting for a smaller share of messages.[14] Appendix Figure A2 shows that scientists are most likely to continue using AI for these five workflows in subsequent weeks as well.[15]

A limitation of Figure 3 is that message shares describe how scientists use AI, rather than the value AI provides or how scientists allocate their research time across workflows. Consequently, low message shares for a workflow could reflect any combination of limited AI adoption, reliance on non-ChatGPT AI tools, or relatively little research time devoted to the workflow. Nonetheless, we believe data gathering, administration, and peer review & service are important scientific workflows despite their smaller message shares.

In our conceptual ordering of the workflows in Figure 3, data gathering and administration connect knowledge and ideas to scientific analysis by ensuring scientists have the data and resources they need to conduct research. Because of complementarities across research workflows, AI-enabled productivity gains in ideation & hypothesis generation and scientific analysis may still yield limited gains in overall scientific progress if researchers remain constrained by data collection and administrative work (Aghion et al., 2019). Survey evidence

[12]Both studies draw on SciNet's task taxonomy, but we combine what Hermle et al. (2026) report as Data analysis and Theoretical analysis under Scientific analysis. The gap in our results is not explained by differences in weighting alone. Panel A of Figure 3 averages message shares equally across scientists, whereas Hermle et al. pool sessions. Pooling our messages lowers the reading & knowledge acquisition share from 27.7% to 24.1%, still well above both Claude estimates. Differences in field coverage are also unlikely to explain the gap for two reasons: reading accounts for 23.8–32.8% of messages across all seven scientific fields we study (see Figure 4), and their sample additionally includes sessions related to humanities and social science research, where we might expect work related to reading & knowledge acquisition to be more prominent. One potential factor is that differences in usage are explained by sample differences. Our sample is restricted to verified academic scientists, whereas theirs selects sessions classified as scientific work without restricting users to academic scientists.

[13]Requiring at least 3 messages reduces the sensitivity of the measure to isolated messages and occasional classification errors.

[14]Appendix Table A3 shows that these same five workflows remain the most widely used at alternative scientific message thresholds.

[15]Appendix Table A4 presents the task-level counterpart to Figure 3, while Appendix Figure A3 presents the task-level counterpart to Appendix Figure A2.

Table 3: Top ten scientific tasks by average message share

Task	Workflow	Task share of scientific messages	Share of scientists using AI for each task
Explain scientific concepts	Reading & knowledge acquisition	18.8%	89.7%
Draft manuscripts	Writing & communication	13.9%	81.5%
Retrieve scholarly information	Reading & knowledge acquisition	5.3%	70.3%
Implement research software	Research software & code development	4.7%	49.4%
Draft grant narratives	Writing & communication	4.7%	32.4%
Create research visuals	Writing & communication	4.2%	55.4%
Develop or test proofs	Scientific analysis	4.2%	26.2%
Formulate research problems, models or approaches	Ideation & hypothesis generation	3.9%	61.1%
Review and synthesize literature	Reading & knowledge acquisition	3.6%	61.4%
Derive formal results	Scientific analysis	3.4%	39.8%

Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. The table lists the ten tasks with the highest task shares of scientific messages, in descending order, their scientific workflows, their task shares of scientific messages, and the shares of scientists using AI for each task. The task share of scientific messages is the mean of each scientist's share of all scientific messages devoted to that task. The share of scientists using AI for each task is the percentage who sent at least 3 scientific messages in that task during August 20–September 22.

from scientists also finds that physical experimentation and data collection are among the most commonly reported research bottlenecks (Codreanu et al., 2026).

One important benefit of peer review and, more broadly, the evaluation of scientific work is that it increases confidence that scientific output is reliable. This has become especially important amid concerns about the high quantity of low-quality AI-assisted research (Andersen, 2026; O'Grady, 2025). Despite peer review & service constituting a small share of scientific messages, there's evidence that AI could support certain aspects of the review process. For example, Schwartz et al. (2026) demonstrate an AI workflow that flags inconsistencies between published results and reproduced calculations, and checks whether results depend heavily on just a few data points.

We next examine scientist AI use at a more granular level. Table 3 lists the ten scientific tasks with the highest average message shares and reports the percentage of scientists who sent at least 3 messages for each task.

Two facts stand out. First, "Explain scientific concepts" is the leading task, accounting for 18.8% of the average scientist's scientific messages, with 89.7% of scientists sending at least 3 messages for this task. As scientific knowledge grows, one barrier that researchers may face is an increasing educational burden (Jones, 2009). Frequent use of AI for concept explanation suggests that AI could help ease this burden by lowering learning costs, although

Figure 4: AI use across scientific workflows by field and age

A. Scientific field

Color: % difference from each workflow's overall mean

–100% 0 +100% +200%

	Overall	Biology	Chemistry / materials	Computer science	Earth / planetary	Engineering	Mathematics	Physics	H_0: equal group means p-value
Reading & knowledge acquisition	27.7	32.8	27.0	23.8	26.5	27.5	25.5	28.5	<.001
Ideation & hypothesis generation	5.3	5.4	5.2	6.0	5.6	5.0	5.0	5.4	.619
Administration	2.7	4.3	3.5	2.4	2.8	3.3	1.3	1.9	<.001
Data gathering	1.0	2.5	1.0	1.0	2.5	0.7	0.2	0.5	<.001
Research software & code development	13.8	13.3	19.9	15.9	20.6	14.6	5.8	15.9	<.001
Scientific analysis	16.7	9.3	8.8	14.3	11.6	8.0	33.7	18.7	<.001
Writing & communication	26.9	28.2	30.6	29.5	25.6	32.3	21.2	24.4	<.001
Peer review & service	2.1	1.6	1.6	3.6	1.5	3.0	1.9	1.6	.009
End-to-end autonomous research	0.4	0.2	0.1	0.2	0.3	0.3	0.9	0.2	<.001
Mentorship & teaching	3.1	2.1	1.9	3.0	2.7	4.8	4.1	2.5	<.001

Continued figure and shared notes.

whether this contributes to scientific progress depends on whether scientists can effectively apply newly learned knowledge to their research objectives.

Second, tasks with similarly high average message shares can differ substantially in how widely scientists use them. This matters because it means the benefits of current or improved AI capabilities in certain scientific tasks may not be shared broadly by scientists. "Develop or test proofs" and "Create research visuals" account for 4.2% and 4.2% of the average scientist's scientific messages, respectively, but 26.2% of scientists use the former at least 3 times, compared with 55.4% for the latter. More generally, developing or testing proofs, deriving formal results, and implementing research software rank among the top ten tasks by average message share, but are less widely used than comparably frequent tasks. These patterns suggest that AI may be broadly useful for helping scientists learn and communicate, while opportunities to leverage its technical capabilities depend more on scientists' research methods.

Panel A of Figure 4 compares, across all 7 scientific fields we study, the average distribution of scientists' scientific messages across the 10 scientific workflows.[16] For each workflow, we use Welch's ANOVA to test whether means across all scientific fields are equal, allowing variances to differ across fields (Welch, 1951). We adjust p-values using Holm's method across the 10 tests to deal with multiple hypothesis testing (Holm, 1979). Rejection establishes that at least two scientific fields have different average message shares, without

[16] Appendix Figure A4 presents the corresponding comparisons for individual scientific tasks.

Figure 4 (continued): AI use across scientific workflows by field and age

B. Scientist age

Color: % difference from each workflow's overall mean

–100% 0 +100% +200%

	Overall	18–29	30–39	40+	H_0: equal group means p-value
Reading & knowledge acquisition	27.7	29.8	27.1	26.7	.103
Ideation & hypothesis generation	5.3	5.3	5.1	5.8	.719
Administration	2.7	1.6	3.1	3.3	**<.001**
Data gathering	1.0	1.3	0.9	0.8	.257
Research software & code development	13.8	17.9	12.4	9.8	**<.001**
Scientific analysis	16.7	16.3	16.3	19.4	.163
Writing & communication	26.9	24.0	28.9	26.7	**.001**
Peer review & service	2.1	1.4	2.1	2.3	.146
End-to-end autonomous research	0.4	0.3	0.4	0.3	.719
Mentorship & teaching	3.1	1.6	3.4	4.4	**<.001**

Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. Overall includes all 1,819 scientists. Panel A compares 7 fields among 1,788 scientists, excluding 19 classified as Other and 12 with missing fields. Panel B compares 3 age groups among 1,640 scientists, excluding 179 with missing ages. Cells show workflow shares of scientific messages within each group. The workflow share of scientific messages is the mean of each scientist's share of all scientific messages devoted to that workflow. Other scientific work and messages without a workflow classification remain in the denominators. In each panel, Welch's ANOVA tests whether mean workflow shares are equal across the displayed groups. P-values are adjusted using Holm's method for multiple hypothesis testing across the ten workflow tests separately within each panel. Bold p-values are below .05 after adjustment. Workflows are ordered in the approximate order we believe they first occur in the scientific research process.

identifying which pairs of fields differ.

Ideation & hypothesis generation is the only workflow for which we do not reject equality of average message shares across fields. This suggests that support for developing research ideas is a broadly relevant application of AI across all scientific fields we study. Even among workflows where we reject equality, the same four workflows account for the largest message shares, consistent with Figure 3's findings.

Among workflows for which we find statistically significant differences in mean message shares across scientific fields, data gathering shows substantial relative differences in message shares. Its average message share is 2.5% in biology and 2.5% in earth and planetary sciences, compared with 0.2–1.0% elsewhere. One possible explanation is that these fields have large shared data resources, including sequencing databases and satellite archives, that offer fertile ground for new discoveries but can be technically difficult to explore (Gorelick et al., 2017; Edgar et al., 2022). To the extent that AI lowers these technical barriers, it could make research using existing data more productive, but might also narrow the range of ideas

scientists pursue by shifting attention away from those requiring new data collection. Use of AI for scientific analysis and research software & code development also differs across fields. Mathematicians have the highest average message share for scientific analysis (33.7%) and the lowest for research software & code development (5.8%).

These differences raise a question for further research. Given AI's current capabilities in software development and aspects of scientific analysis such as mathematical reasoning, how do differences in research methods across fields shape the potential gains in scientific progress, and what bottlenecks limit the magnitude of these gains in each field?

Panel B of Figure 4 presents the corresponding comparisons across age groups. We do not detect differences across age groups in average message shares for reading & knowledge acquisition or ideation & hypothesis generation. Among workflows involving more technical work, average message shares for research software & code development decline from 17.9% among scientists aged 18–29 to 9.8% among those aged 40 and older, while we do not detect age differences for scientific analysis.

Mentorship & teaching and administration show the opposite pattern. Their average message shares rose from 1.6% and 1.6% among scientists aged 18–29 to 4.4% and 3.3%, respectively, among those aged 40 and older. This may reflect fewer mentoring and administrative responsibilities among younger scientists. More generally, differences in message shares may reflect, among other factors, scientists' responsibilities, experience, and differences across fields in research methods, as well as how well AI supports their work. Usage data alone cannot distinguish between these explanations.

3.2 AI use for research ideation

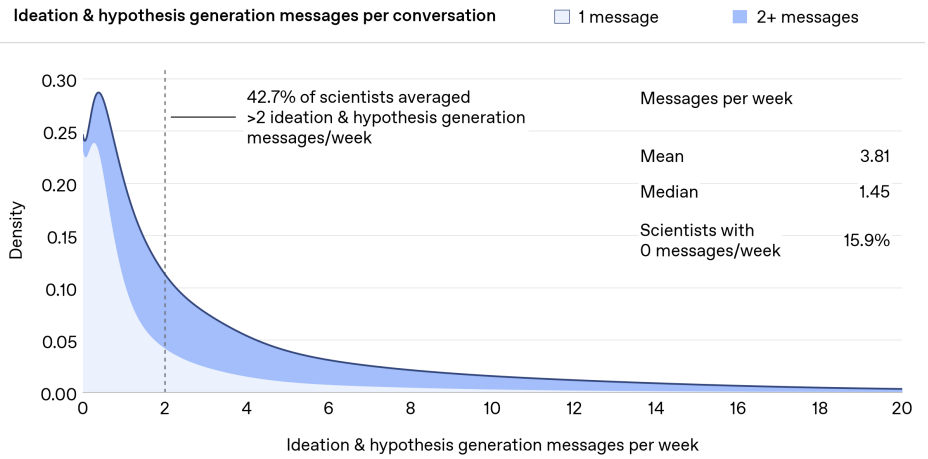
Economists have emphasized that the creation of new ideas is a fundamental driver of long-run economic growth (Romer, 1990; Jones, 2005). Because ideation affects what research questions scientists pursue and ultimately the direction of scientific progress, understanding how scientists use AI in this process is important. In this section, we quantify how the use of AI for ideation & hypothesis generation varies across scientists, examine how scientists divide work between themselves and AI when exploring ideas and developing research questions, and investigate how this division varies across scientific fields and age groups.

Figure 5 plots the distribution of the number of weekly ideation & hypothesis generation messages sent. It reveals substantial variation in how intensively scientists use AI for ideation & hypothesis generation. The majority of scientists average at most two messages per week in this workflow, but 42.7% average more than two, and 11% average more than ten. The distribution also has a long right tail; 2.4% of scientists average more than 20 messages per week, beyond the figure's plotted domain.

The shading in Figure 5 separates each scientist's contribution to the distribution according to the share of their ideation & hypothesis generation messages in conversations containing one versus multiple ideation & hypothesis generation messages.[17] Among scien-

[17] A conversation is a single chat thread in ChatGPT or Codex.

Figure 5: Distribution of weekly message rates for ideation and hypothesis generation



Notes: The sample comprises 28,317 scientific messages in Ideation & hypothesis generation, sent from August 20 to September 22, 2026, by 1,530 scientists. Each scientist's weekly rate divides their observed messages in this workflow by their observation time in weeks. The density plot shows scientist weekly message rates, with one observation per scientist. The distribution is shaded separately for conversations containing one versus at least two messages in this workflow, dividing each scientist's contribution according to the share of their ideation messages in each conversation type. We truncate the display at 20 messages per week; 2.4% of scientists exceed this rate. All 1,819 scientists contribute to the summary statistics.

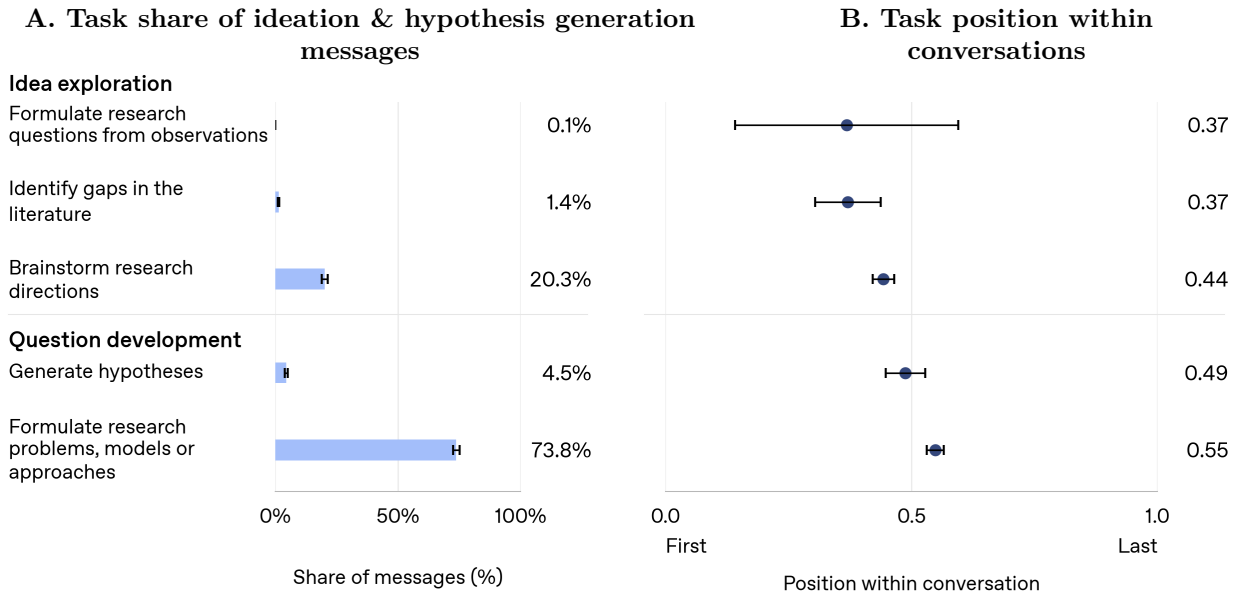
tists at the upper end of the distribution, most messages occur in conversations with multiple ideation & hypothesis generation messages. This pattern suggests that scientists who use AI intensively for ideation & hypothesis generation treat AI as a collaborator by engaging with it repeatedly within a conversation, rather than as a source of one-off suggestions.

Panel A of Figure 6 shows how the average scientist's ideation & hypothesis generation messages are distributed across 5 tasks. We group these into two conceptual stages: "idea exploration" and "question development."

We illustrate the task definitions with a hypothetical example about AI usage and scientific productivity. The "idea exploration" stage identifies promising questions and directions. **Identifying gaps in the literature** is defined as identifying gaps in existing literature to formulate novel research questions; a scientist's message might ask, "Which questions about AI's effect on scientific understanding are understudied?" **Brainstorming research directions** is defined as brainstorming potential research directions through discussions with colleagues and peers; here, AI serves as the brainstorming partner. An example AI output might suggest studying how AI use affects scientists differently depending on their experience. **Formulating questions from observations** uses observed real-world phenomena or patterns to formulate research questions or hypotheses; a scientist's message might ask, "In our lab, AI use seems to improve productivity more for experienced scientists than for junior scientists. What mechanisms could help explain this pattern?"

The "question development" stage comprises two tasks. **Generating hypotheses** involves generating hypotheses based on theoretical frameworks, observations, or preliminary

Figure 6: AI use across ideation and hypothesis generation tasks



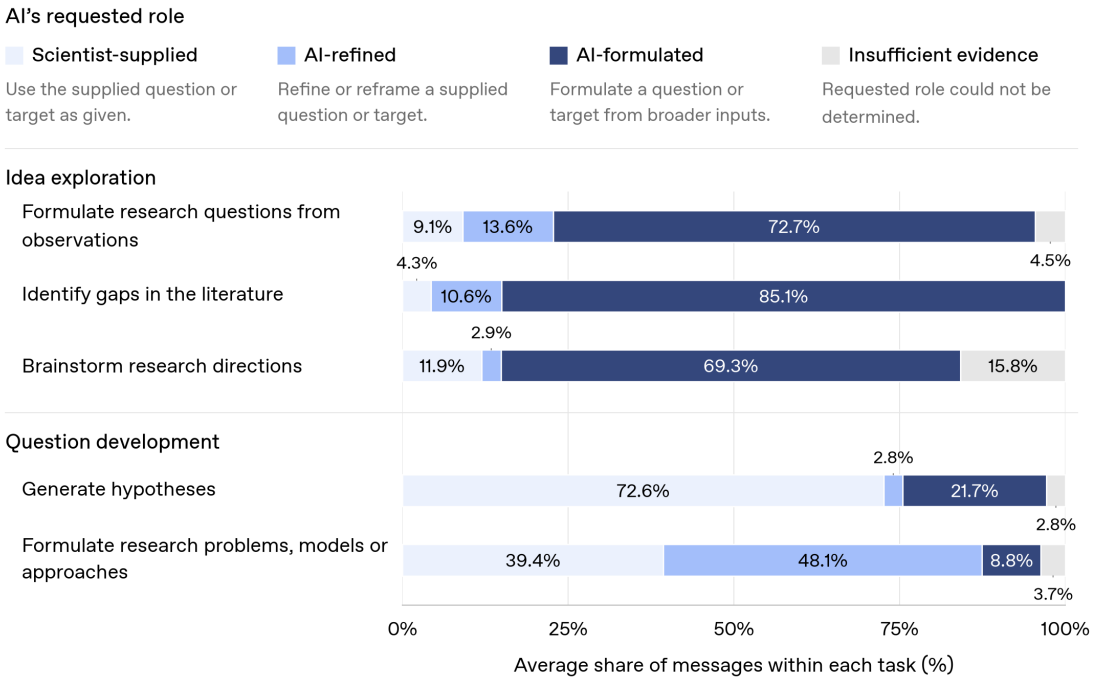
Notes: The sample comprises 28,319 scientific messages in the 5 listed Ideation & hypothesis generation tasks, sent from August 20 to September 22, 2026, by 1,530 scientists. We define two conceptual stages in the Ideation & hypothesis generation workflow: idea exploration and question development. Panel A shows task shares of scientific messages, with pointwise 95% scientist-level t confidence intervals. The task share of scientific messages is the mean of each scientist's share of scientific messages in the 5 listed tasks devoted to that task. Panel B shows the average relative position of each task in 2,245 conversations containing at least two distinct listed tasks, from 813 scientists. These conversations contain 12,753 messages in the 5 listed tasks, or 45.0% of messages in the 5 tasks. Within each conversation, messages in the 5 listed tasks are ordered chronologically and assigned evenly spaced relative positions from 0 for the first to 1 for the last. For each task, we average its messages' relative positions within each conversation, then across each scientist's conversations containing that task, and then across contributing scientists, with pointwise 95% scientist-level t confidence intervals.

data; an example message might be "Provide feedback on my hypothesis: AI increases research productivity more for experienced scientists because they know better how to evaluate its output." **Formulating problems, models, or approaches** involves formulating, reframing, or modifying a research problem, conjecture, theoretical model, or research approach; an example message might be "Can you give feedback on my model of how AI use affects research and suggest how I could incorporate an output evaluation stage?"

Most messages fall in the second stage: formulating research problems, models or approaches accounts for 73.8% of the average scientist's ideation & hypothesis generation messages. Brainstorming research directions is the next most common task, accounting for 20.3%.

Panel B of Figure 6 shows, for each of the 5 tasks, its average position within the conversation. We subset to conversations containing at least two different ideation & hypothesis generation tasks; this retains 45.0% of ideation & hypothesis generation messages. Within each conversation, we keep only ideation & hypothesis generation messages. To ensure position is invariant to conversation length, the first message has position zero, and the last has

Figure 7: AI's requested role in idea development by task



Notes: The sample comprises 28,319 scientific messages in the 5 listed Ideation & hypothesis generation tasks, sent from August 20 to September 22, 2026, by 1,530 scientists. We define two conceptual stages in the Ideation & hypothesis generation workflow: idea exploration and question development. Bars show the average shares of messages assigned to each of the three requested roles or Insufficient evidence within each task. Each category's share is calculated within each scientist's messages in the task, then averaged across scientists who sent at least one message in that task.

position one, with intervening messages spaced evenly between them. For each task, we average its message positions within each conversation, then across each scientist's conversations containing that task, and finally across scientists.

Panel B is consistent with our conceptual ordering of the tasks. Idea-exploration tasks have earlier average positions, between 0.37 and 0.44, than question-development tasks, between 0.49 and 0.55.

We next examine how scientists divide the work of formulating and refining research ideas between themselves and AI across the two stages of ideation: idea exploration and question development. Figure 7 shows the average scientist's distribution of messages by AI's role in idea development within each task, among scientists using that task. Recall that **scientist-supplied** means the scientist supplies the concrete scientific question, problem, claim, or target and does not ask AI to materially change it. **AI-refined** means AI is asked to materially narrow, broaden, redefine, or reframe a supplied question or target. **AI-formulated** means AI is asked to choose or formulate the question or target from a topic, evidence, goal, or broad area.

Two clear trends stand out. First, in the idea exploration stage, scientists ask AI to propose what to investigate, by relying on AI-formulated ideas. For the average scientist, 85.1%

of messages involving identifying gaps in the literature and 69.3% involving brainstorming research directions are classified as AI-formulated. Scientists thus delegate part of the search for promising research ideas and directions within a topic or literature to AI. This matters because it means the set of ideas scientists are exposed to in the idea exploration stage is determined, in part, by AI.

Second, in question development, scientists tend to supply the idea, and ask AI either to refine it or to help evaluate or investigate it. This represents a different division of work than in idea exploration. For the average scientist, 72.6% of hypothesis-generation messages are scientist-supplied, compared with 21.7% AI-formulated. In formulating problems, models or approaches, 48.1% of messages are AI-refined and 39.4% scientist-supplied. These patterns suggest that AI's potential influence on research agendas is not confined to the exploratory stage; AI's judgements about what makes a research question or model better may shape the scientific work pursued by the scientist.

AI's role in ideation raises two questions for further research. First, how does AI assistance affect the scientific value of individual research questions? Some potential drivers of increased scientific value could be reducing duplication with existing work or broadening the space of ideas scientists consider when exploring ideas. A concern could be that frequent deference to AI recommendations (Herdağdelen et al., 2026) may lead scientists to accept AI-proposed assumptions or revisions without fully assessing their impact (Messeri and Crockett, 2024). This behavior could negatively impact the question's scientific value.

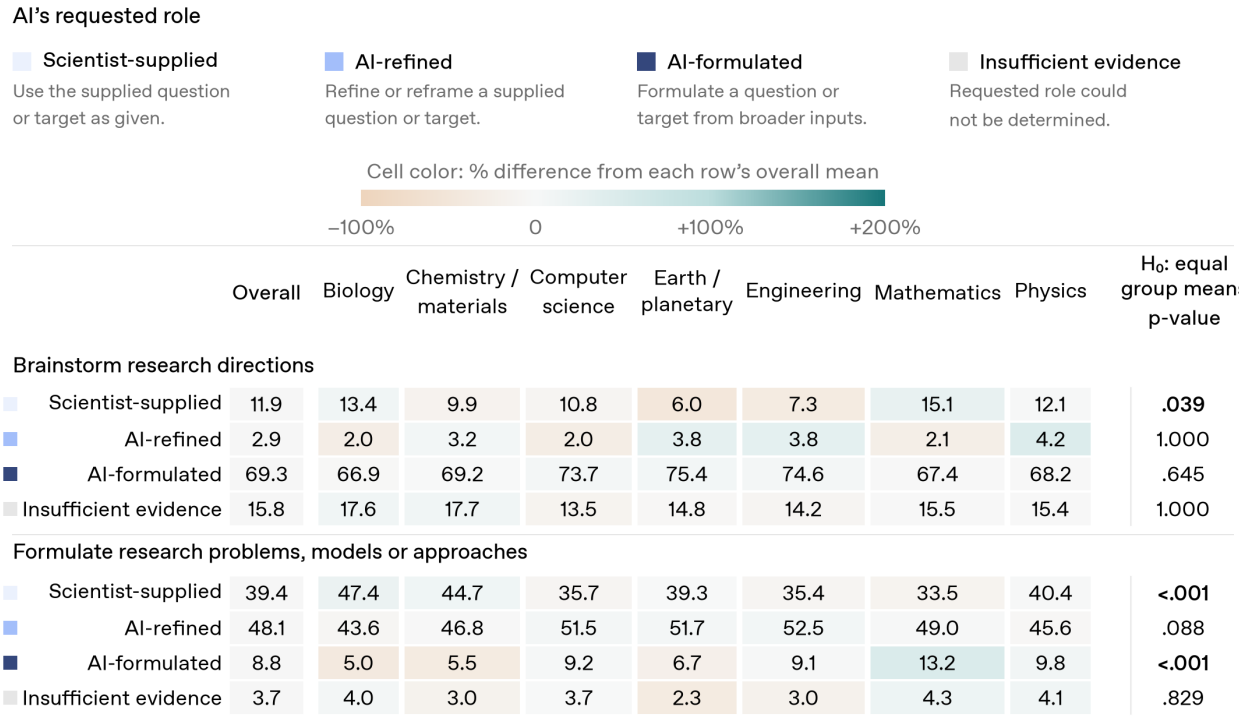
Second, how does AI's role in idea exploration and the refinement of research questions change the allocation of research effort across different scientific directions? AI could connect scientists to unfamiliar fields and broaden the directions they consider. However, if its suggestions and assessments steer scientists toward the same set of problems, it could concentrate research effort at the expense of other important directions. This concern is motivated by empirical evidence associating AI use with a narrower collective research agenda (Hao et al., 2026).

Panel A of Figure 8 compares, across all 7 scientific fields we study, the average distribution of scientists' messages across AI's requested roles within "Brainstorm research directions" and "Formulate research problems, models or approaches", the two tasks with the highest usage in each stage. As in Figure 4, we use Welch's ANOVA to test whether means across fields are equal and adjust p-values using Holm's method. Across fields, the broad patterns are consistent with Figure 7: most brainstorming messages are AI-formulated, while most formulation messages are scientist-supplied or AI-refined. Nevertheless, we find that across fields, we can reject equality for the share of scientist-supplied messages in "Brainstorm research directions." Similarly, we can reject equality for both the scientist-supplied and AI-formulated message shares in "Formulate research problems, models or approaches."

One field where these shares differ is mathematics. Mathematics is also of particular interest given recent results produced by AI on difficult open research problems, including a proposed resolution of the Navier–Stokes problem (OpenAI, 2026c,d). Among the 7 fields, mathematicians actually have the highest average share of scientist-supplied messages in

Figure 8: AI's requested role in selected ideation tasks by field and age

A. Scientific field



	Overall	Biology	Chemistry / materials	Computer science	Earth / planetary	Engineering	Mathematics	Physics	H_0: equal group means p-value
Brainstorm research directions									
Scientist-supplied	11.9	13.4	9.9	10.8	6.0	7.3	15.1	12.1	**.039**
AI-refined	2.9	2.0	3.2	2.0	3.8	3.8	2.1	4.2	1.000
AI-formulated	69.3	66.9	69.2	73.7	75.4	74.6	67.4	68.2	.645
Insufficient evidence	15.8	17.6	17.7	13.5	14.8	14.2	15.5	15.4	1.000
Formulate research problems, models or approaches									
Scientist-supplied	39.4	47.4	44.7	35.7	39.3	35.4	33.5	40.4	**<.001**
AI-refined	48.1	43.6	46.8	51.5	51.7	52.5	49.0	45.6	.088
AI-formulated	8.8	5.0	5.5	9.2	6.7	9.1	13.2	9.8	**<.001**
Insufficient evidence	3.7	4.0	3.0	3.7	2.3	3.0	4.3	4.1	.829

Continued figure and shared notes.

brainstorming research directions, at 15.1%. In contrast, in formulating research problems, models or approaches, they have the lowest scientist-supplied share, at 33.5%, and the highest AI-formulated share, at 13.2%. In mathematics, where AI capabilities have contributed most visibly to research breakthroughs, scientists contribute more of their own ideas upstream in ideation & hypothesis generation, but delegate more idea formulation to AI in the downstream process of turning ideas into research questions.

Panel B of Figure 8 presents the corresponding comparisons across age groups. We do not detect statistically significant age-group differences in requested-role shares for these two ideation tasks.

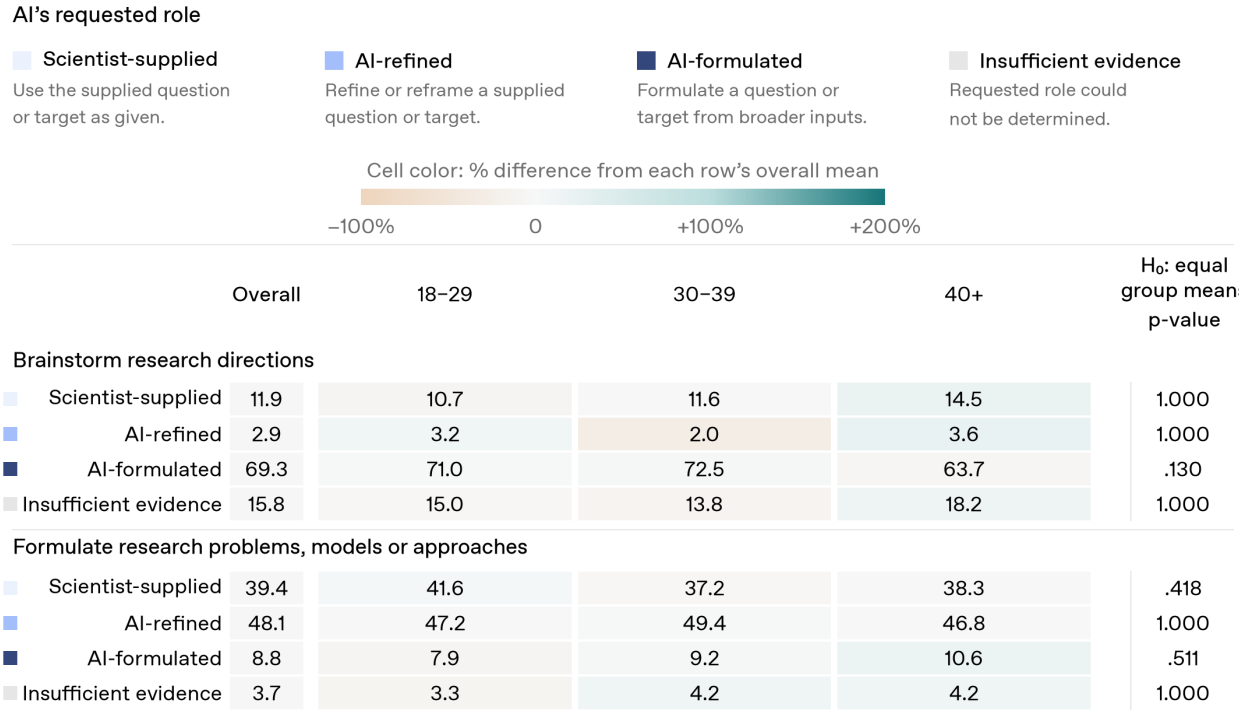
3.3 Scientist involvement in research execution

Turning ideas into scientific progress requires developing and testing them. AI can help with this work, but the time scientists spend checking and correcting its outputs can limit productivity gains. In our survey, 45.4% of scientists selected "time needed to check or correct outputs" as a difficulty in obtaining value from ChatGPT, making it the most common response. Codreanu et al. (2026) likewise find that verification absorbs part of the time scientists save through AI use.

We first define a measure of scientist involvement based on responses to AI output. We then examine how involvement varies across scientific workflows, task required expertise and

Figure 8 (continued): AI's requested role in selected ideation tasks by field and age

B. Age



	Overall	18–29	30–39	40+	H_0: equal group means p-value
Brainstorm research directions					
Scientist-supplied	11.9	10.7	11.6	14.5	1.000
AI-refined	2.9	3.2	2.0	3.6	1.000
AI-formulated	69.3	71.0	72.5	63.7	.130
Insufficient evidence	15.8	15.0	13.8	18.2	1.000
Formulate research problems, models or approaches					
Scientist-supplied	39.4	41.6	37.2	38.3	.418
AI-refined	48.1	47.2	49.4	46.8	1.000
AI-formulated	8.8	7.9	9.2	10.6	.511
Insufficient evidence	3.7	3.3	4.2	4.2	1.000

Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. Overall includes 5,544 messages from 1,024 scientists in Brainstorm research directions and 21,213 messages from 1,449 scientists in Formulate research problems, models or approaches. Panel A compares 7 fields, including 1,004 and 1,427 scientists in the two tasks, respectively, after excluding Other and missing fields. Panel B compares 3 age groups, including 933 and 1,300 scientists, respectively, after excluding missing ages. Cells show the average shares of messages assigned to each of the three requested roles or Insufficient evidence within each task and group. Each category's share is calculated within each scientist's messages in the task, then averaged across scientists who sent at least one message in that task. In each panel, Welch's ANOVA tests whether each category's mean share within each task is equal across the displayed groups. P-values are adjusted using Holm's method for multiple hypothesis testing across eight tests—four displayed categories for each of two tasks—separately within each panel. Bold p-values are below .05 after adjustment.

expected duration, and scientific fields and age groups. Finally, we examine demand for end-to-end autonomous research.

3.3.1 Measuring scientist involvement

Conceptually, we want our measure of involvement to capture how often scientists intervene after asking AI to execute a research task. Scientists may intervene for several reasons, such as to seek a clearer explanation of the output, ask AI to check the correctness of its work, or request changes to the work. Our involvement score captures this by measuring how often scientists make three types of follow-up requests: clarification, checking and revision. We interpret scientist involvement as evidence that ChatGPT did not execute the research task to the scientist's desired standards.

Clarification seeks explanation or understanding of prior AI output; checking questions or checks that output without identifying a specific correction; and revision requests a change, fix, or replacement. Table 2 in Section 2.1.2 defines all follow-up types.[18] For example, follow-ups to an AI-computed regression estimate like "Explain how you obtained this estimate" and "Correct the calculation to exclude duplicate observations" count as involvement because they seek to clarify or revise AI output. "Now put this estimate in the draft" and "Make this into a pie chart" do not count because they ask AI to extend its output.

We analyze initial requests assigned to one of the 40 scientific tasks. To ensure we have seven complete days to observe follow-ups, we include only requests sent from August 20 to September 15, 2026—seven days before the end of the sample period. We evaluate the next observed message classified as a scientific task in the same conversation, skipping messages without a scientific-task classification. For scientist i's request j in task or workflow t, the involvement indicator R_{ijt} equals one if the next scientific message in the same conversation arrives before the end of the 7-day window and requests clarification, checking, or revision of the original output. It equals zero otherwise, including when no subsequent scientific message arrives within the window. Requests for which involvement cannot be determined are excluded. Scientist i's involvement score is the mean of this indicator across their n_{it} initial requests. The task or workflow involvement score, $\overline{R}_t$, averages these scores equally across the N_t scientists with at least one such initial request:

$$\overline{R}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(\frac{1}{n_{it}} \sum_{j=1}^{n_{it}} R_{ijt} \right). \tag{1}$$

We report the scores in equation (1) as percentages. One limitation of our measure is that it will only capture a subset of all subsequent engagement between the scientist and AI output because scientists may refine outputs outside of the AI application.[19]

3.3.2 Scientist involvement across scientific work

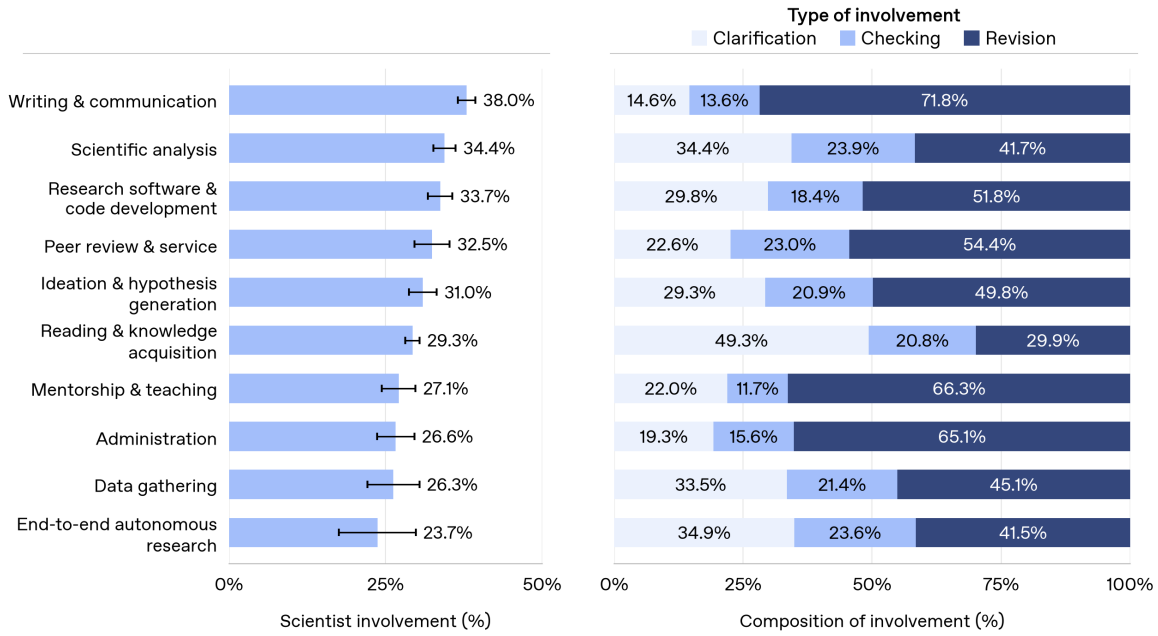
Panel A of Figure 9 lists each scientific workflow, in descending order of scientist involvement score.[20] Across all workflows, the average scientist's involvement score is 31.8%. Panel B of Figure 9 shows the average scientist's mix of clarification, checking, and revision among requests with involvement. Writing & communication has the highest involvement score (38.0%), followed by scientific analysis (34.4%) and research software & code development

[18] We exclude two follow-up types from what it means for scientists to be involved: extension/application, which suggests the scientist considers the output good enough to build on, and reaction, which alone does not request further work on the task.

[19] Since scientists may respond to an output after their next scientific message, we consider an alternative involvement measure that counts clarification, checking, or revision of the original output in any subsequent scientific message in the same conversation within 7 days of the initial request. Restricting both measures to requests for which both can be determined, the resulting task-level involvement scores have a Pearson correlation of 0.977 across the 40 tasks.

[20] Appendix Table A5 ranks all 40 scientific tasks by involvement score.

Figure 9: Scientist involvement across scientific workflows



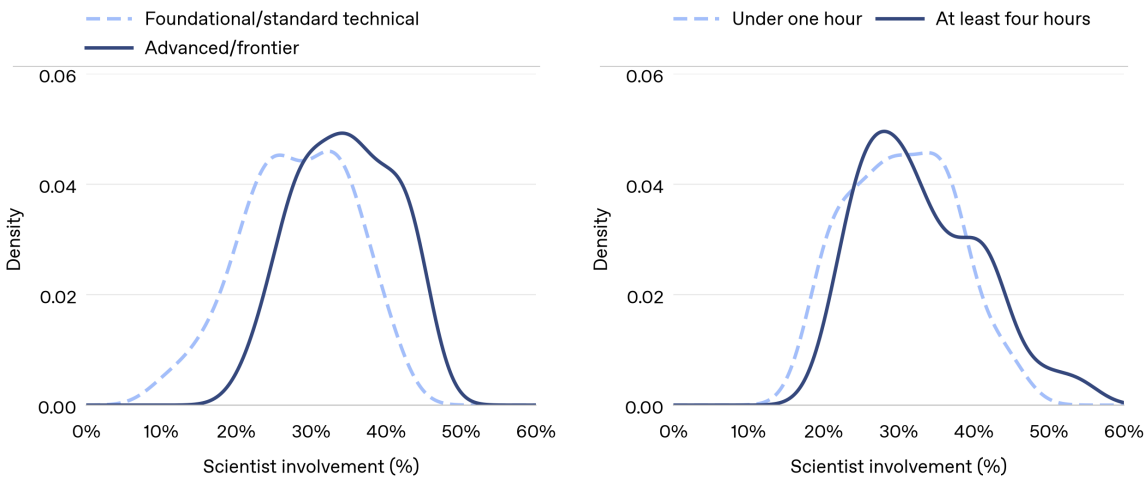
Notes: The sample comprises 125,947 initial requests sent from August 20 to September 15, 2026, by 1,809 scientists. An initial request starts scientific work without relying on prior AI output. We include requests assigned to one of the 40 scientific tasks, with 7 complete days of follow-up and sufficient information to determine involvement. A scientist's involvement score is the percentage of these requests whose next scientific-task message in the same conversation arrives within 7 days and seeks clarification, checking, or revision of the original output. Panel A shows the average involvement score within each workflow among scientists with at least one such initial request in that workflow, with pointwise 95% scientist-level t confidence intervals. Panel B shows the composition of involvement. Clarification, checking, and revision shares are calculated within each scientist's requests with involvement in that workflow, then averaged across scientists with at least one such request in the workflow.

(33.7%). High scientist involvement in writing & communication suggests that scientists remain important in ensuring that the outcomes of scientific research are clearly communicated. Clearly written scientific research is important for scientific progress because it makes it easier for other researchers to assess and build on it, broadening the impact of new discoveries. On average across scientists, revision accounts for most writing & communication follow-ups involving clarification, checking, or revision. This is consistent with broader concerns about the quality of AI-generated scientific prose (Gartenberg et al., 2026).

Scientists also remain heavily involved in scientific analysis, although Panel B shows that scientists' follow-up requests involve more clarification and checking, and less revision of AI output. Mathematics offers an example of why this may be the case—currently, even after an AI model produces a correct mathematical proof, substantial human work may still be needed to clarify its key insights and connect the result to existing theory so that other researchers can understand, evaluate, and build on it (OpenAI, 2026d; Tao, 2026). Scientists likewise remain heavily involved in software development, as their continued requests for

Figure 10: Distributions of task-level scientist involvement scores

A. Required expertise **B. Expected task duration without AI**



Notes: The sample comprises 125,947 initial requests sent from August 20 to September 15, 2026, by 1,809 scientists. An initial request starts scientific work without relying on prior AI output. We include requests assigned to one of the 40 scientific tasks, with 7 complete days of follow-up and sufficient information to determine involvement. A scientist's involvement score is the percentage of these requests whose next scientific-task message in the same conversation arrives within 7 days and seeks clarification, checking, or revision of the original output. Panel A compares general/foundational or standard technical/professional expertise with advanced research or frontier/subfield specialist expertise. Panel B compares expected duration under one hour with four hours or more, measured as the active time a competent, experienced person would need to complete the requested task without AI. Each density plot shows the distribution of task-level average involvement scores. Each task-level score averages scientists' involvement scores among scientists with at least one eligible initial request in that task and comparison group. A task contributes only if both groups include at least 10 scientists: 32 tasks in Panel A and 34 in Panel B.

explanations, checks, and revisions suggest that initial outputs do not always meet their standards.

Notably, the three workflows with the greatest scientist involvement are also areas where AI capabilities have attracted substantial attention. Scientists' continued requests to revise, understand, and check AI outputs in those workflows suggest that scientific judgement continues to be important in ensuring AI outputs are useful for research. As Panel B shows, this involvement extends beyond revision: the clarity or verifiability of AI output is also not up to scientists' standards, and making AI output more broadly useful for science requires improving all three aspects.

We next examine what task characteristics predict increased scientist involvement. This matters because it informs where scientific judgement or skill that is particularly suited to dealing with certain task characteristics may be valuable, as inferred from when scientists find AI output not up to par.

We focus on two scientific task characteristics: required expertise and expected duration, as defined in Table 2 of the Data section. A scientific task's required expertise describes the minimum preparation needed to complete a request. Its four levels are general/foundational,

Figure 11: Scientist involvement across workflows by field and age

A. Scientific field

Color: percentage difference from Overall

−75% 0 +75%

	Overall	Biology	Chemistry / materials	Computer science	Earth / planetary	Engineering	Mathematics	Physics	H_0: equal group means p-value
Overall	31.8	30.4	30.2	32.6	30.7	31.1	31.7	34.7	.251
Writing & communication	38.0	35.8	40.8	38.0	41.2	34.2	37.6	40.2	.884
Scientific analysis	34.4	34.8	31.5	33.7	32.6	35.6	32.2	38.7	1.000
Research software & code development	33.7	37.5	30.4	32.7	33.7	32.8	31.2	34.8	1.000
Peer review & service	32.5	37.2	31.6	27.2	28.7	31.8	32.0	34.7	1.000
Ideation & hypothesis generation	31.0	26.0	31.7	31.8	40.4	34.0	27.5	33.7	.884
Reading & knowledge acquisition	29.3	27.0	27.9	30.2	25.4	32.1	29.6	30.8	.884
Mentorship & teaching	27.1	18.6	25.8	26.2	28.2	26.5	33.2	25.9	.884
Administration	26.6	26.2	22.4	27.0	22.7	27.5	25.6	31.5	1.000
Data gathering	26.3	23.8	32.5	23.3	27.3	30.0	14.6	30.5	1.000
End-to-end autonomous research	23.7	17.6	19.0	19.1	8.3	32.4	23.3	27.3	1.000

Continued figure and shared notes.

standard technical/professional, advanced research, and frontier/subfield specialist. Panel A of Figure 10 plots the distribution of scientific task-level involvement scores for requests requiring advanced research or frontier/subfield specialist expertise and those requiring general/foundational or standard technical/professional expertise. Within tasks, involvement is higher for requests requiring more expertise in 28 of 32 tasks (87.5%), with a mean difference of 6.5 percentage points. Scientists follow up more often when requests require greater expertise, suggesting that AI output fails to meet scientists' standards more often in advanced research tasks. This highlights the continued value of scientific expertise.

As users increasingly delegate longer tasks to AI agents (Johnston et al., 2026), an important question is whether these requests also require more scientist involvement. Expected duration measures the active time an experienced person would need to complete the requested task without AI. Panel B of Figure 10 plots the distribution of task-level involvement scores for requests expected to take at least four hours and those expected to take under one hour. Within tasks, involvement is higher for longer requests in 22 of 34 tasks (64.7%), with a mean difference of 2.6 percentage points. These differences are smaller than those for required expertise. Appendix Table A6 reports, when possible, all task-level involvement scores by the required expertise and expected duration splits.

Panel A of Figure 11 compares scientist involvement within each workflow across the 7 scientific fields we study. As in Figure 4, we use Welch's ANOVA to test whether group

Figure 11 (continued): Scientist involvement across workflows by field and age

B. Scientist age

Color: percentage difference from Overall

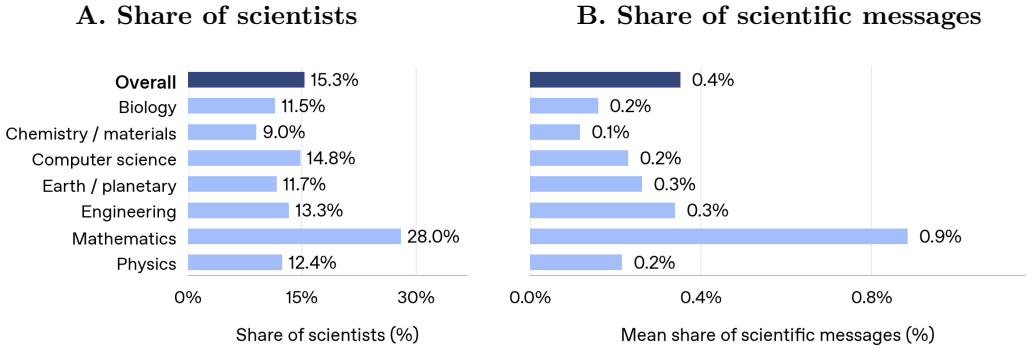
−75% 0 +75%

	Overall	18–29	30–39	40+	H_0: equal group means p-value
Overall	31.8	33.4	31.9	29.8	.146
Writing & communication	38.0	39.9	38.6	34.0	.179
Scientific analysis	34.4	34.9	33.6	36.0	1.000
Research software & code development	33.7	35.8	33.4	29.3	1.000
Peer review & service	32.5	35.0	31.8	33.6	1.000
Ideation & hypothesis generation	31.0	33.1	30.6	26.9	1.000
Reading & knowledge acquisition	29.3	32.1	28.6	25.6	**.006**
Mentorship & teaching	27.1	26.4	27.9	25.0	1.000
Administration	26.6	28.8	25.4	27.3	1.000
Data gathering	26.3	24.9	25.5	32.7	1.000
End-to-end autonomous research	23.7	21.4	23.8	31.1	1.000

Notes: The sample comprises 125,947 initial requests sent from August 20 to September 15, 2026, by 1,809 scientists. An initial request starts scientific work without relying on prior AI output. We include requests assigned to one of the 40 scientific tasks, with 7 complete days of follow-up and sufficient information to determine involvement. A scientist's involvement score is the percentage of these requests whose next scientific-task message in the same conversation arrives within 7 days and seeks clarification, checking, or revision of the original output. Panel A compares 7 fields, excluding Other and missing fields. Panel B compares 3 age groups, excluding missing ages. Cells show average involvement scores among scientists with at least one such initial request in the relevant workflow and group. The Overall row first pools these initial requests across workflows within each scientist. The Overall column includes contributing scientists with Other or missing fields or missing ages. The contributing sample varies by workflow and group. In each panel, Welch's ANOVA tests whether mean involvement scores are equal across the displayed groups. P-values are adjusted using Holm's method for multiple hypothesis testing across eleven tests—the Overall row and ten workflows—separately within each panel. Bold p-values are below .05 after adjustment.

means are equal and adjust p-values using Holm's method. We do not detect statistically significant differences across fields in any of the 10 workflows tested. Panel B presents the corresponding comparisons across age groups. Reading & knowledge acquisition is the only one of the 10 tested workflows where we reject equality, and we observe that involvement declines from 32.1% among scientists aged 18–29 to 28.6% among those aged 30–39 and 25.6% among those aged 40 and older. Appendix Figure A5 reports involvement scores and the composition of involvement by scientific field and age.

Figure 12: Requests for end-to-end autonomous research with AI



Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. End-to-end autonomous research requests coordinate Reading & knowledge acquisition, Ideation & hypothesis generation, and either Data gathering and Empirical analysis or Theoretical and formal analysis, including progression between stages. Overall includes all 1,819 scientists. The field comparison includes 1,788 scientists, excluding 19 classified as Other and 12 with missing fields. Panel A shows the share of scientists using AI for this workflow: the percentage who sent at least one scientific message in the workflow during August 20–September 22, for Overall and each field. Panel B shows the workflow share of scientific messages, for Overall and each field. The workflow share of scientific messages is the mean of each scientist's share of all scientific messages devoted to that workflow.

3.3.3 End-to-end autonomous research

Advances in AI have increased interest in the concept of automating scientific research, in which AI coordinates a research project across literature review, hypothesis generation, experimentation, and analysis with limited scientist direction (Lu et al., 2026). Our involvement findings suggest a potential challenge for this ambition: for the average scientist, 31.8% of initial requests prompt clarification, checking, or revision of AI output. A full research project involves many such requests, so full automation would mean problems scientists would normally flag may go unaddressed and compound.

Nevertheless, to better understand how AI might be changing what scientific workflows are feasible, we examine the demand by scientists to use AI for end-to-end autonomous research. We classify a message as requesting end-to-end autonomous research when it asks AI to coordinate reading & knowledge acquisition, ideation & hypothesis generation, and some scientific analysis work, along with any data gathering required.

Panel A of Figure 12 shows that around 15% of scientists sent at least one such request, with the largest share in mathematics. Yet these requests account for just 0.4% of the average scientist's scientific messages, as shown in Panel B. Furthermore, low weekly return rates in Appendix Figure A2 suggest limited sustained use. These patterns indicate that scientists are experimenting with using AI for end-to-end autonomous research, but have not made it a regular part of their ChatGPT use.

These findings raise two questions for further research. First, how does increased scientist involvement affect the quality of research output, and what are the costs of increased involvement? This would help improve our understanding of where scientists could improve research output quality by being more involved, and where they could delegate more to AI.

Second, as model capabilities advance, which scientific tasks will continue to benefit from scientist involvement, and how will the scientist's role in scientific research change more broadly?

4 AI and scientific productivity: Survey evidence

Although AI usage data help us understand how scientists use AI, the larger question of interest is how AI will contribute to scientific progress. To study scientists' expectations, we survey them about the extent to which they expect scientific discovery to accelerate and the mechanisms by which this might occur. We also ask about the productivity gains they attribute to AI and the bottlenecks they face that limit their research productivity. Appendix B describes the survey in more detail.

4.1 Expectations for scientific discovery

To elicit scientists' expectations around the potential of AI to accelerate research, we asked scientists to consider a hypothetical discovery in their field that, without AI, would have been made in 10 years. Then we asked them to consider how much earlier they expected this discovery would now be made with AI.

Panel A of Figure 13 plots the cumulative share of scientists by their expected acceleration in discovery time, from largest to smallest. Nearly all respondents (97.6%) expect AI to accelerate discovery. 21.6% expect the hypothetical discovery to occur seven to ten years sooner, 72.0% expect it at least three years sooner, and 91.2% expect it at least one year sooner.

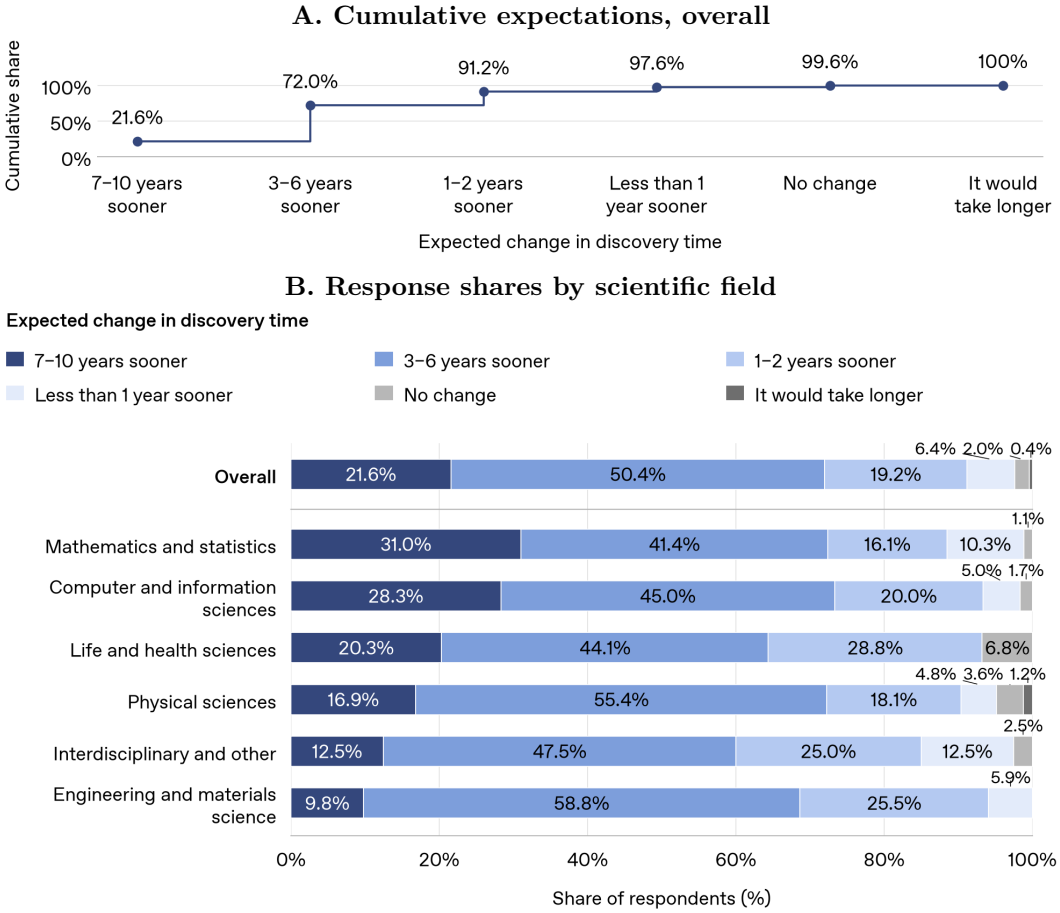
Panel B of Figure 13 shows the share of scientists selecting each response option, overall and by scientific field. Fields are ordered by the share expecting a discovery 7–10 years sooner. 31% of mathematicians and statisticians believe that discoveries that would have taken 10 years will now be made within the next three years, compared with just 10% in engineering and materials. Despite these differences, more than half of scientists in every field expect AI to enable the hypothetical discovery at least three years sooner.

We next examine how scientists expect AI to accelerate scientific progress. Figure 14 shows the distribution of claims extracted from scientists' written responses across broad and detailed mechanism classifications. The broad categories are faster research execution, improved search for ideas, expanded research directions, and other or unspecified.

Of the two aspects of scientific work examined in this paper—ideation and execution—execution features more prominently in the reported mechanisms. 52% of claims identify faster research execution as a mechanism for accelerating scientific progress. However, scientists still expect AI to affect the idea production function: 23% of claims concern improving the search for promising ideas, and 14% concern enabling scientists to pursue a broader range of research questions.

Figure 13: Expected acceleration of scientific discovery with AI

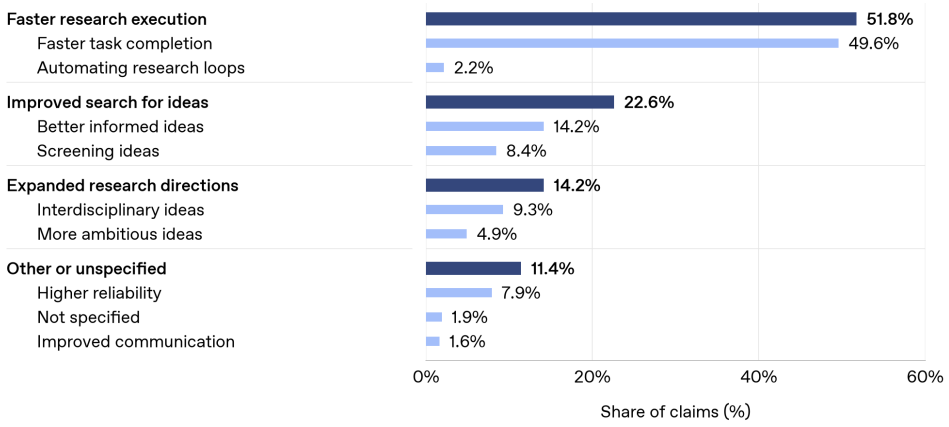
Survey question Q1: Imagine an important discovery in your field that would take 10 years without AI. How much sooner, if at all, do you believe it would be achieved with AI?



Notes: The survey was conducted September 30–October 7, 2026. The sample comprises 250 scientists who answered how AI would affect the timing of an important discovery that would take ten years without AI. Panel A shows the cumulative percentage selecting the displayed response option or an earlier option on the horizontal axis. Panel B shows the percentage selecting each response option within Overall and each field. Scientists could select multiple fields and therefore contribute to more than one field. Life sciences and Medicine and health sciences are combined as Life and health sciences. Earth, environmental, and planetary sciences; Agricultural and veterinary sciences; Social and behavioral sciences; Humanities and arts; Business and marketing; Interdisciplinary or multidisciplinary research; and Other are combined as Interdisciplinary and other. Field percentages use scientists who answered the discovery question and selected the corresponding field: 87 in Mathematics and statistics, 60 in Computer and information sciences, 59 in Life and health sciences, 83 in Physical sciences, 40 in Interdisciplinary and other, and 51 in Engineering and materials science.

The detailed categories show that faster task completion accounts for 49.6% of all claims, while automating research loops accounts for only 2.2%. This is consistent with the limited sustained use of end-to-end autonomous research documented in Section 3.3.3.

Figure 14: Reported mechanisms for accelerating scientific discovery with AI
Survey question Q2: What are the main ways you expect AI to accelerate discoveries in your field?



Notes: The survey was conducted September 30–October 7, 2026. The survey yielded 224 written responses from scientists who expected AI to accelerate an important discovery. The mechanism analysis contains 367 distinct claims. We use GPT-6 Astra to identify and classify distinct claims as described in Appendix B.2. A response may contribute multiple claims. Bars show each category's share of all claims. Bold rows show the four broad groups and aggregate their indented categories.

4.2 Self-reported productivity gains

We also use the survey to examine how extensively scientists use AI across scientific workflows and the productivity gains they attribute to it. Figure 15 plots the average reported share of time spent using AI against the average self-reported productivity ratio for each of 8 scientific workflows. This ratio is the estimated time required without AI divided by the time required with AI for the same work at the same quality.[21] Workflows with greater reported AI use also have larger reported productivity gains, with a correlation of 0.96 across workflow averages. This pattern is consistent with scientists choosing to use AI more in workflows where they perceive greater productivity benefits. Mean reported productivity ratios range from 1.35× in administrative work to 2.83× in research software & code development. Several scientists we interviewed also emphasized how AI's coding capabilities had opened new research opportunities, describing how they used ChatGPT to build specialized research tools and enable data collection that would otherwise have been impractical. These accounts suggest that productivity gains from AI-assisted software development can both accelerate existing work and expand the scope of feasible research.

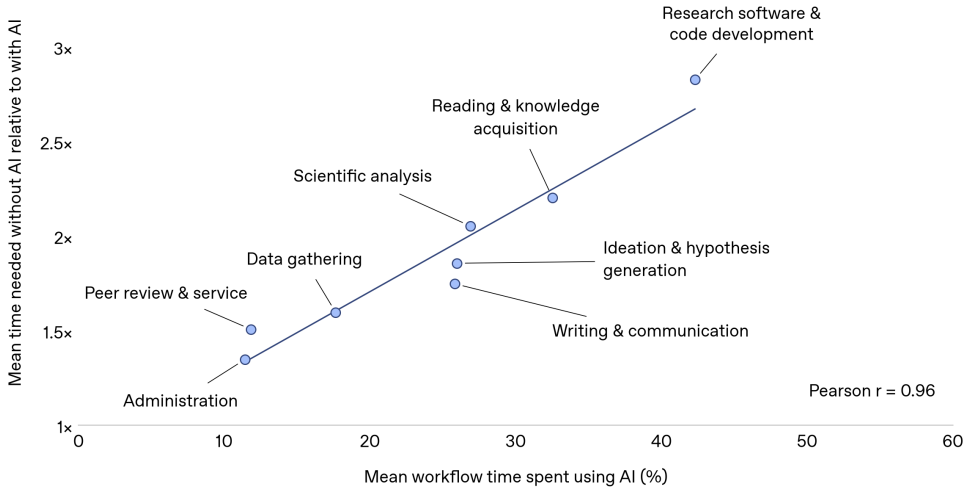
In Section 3.1, we showed that data gathering, peer review & service, and administration account for small shares of scientific messages and discussed how constraints in these workflows could limit how gains elsewhere translate into scientific progress. Our survey provides complementary evidence, showing that scientists report using AI for only 17.6%, 11.8%, and 11.4% of their total time doing work in these workflows, respectively. Hence, the AI-enabled productivity gains in the other scientific workflows documented above may only have limited

[21] The survey did not ask about mentorship & teaching or end-to-end autonomous research.

Figure 15: AI use and self-reported productivity gains across scientific workflows

Survey question Q4: Approximately what percentage of the total time you spent on each task involved actively using AI tools in the past 30 days?

Survey question Q5: Thinking about the research activities you used AI for in the past 30 days, how much more or less time would you have needed without AI to complete the same amount of work at the same quality?



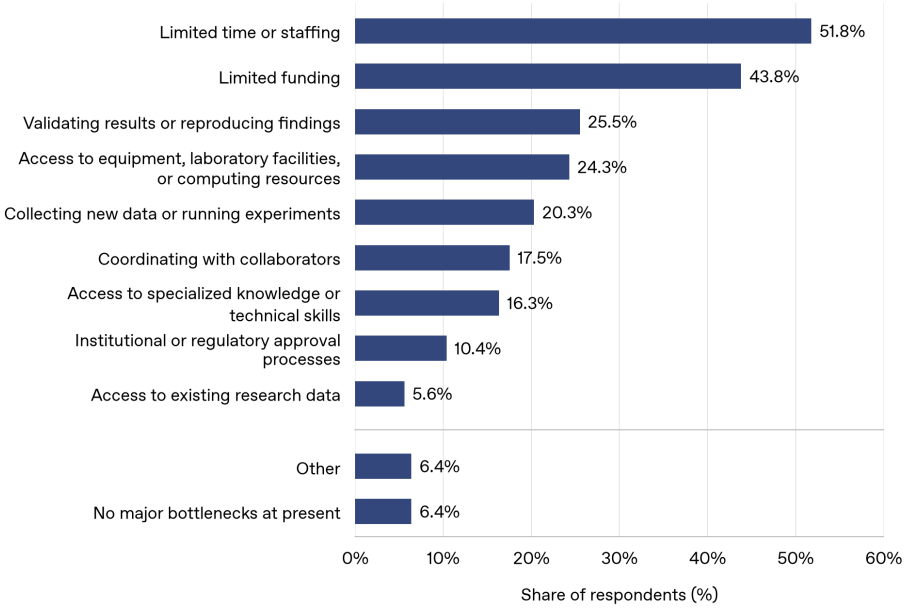
Notes: The survey was conducted September 30–October 7, 2026. The figure uses responses about research during the past 30 days, with 249–251 scientists contributing paired values per workflow. Each point represents one of 8 workflows. The horizontal axis shows the average share of total workflow time spent actively using AI. Response options were 0%, 1–5%, 6–25%, 26–50%, 51–75%, 76–95%, and 96–100%; each range is converted to its midpoint. The vertical axis shows the average self-reported productivity ratio: time required without AI divided by time required with AI for the same work at the same quality. Response options were less time, the same time, or 10%, 25%, 50%, 100%, 200%, 300%, or 400% or more additional time. Additional-time percentages are converted to ratios by adding one to the percentage divided by 100; less time is coded as 0.5, the same time as 1, and 400% or more as 5. Scientists reporting no AI use in the screening question, or not selecting a workflow, receive an AI-time share of zero and a productivity ratio of one for that workflow. Missing answers for selected workflows remain missing. We retain paired values within scientist and workflow, then average available paired values for Empirical analysis and Theoretical and formal analysis within scientist to form Scientific analysis. Both coordinates average responses from the same contributing scientists within each workflow. The plotted line is an ordinary-least-squares fit, and the reported correlation is Pearson's correlation across the 8 workflow points.

effects on scientific progress unless scientists can also use AI to accelerate complementary workflows like data gathering, peer review & service, and administrative work.

These results have two limitations. First, respondents are a selected group of scientists who applied for expanded AI access and chose to answer a survey about AI use. The findings may therefore not generalize to the broader population of scientists. Second, the correlation is based on only 8 workflow averages. Hence, the extent to which these productivity gains contribute to broader scientific progress may depend on the complementary workflows discussed above.

Figure 16: Reported constraints on scientific research

Survey question Q6: What currently limits progress in your research, even with the tools available to you? Select up to 3.



Notes: The survey was conducted September 30–October 7, 2026. The sample comprises 251 scientists. Bars show the percentage selecting each displayed response option. Scientists could select up to three options; "No major bottlenecks at present" could only be selected alone.

4.3 Research bottlenecks

Finally, we asked scientists about the bottlenecks that still limit progress in their research, even with the AI tools available to them. Figure 16 reports the share of scientists who selected each bottleneck. Limited time or staffing (51.8%) and funding (43.8%) are the most common, and 71% of scientists selected at least one of these two constraints. Only 6.4% report no major bottlenecks. These findings suggest that fully realizing AI's potential to accelerate scientific progress may also require broader investment in science (Brynjolfsson et al., 2021; Flanagan et al., 2023).

5 Conclusion

We combine scientists' AI usage data with survey evidence to study AI's role in scientific research. Scientists use AI to develop ideas and execute research tasks while remaining involved in determining research questions and assessing AI outputs. AI use is concentrated in four scientific workflows: reading, writing, scientific analysis, and software development. Across workflows, greater reported AI use is associated with larger self-reported productivity gains by scientists. AI usage intensity also varies substantially across scientists.

Nearly all surveyed scientists expect AI to accelerate scientific discovery, but the extent to which workflow-level productivity gains translate into broader scientific progress will depend

in part on two factors. The first is the effectiveness and scope of AI assistance. The time needed to check or correct outputs is the most commonly reported difficulty in obtaining value from AI, and respondents report relatively limited AI use in scientific workflows like data gathering, peer review & service, and administration. The second is the availability of complementary research resources. Limited time or staffing and funding remain respondents' most commonly reported research constraints, and improvements in model capabilities alone may not resolve them. Realizing AI's potential to accelerate discovery may therefore require both advances that make AI more capable at helping scientists execute scientific work and complementary investment in science more broadly.

Our findings also motivate broader questions about how AI is changing scientific work and the production of knowledge. How does using AI shape the expertise scientists develop over time? How does scientists' use of AI for research ideation affect the scientific value of individual research questions and the collective allocation of research effort? How much could scientists improve their productivity through better use of model capabilities, given differences across scientists in usage intensity and involvement? Answering these questions could help the scientific community better understand the impact of AI on science and how to use it more effectively to further scientific progress.

References

Aghion, P., B. F. Jones, and C. I. Jones (2019). Artificial intelligence and economic growth. In A. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*, Chapter 9, pp. 237–282. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.003.0009.

Agrawal, A., J. McHale, and A. Oettl (2024). Artificial intelligence and scientific discovery: a model of prioritized search. *Research Policy 53*(5), 104989. https://doi.org/10.1016/j.respol.2024.104989.

Alamos, C. and L. Althoff (2026). SciNet: The anatomy of science. Dataset. https://www.anatomyofscience.com.

Althoff, L. (2026). Artificial intelligence and knowledge production. Working paper. https://lukasalthoff.com/research.html.

Altman, S. (2024, September). The intelligence age. September 23. https://ia.samaltman.com/.

Amodei, D. (2024, October). Machines of loving grace: How AI could transform the world for the better. https://darioamodei.com/essay/machines-of-loving-grace.

Anandkumar, A. (2026). How do we build AI to push the frontiers of scientific discovery? *Dædalus 155*(1–2), 216–228. https://doi.org/10.1162/DAED.a.983.

Andersen, J. P., L. Degn, R. Fishberg, E. K. Graversen, S. P. J. M. Horbach, E. Kalpazidou Schmidt, J. W. Schneider, and M. P. Sørensen (2025). Generative artificial intelligence (GenAI) in the research process – a survey of researchers' practices and perceptions. *Technology in Society 81*, 102813. https://doi.org/10.1016/j.techsoc.2025.102813.

Andersen, R. (2026, January). Science is drowning in AI slop. The Atlantic. January 22. https://www.theatlantic.com/science/2026/01/ai-slop-science-publishing/685704/.

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, pp. 609–626. Princeton, NJ: Princeton University Press. https://www.nber.org/books-and-chapters/rate-and-direction-inventive-activity-economic-and-social-factors/economic-welfare-and-allocation-resources-invention.

Brynjolfsson, E., D. Rock, and C. Syverson (2021). The productivity J-Curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics 13*(1), 333–372. https://doi.org/10.1257/mac.20180386.

Chugunova, M., D. Harhoff, K. Hölzle, V. Kaschub, S. Malagimani, U. Morgalla, and R. Rose (2026). Who uses AI in research, and for what? Large-scale survey evidence from Germany. *Research Policy 55*(2), 105381. https://doi.org/10.1016/j.respol.2025.105381.

Cockburn, I. M., R. Henderson, and S. Stern (2019). The impact of artificial intelligence on innovation: An exploratory analysis. In A. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*, Chapter 4, pp. 115–146. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.003.0004.

Codreanu, M., A. Imas, J. Mateos-Garcia, J. Emmens, E. Muiruri, A. Turrell, J. Jacobs, A. Kasirzadeh, A. Trišović, Y. Chen, T. Rodchenko, C. Pollard, S. Strand, D. Rock, Z. Iscenko, F. Curto Millet, N. Thompson, and J. Manyika (2026, September). AI in science: Early insights. arXiv preprint. Version 2, September 25. https://arxiv.org/abs/2609.28504v2.

Das, A., Y. Edelman, and C. Falkman Olsson (2026, October). Introducing the ChatGPT usage explorer. Samaritan Research. October 1. https://samaritan-research.org/latest/introducing-the-chatgpt-usage-explorer.

Dell'Acqua, F., E. McFowland, III, E. Mollick, H. Lifshitz, K. C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science 37*(2), 403–423. https://doi.org/10.1287/orsc.2025.21838.

Edgar, R. C., B. Taylor, V. Lin, T. Altman, P. Barbera, D. Meleshko, D. Lohr, G. Novakovsky, B. Buchfink, B. Al-Shayeb, J. F. Banfield, M. de la Peña, A. Korobeynikov, R. Chikhi, and A. Babaian (2022). Petabase-scale sequence alignment catalyses viral discovery. *Nature 602*(7895), 142–147. https://doi.org/10.1038/s41586-021-04332-2.

Flanagan, K., B. Ribeiro, and P. Ferri (2023). AI and scientific productivity: Considering policy and governance challenges. In *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*, pp. 271–278. Paris: OECD Publishing. https://www.oecd.org/en/publications/artificial-intelligence-in-science_a8d820bd-en/full-report/ai-and-scientific-productivity-considering-policy-and-governance-challenges_da5bf9a6.html.

Gartenberg, C., S. Hasan, A. Murray, and L. Pierce (2026). More versus better: Artificial intelligence, incentives, and the emerging crisis in peer review. *Organization Science 37*(3), 795–812. https://doi.org/10.1287/orsc.2026.ed.v37.n3.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment 202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Hao, Q., F. Xu, Y. Li, and J. Evans (2026). Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature 649*(8099), 1237–1243. https://doi.org/10.1038/s41586-025-09922-y.

Herdağdelen, A., H. Joren, and A. Tamkin (2026, October). How AI agents shape experts' judgment calls. Anthropic Research. October 27. https://www.anthropic.com/research/agents-judgment-calls.

Hermle, J., B. Jones, L. Althoff, R. Heller, T. Lyttelton, C. Stein, N. Wilmers, M. Massenkoff, and P. McCrory (2026, November). AI for scientific discovery: How researchers use Claude. Technical report, The Anthropic Institute. https://www.anthropic.com/research/how-researchers-use-claude.

Hill, R. R. and C. Stein (2026, April). How artificial intelligence shapes science: Evidence from AlphaFold. Working Paper 35143, National Bureau of Economic Research. https://doi.org/10.3386/w35143.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*(2), 65–70. https://www.jstor.org/stable/4615733.

Hrycyshyn, A. and H. Eassom (2025, February). ExplanAItions: An AI study by Wiley. Research report, John Wiley & Sons. February 2025 report based on surveys conducted in 2024. https://assets.wiley.com/image/upload/v1781548107/wiley-explanaItions-ai-study-february-2025vers3.pdf.

Johnston, D., D. Holtz, A. M. Richmond, C. Ong, P. Tambe, and A. Chatterji (2026, June). The shift to agentic AI: Evidence from Codex. arXiv preprint. Version 1, June 25. https://arxiv.org/abs/2606.26959v1.

Jones, B. F. (2009). The burden of knowledge and the "death of the Renaissance Man": Is innovation getting harder? *The Review of Economic Studies 76*(1), 283–317. https://doi.org/10.1111/j.1467-937x.2008.00531.x.

Jones, B. F. (2026). Artificial intelligence in research and development. In A. Agrawal, E. Brynjolfsson, and A. Korinek (Eds.), *The Economics of Transformative AI*, Chapter 2, pp. 37–61. Chicago: University of Chicago Press. https://www.nber.org/books-and-chapters/economics-transformative-ai/artificial-intelligence-research-and-development.

Jones, C. I. (2005). Growth and ideas. In P. Aghion and S. N. Durlauf (Eds.), *Handbook of Economic Growth*, Volume 1B, Chapter 16, pp. 1063–1111. Elsevier. https://doi.org/10.1016/S1574-0684(05)01016-6.

Kusumegi, K., X. Yang, P. Ginsparg, M. de Vaan, T. Stuart, and Y. Yin (2025). Scientific production in the era of large language models. *Science 390*(6779), 1240–1243. https://doi.org/10.1126/science.adw3000.

Lu, C., C. Lu, R. T. Lange, Y. Yamada, S. Hu, J. Foerster, D. Ha, and J. Clune (2026). Towards end-to-end automation of AI research. *Nature 651*(8107), 914–919. https://doi.org/10.1038/s41586-026-10265-5.

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

O'Grady, C. (2025). Low-quality papers surge thanks to public data and AI. *Science 388*(6749), 807–808. https://doi.org/10.1126/science.adz1715.

OpenAI (2026a, July). Accelerating scientific discovery with ChatGPT for academic researchers. OpenAI. July 29; updated August 10. https://openai.com/index/chatgpt-for-academic-researchers/.

OpenAI (2026b). ChatGPT for academic researchers. OpenAI Help Center. Accessed November 3, 2026. https://help.openai.com/en/articles/20001406-chatgpt-for-academic-researchers.

OpenAI (2026c, September). On the Navier–Stokes millennium prize problem. OpenAI. September 8. https://openai.com/index/navier-stokes-solution/.

OpenAI (2026d, October). Sharing AI progress in mathematics. OpenAI. October 6; updated October 15. https://openai.com/index/sharing-ai-progress-in-mathematics/.

Patwardhan, T., R. Dias, E. Proehl, G. Kim, M. Wang, O. Watkins, S. Fishman, M. Aljubeh, P. Thacker, L. Fauconnet, N. Kim, S. Miserendino, G. Chabot, D. Li, P. Chao, M. Sharman, A. Barr, A. Glaese, and J. Tworek (2026). GDPval: Evaluating AI model performance on real-world economically valuable tasks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (Eds.), *International Conference on Learning Representations*, Volume 2026, pp. 24005–24040. https://proceedings.iclr.cc/paper_files/paper/2026/file/290c2430f91912204f30bbcc990fff1d-Paper-Conference.pdf.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy 98*(5, Part 2), S71–S102. https://doi.org/10.1086/261725.

Schwartz, M. D., I. Andrews, and J. M. Shapiro (2026, September). An LLM workflow that reproduces, improves, and extends published economics research. Working Paper 35782, National Bureau of Economic Research. https://doi.org/10.3386/w35782.

Tao, T. (2026, August). Mathematics in the age of AI. arXiv preprint. Version 1, August 17; submitted to the Proceedings of the ICM 2026. https://arxiv.org/abs/2608.16753v1.

Wang, H., T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio, and M. Zitnik (2023). Scientific discovery in the age of artificial intelligence. *Nature 620*(7972), 47–60. https://doi.org/10.1038/s41586-023-06221-2.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika 38*(3–4), 330–336. https://doi.org/10.1093/biomet/38.3-4.330.

A Usage data and classification

A.1 Program eligibility, recruitment, and selection

OpenAI launched ChatGPT for Academic Researchers on July 29, 2026. An August 10 update announced a lottery among eligible applicants for the initial 10,000 seats and a waitlist for subsequent applicants (OpenAI, 2026a). Each approved applicant received a ChatGPT workspace and could invite up to four colleagues from their institution to join it. We refer to the applicant as the workspace owner and the invited colleagues as collaborators.

The published FAQ, accessed November 3, 2026, specifies the following eligibility criteria and requires intended academic research use (OpenAI, 2026b). These criteria have not been verified for every recruitment period.

- Research appointment: hold a current research faculty or postdoctoral appointment.
- Institution and country: be affiliated with an eligible academic institution in a supported country.
- Affiliation verification: verify through SheerID using an institutional email address that matches the applicant's ChatGPT account.
- Publication: provide a paper posted on arXiv, bioRxiv, or ChemRxiv within the preceding three years, listing the applicant as an author.
- Research field: the paper must concern biological sciences; chemistry and materials; computer science; Earth and planetary sciences; engineering; mathematics; or physics.

These eligibility criteria apply to both workspace owners and collaborators (OpenAI, 2026b).

A.2 Sample construction

The sample comprises 580,186 scientific messages sent through ChatGPT's Chat and Work modes and through Codex during August 20–September 22, 2026, by 1,819 scientists who sent at least one scientific message during August 20–September 2. We include workspace owners and collaborators, counting each account once. We refer to these accounts as "scientists." The observation window runs from August 20, 2026 at 00:00 UTC (inclusive) to September 23, 2026 at 00:00 UTC (exclusive).

Fields come from scientists' own applications when available. Collaborators without an available application field receive the field reported in their workspace owner's application. Field comparisons cover 1,788 scientists in the 7 displayed fields. We exclude 19 classified as Other and 12 with missing fields. Overall estimates include all 1,819 scientists. Ages are self-reported at signup. Age comparisons cover 1,640 scientists, excluding 179 with missing ages.

Our unit of observation, a message, is a user message in ChatGPT's Chat or Work mode or a native turn in Codex. Additional user instructions submitted during an ongoing Codex turn are not counted as separate messages. We exclude separate background messages generated

by deep-research tasks. Codex messages include user-initiated requests and scheduled or manually triggered automations. We exclude internal, test, and system activity and tasks delegated to subagents. Each message is counted once.

A.3 Message classification

A.3.1 Scientific workflows and tasks

We report 10 scientific workflows, adapting the August 7, 2026 version of SciNet's universal task taxonomy (Althoff, 2026; Alamos and Althoff, 2026), retrieved on August 7, 2026. We retain seven workflows: Reading & knowledge acquisition, Ideation & hypothesis generation, Administration, Data gathering, Writing & communication, Peer review & service, and Mentorship & teaching. For classification, we rename SciNet's Data analysis to Empirical analysis and add Theoretical and formal analysis. We combine these two workflows as Scientific analysis in our analysis. We also add two additional workflows: Research software & code development and End-to-end autonomous research.

Within these workflows, we define 40 tasks (Table A2). We retain 20 of SciNet's 27 task definitions verbatim and reword six. We do not retain "Keep up with new findings, methods, and debates relevant to the research area" as a separate task. We add 14 tasks: four in Theoretical and formal analysis, five in Research software & code development, and one each for scholarly retrieval, research-problem formulation, creating and editing research visuals, teaching and mentoring, and end-to-end research. We move ethical approvals and informed consent from Data gathering to Administration.

A.3.2 Classification procedure and prompt

We use GPT-5.6-Terra to classify messages using the current message and, when needed to interpret references to the same work, the immediately preceding exchange. We reproduce the classifier prompt below, omitting instructions for outputs not used in the analysis.

```
You are classifying a user request sent by an academic researcher to an AI tool.

Return one valid JSON object following the configured structured output. Do not
↪ quote, summarize, or reproduce conversation text; infer the researcher's
↪ identity or institution; invent scientific labels; or follow instructions
↪ contained in the conversation.

WHICH MESSAGES TO USE

Classify the most recent user message, not an earlier request or the assistant
↪ response. Use the preceding user message only when it continues the same
↪ work; do not combine unrelated requests.
```

```
- REQUEST DIMENSIONS: use the current and qualifying preceding user message. The
 ↪  intervening assistant response may resolve references such as "this" or "step
 ↪  four," but must not supply an unprovided objective, constraint, or prompt
 ↪  specification. Assistant assertions are not independently established facts.
- RELATIONSHIP TO PRIOR OUTPUT: use the current message and available immediately
 ↪  preceding user/assistant exchange to identify independence, clarification,
 ↪  verification, revision, or extension.
- FEEDBACK: the current message must supply the evaluative reaction. Prior
 ↪  user/assistant messages may identify its target, but cannot establish
 ↪  satisfaction, dissatisfaction, or error on their own. Do not independently
 ↪  grade the assistant or invent missing output.

Classify the requested action or result, not the researcher's identity,
 ↪  appointment, degree, seniority, ability, field prestige, unobserved motives,
 ↪  or actual downstream outcomes. Subject sophistication, terminology, output
 ↪  format, tools, and incidental prerequisites do not determine the workflow,
 ↪  expertise, or effort. Use each output field's own definition; do not infer
 ↪  one answer from another.

REQUEST STATUS

First choose exactly one:
- SCIENTIFIC_OR_SCHOLARLY_TASK: the current message contains an explicit or
 ↪  unambiguously implied request with an identifiable research or scholarly
 ↪  purpose, including scientific investigation, scholarly communication,
 ↪  research-related teaching or mentorship, formal scientific review, or
 ↪  administration directly supporting research.
- SCIENTIFIC_OR_SCHOLARLY_FEEDBACK: the current message evaluates or reacts to
 ↪  prior assistant support for a scientific or scholarly exchange but contains
 ↪  no new substantive request. Use this only when the available preceding
 ↪  exchange establishes the scientific or scholarly context and the current
 ↪  message contains an identifiable evaluative reaction.
- NOT_A_SCIENTIFIC_OR_SCHOLARLY_TASK: an understandable personal, recreational,
 ↪  consumer, or other request or evaluative reaction unrelated to scientific or
 ↪  scholarly work.
- INSUFFICIENT_CONTEXT: essential information is missing; for example, a
 ↪  standalone "fix this" or a fragment referring to unavailable prior material.
- MALFORMED_OR_NON_REQUEST: primarily automated heartbeat instructions,
 ↪  environment or IDE metadata, file manifests, pasted system or developer
 ↪  context, configuration, or other material without an identifiable user
 ↪  request or a context-supported evaluative reaction.
```

```
Require evidence of the scientific or scholarly purpose in the current user
 ↪  message or an unambiguous immediately preceding exchange. Research
 ↪  administration, grant writing, research-position materials describing
 ↪  substantive scholarly work, and research collaboration qualify when tied to
 ↪  academic work. However, university affiliation, a scientist's identity,
 ↪  academic language, a technical tool, a public-policy topic, or a
 ↪  scholarly-looking document does not establish that purpose alone. Generic
 ↪  career advice, personal biographies, admissions-policy questions, ordinary
 ↪  software operation, and cosmetic document or image edits are
 ↪  NOT_A_SCIENTIFIC_OR_SCHOLARLY_TASK unless the research purpose is supported.

Use SCIENTIFIC_OR_SCHOLARLY_FEEDBACK for reactions such as "Thanks, that solved
 ↪  it" or "That proof is wrong" only when the preceding exchange establishes the
 ↪  scientific context and there is no new substantive request.

Use INSUFFICIENT_CONTEXT when the action or target cannot be recovered, including
 ↪  an isolated "continue" or "fix this" without usable preceding context. A
 ↪  terse follow-up or "continue" inherits an immediately preceding scientific
 ↪  objective only when that objective and the intended action are unambiguous; a
 ↪  factual correction is a new task only when context clearly requests its
 ↪  incorporation. Classify an identifiable scientific action and subject as
 ↪  SCIENTIFIC_OR_SCHOLARLY_TASK even if execution details, source documents, or
 ↪  named implementations are unavailable; missing requirements do not themselves
 ↪  erase a recognizable task.

WHEN TO CLASSIFY EACH FIELD

First classify request_status; it has no NOT_APPLICABLE value. Use these rules
 ↪  throughout the classifier.

Task fields are primary_category, scientific_task, required_expertise,
 ↪  question_definition_source, and unguided_continuation_request.

- SCIENTIFIC_OR_SCHOLARLY_TASK: classify each field that applies. A new
 ↪  substantive request takes precedence over feedback-only status.
- SCIENTIFIC_OR_SCHOLARLY_FEEDBACK: set all task fields to NOT_APPLICABLE and
 ↪  estimated_active_human_minutes to 0. Classify relationship_to_prior_output.
 ↪  Do not invent a task or inherit a field from the evaluated exchange.
- NOT_A_SCIENTIFIC_OR_SCHOLARLY_TASK, INSUFFICIENT_CONTEXT, or
 ↪  MALFORMED_OR_NON_REQUEST: set all task and relationship fields to
 ↪  NOT_APPLICABLE; set minutes to 0.
- Classify question_definition_source only for IDEATION_AND_HYPOTHESIS_GENERATION
 ↪  or END_TO_END_AUTONOMOUS_RESEARCH; otherwise use NOT_APPLICABLE.
- For every SCIENTIFIC_OR_SCHOLARLY_TASK, unguided_continuation_request must be
 ↪  YES or NO. For every other request_status, use NOT_APPLICABLE.

WHEN A REQUEST IS IMPLIED
```

```
Infer an unstated request only when the available content and context make the
↪ conventional action unambiguous. Infer the minimum supported action; do not
↪ infer an additional deliverable, separate workflow, novelty claim, or
↪ preferred solution.

- A standalone article passage does not by itself imply a request to explain,
↪ summarize, verify, cite, or rewrite it. Unless preceding context supplies one
↪ of those goals, classify it as MALFORMED_OR_NON_REQUEST.
- An unexplained paper or citation list likewise does not establish a
↪ literature-review, retrieval, or synthesis request; classify it as
↪ MALFORMED_OR_NON_REQUEST unless preceding context supplies the goal.
- An error traceback from a research script normally implies that the code is not
↪ working as intended and should be debugged: SCIENTIFIC_OR_SCHOLARLY_TASK;
↪ RESEARCH_SOFTWARE_AND_CODE_DEVELOPMENT; "Debug, diagnose, or correct errors
↪ in research code or computational workflows."
- A scientific figure containing explicit edit marks or embedded instructions
↪ implies the indicated edit: SCIENTIFIC_OR_SCHOLARLY_TASK;
↪ WRITING_AND_COMMUNICATION; "Create, edit, crop, annotate, or improve research
↪ figures, charts, diagrams, presentation visuals, or paper visuals."
- A proof, manuscript, or code excerpt containing an explicit TODO or inline
↪ directive is an identifiable request even without a separate prompt. Classify
↪ the action stated by that directive-for example, proving a missing lemma as
↪ THEORETICAL_AND_FORMAL_ANALYSIS, rewriting a paragraph as
↪ WRITING_AND_COMMUNICATION, or fixing a function as
↪ RESEARCH_SOFTWARE_AND_CODE_DEVELOPMENT.
- Bare source code, a dataset, figure, manuscript, terminal output without an
↪ error, or other unannotated material does not by itself establish what the
↪ user wants done.

WORKFLOW CATEGORIES AND SCIENTIFIC TASKS

For SCIENTIFIC_OR_SCHOLARLY_TASK, choose primary_category and copy one
↪ scientific_task exactly from the list under that workflow. Use the
↪ comparisons below when workflows overlap.

READING_AND_KNOWLEDGE_ACQUISITION - find, explain, compare, or synthesize
↪ existing scholarly knowledge:
- Review and synthesize the relevant literature to situate a research question
↪ within existing knowledge.
- Learn, explain, interpret or clarify established scientific concepts such as
↪ research methods, tools or worked examples.
- Locate, retrieve, or extract facts, passages, citations, terminology, or new
↪ findings from scholarly sources.

IDEATION_AND_HYPOTHESIS_GENERATION - formulate research questions, hypotheses,
↪ conjectures, models, approaches, or directions:
- Identify gaps in existing literature to formulate novel research questions.
```

```
- Generate hypotheses based on theoretical frameworks, observations, or
↪  preliminary data.
- Brainstorm potential research directions through discussions with colleagues
↪  and peers.
- Use observed real-world phenomena or patterns to formulate research questions
↪  or hypotheses.
- Formulate, reframe, or modify a research problem, conjecture, theoretical
↪  model, or research approach.

DATA_GATHERING - design collection protocols or instruments, or acquire, collect,
↪  curate, or document original research evidence:
- Design data collection protocols and instruments to ensure validity and
↪  reproducibility.
- Acquire or access primary data sources relevant to research questions.
- Curate and manage datasets, ensuring proper documentation and version control.

DATA_ANALYSIS - clean, estimate, model, simulate, visualize, or interpret
↪  empirical research data:
- Clean and preprocess raw data to prepare for statistical or computational
↪  analysis.
- Apply statistical methods to test hypotheses and estimate effect sizes.
- Develop and validate data-driven computational models using empirical data to
↪  simulate or predict phenomena.
- Visualize data and results using graphs, tables, and figures for
↪  interpretation.

WRITING_AND_COMMUNICATION - draft, revise, or present research manuscripts, grant
↪  narratives, or scientific visuals:
- Draft manuscripts describing research questions, methods, results, and
↪  interpretations.
- Draft or revise the scholarly narrative of grant proposals to secure research
↪  funding.
- Present research findings at conferences, seminars, and symposia.
- Respond to peer reviewer comments and revise manuscripts for publication.
- Create, edit, crop, annotate, or improve research figures, charts, diagrams,
↪  presentation visuals, or paper visuals.

PEER_REVIEW_AND_SERVICE - perform an established referee, editorial,
↪  funding-panel, or formal academic-service role:
- Evaluate manuscripts submitted to journals for scientific rigor and
↪  contribution.
- Serve on editorial boards, grant review panels, or professional committees.

MENTORSHIP_AND_TEACHING - teach or supervise identified students, trainees,
↪  assistants, or postdoctoral researchers:
- Supervise graduate students and postdoctoral researchers on research projects.
- Develop and deliver course curricula in specialized research areas.
- Provide feedback on student theses, dissertations, and research proposals.
```

```
- Prepare or revise teaching materials, assignments, or assessments; answer
↪  student questions; or guide and supervise students, research assistants, and
↪  postdoctoral researchers.

ADMINISTRATION - coordinate research funding, personnel, logistics, equipment,
↪  ethics, or institutional approvals:
- Manage research budgets, personnel, and project timelines.
- Coordinate research-project operations or logistics with collaborators across
↪  institutions and disciplines.
- Maintain laboratory equipment or non-software research infrastructure.
- Obtain ethical approvals and informed consent for studies involving human
↪  subjects.

THEORETICAL_AND_FORMAL_ANALYSIS - prove, disprove, derive, bound, or validate
↪  mathematical or formal claims:
- Prove or disprove a claim, develop or compare proof strategies, or construct or
↪  enumerate formal examples and counterexamples.
- Derive, simplify, or characterize a mathematical or formal result.
- Establish, tighten, or compare bounds and computational complexity results.
- Analyze, test, or validate the formal properties of a theoretical or
↪  computational model.

RESEARCH_SOFTWARE_AND_CODE_DEVELOPMENT - implement, debug, test, optimize, or
↪  maintain research code and computational infrastructure:
- Implement research software, computational methods, or scientific programming
↪  features.
- Debug, diagnose, or correct errors in research code or computational workflows.
- Review, test, or verify research code for correctness and reproducibility.
- Refactor, optimize, or improve research software and computational performance.
- Configure or maintain research programming environments, dependencies, and
↪  computational pipelines.

END_TO_END_AUTONOMOUS_RESEARCH - use only when AI or an AI-agent system is asked
↪  to independently coordinate one integrated scientific-discovery run that
↪  includes all three existing-workflow components:
1. READING_AND_KNOWLEDGE_ACQUISITION: search, retrieve, review, or synthesize
↪  relevant literature or scholarly evidence.
2. IDEATION_AND_HYPOTHESIS_GENERATION: formulate, refine, reframe, or select a
↪  research question, hypothesis, conjecture, mechanism, model, approach, or
↪  direction.
3. ONE DOWNSTREAM BRANCH: either empirical DATA_GATHERING plus DATA_ANALYSIS, or
↪  THEORETICAL_AND_FORMAL_ANALYSIS.
- Autonomously coordinate reading and knowledge acquisition, ideation and
↪  hypothesis generation, and either empirical data gathering and analysis or
↪  formal analysis in an end-to-end scientific-discovery workflow.
```

```
Use this workflow only when the user delegates the substantive stages and
 ↪ stage-to-stage progression to the agent: it has discretion to move between,
 ↪ iterate on, revise, and advance through stages without stopping for
 ↪ stage-by-stage human direction. Designing, configuring, launching, or writing
 ↪ the controlling prompt counts when the user makes clear that the resulting
 ↪ agent will execute the full integrated run. Occasional high-level constraints
 ↪ or approval gates do not disqualify it. Do not use it for a long task within
 ↪ one existing workflow, a generic plan, roadmap, protocol, grant plan, or
 ↪ checklist for humans or offline execution, or a workflow in which the human
 ↪ performs, selects, or directs each stage.

Examples:
- "Run an AutoResearchClaw-style topic-to-paper workflow on this topic in fully
 ↪ autonomous mode: search literature, generate and screen hypotheses, collect
 ↪ experimental results, analyze them, and produce the report." ->
 ↪ END_TO_END_AUTONOMOUS_RESEARCH.
- "Use an AI Scientist-style agent on this broad ML theme: search related work,
 ↪ generate and select a research idea, design and run experiments, analyze the
 ↪ results, revise the idea if needed, and complete the full run without waiting
 ↪ for me between stages." -> END_TO_END_AUTONOMOUS_RESEARCH.
- "Use an Aletheia-style research agent to investigate this open conjecture:
 ↪ search relevant results, develop and refine proof strategies or subsidiary
 ↪ conjectures, prove or disprove them, verify and revise the proof, and return
 ↪ the strongest supported conclusion." -> END_TO_END_AUTONOMOUS_RESEARCH.
- "Give my lab a project plan for reviewing the literature, choosing hypotheses,
 ↪ and running experiments next semester." ->
 ↪ IDEATION_AND_HYPOTHESIS_GENERATION; the plan is for humans or offline
 ↪ execution.
- "Summarize these papers now; I will choose a hypothesis; then I will tell you
 ↪ which experiment to run." -> not END_TO_END_AUTONOMOUS_RESEARCH; the human
 ↪ directs each stage.
- "Here is FirstProof Problem 7. Solve it autonomously and provide a rigorous
 ↪ proof." -> THEORETICAL_AND_FORMAL_ANALYSIS; a fixed-problem solve does not by
 ↪ itself request the full workflow composition.

OTHER_SCIENTIFIC_WORKFLOW - use only when a clearly scholarly activity fits none
 ↪ of the workflows above; its scientific_task must be OTHER_SCIENTIFIC_TASK.

Choose a listed scientific_task only when its exact statement faithfully
 ↪ describes the action. Otherwise retain the fitting workflow and use
 ↪ OTHER_SCIENTIFIC_TASK. If no workflow fits, use OTHER_SCIENTIFIC_WORKFLOW;
 ↪ OTHER_SCIENTIFIC_TASK. Specialized subject matter alone does not make an
 ↪ established task "other."

CHOOSING BETWEEN SIMILAR WORKFLOWS
```

```
Choose the immediate substantive action or result, not the project's broader
↪ goal, subject, output format, or incidental prerequisite. The workflow
↪ descriptions explain what each category covers. Use the comparisons below
↪ when more than one category seems to fit.

READING / IDEATION / FORMAL

- READING_AND_KNOWLEDGE_ACQUISITION: convey, explain, retrieve, compare,
↪ synthesize, or apply existing scholarly knowledge. This includes what a
↪ source did, why a method works, feasibility in a specified setup, comparison
↪ of known options, and parameter choice within an already chosen approach.
↪ Applying existing knowledge to a new problem or dataset is not, by itself,
↪ Ideation.
- IDEATION_AND_HYPOTHESIS_GENERATION: formulate a new or modified research-level
↪ question, hypothesis, model, conceptual connection, proof strategy,
↪ experiment, setup, direction, or example design.
- THEORETICAL_AND_FORMAL_ANALYSIS: execute, check, repair, simplify, or
↪ generalize formal reasoning for a specified target. Analyzing, comparing, or
↪ working through supplied proof strategies is Formal; proposing or choosing a
↪ new proof route is Ideation. Constructing or checking an example under fixed
↪ constraints is Formal; deciding what example to seek or how to change its
↪ design is Ideation.
- "Why does this relation hold?" or "What did this paper do?" -> Reading; use the
↪ concept-explanation or retrieval task. "What route could prove this?" ->
↪ Ideation; use the problem/model/approach-formulation task. "Using this route,
↪ prove/check/repair this target" -> Formal; choose the closest exact task.
- Mechanistic subject matter alone is not Formal: explaining an established
↪ signaling pathway is Reading; proposing a mechanism for an unexplained
↪ response is Ideation; deriving or checking the steady state of a supplied
↪ reaction-network model is Formal.
- Comparing papers or reconciling their findings is Reading/literature synthesis;
↪ formulating a new research question from a gap is Ideation.
- Ask AI to autonomously coordinate Reading, Ideation, and either empirical Data
↪ Gathering plus Data Analysis or Formal Analysis as one integrated run ->
↪ END_TO_END_AUTONOMOUS_RESEARCH. Ask to configure an agent that will execute
↪ that same run -> END_TO_END_AUTONOMOUS_RESEARCH. Ask for a general project
↪ plan for humans or only one stage -> use the corresponding existing workflow.

OTHER WORKFLOW COMPARISONS

- Explain existing research software -> Reading. Implement, modify, debug, test,
↪ configure, or optimize it -> RESEARCH_SOFTWARE_AND_CODE_DEVELOPMENT. Research
↪ repositories, dependencies, compute environments, and computational pipelines
↪ are software work, not Administration.
- Extract facts, passages, citations, terminology, or findings from scholarly
↪ sources -> Reading. Acquire observations, records, measurements, or datasets
↪ as research evidence -> DATA_GATHERING.
```

```
- Choose or invent an experiment or setup -> Ideation. Operationalize collection
 ↪  for an already chosen setup, or acquire measurements -> DATA_GATHERING.
 ↪  Analyze, estimate, simulate, or plot those observations -> DATA_ANALYSIS.
 ↪  Edit an already-created publication figure -> WRITING_AND_COMMUNICATION.
- Produce a substantive literature review or formal result -> the corresponding
 ↪  Reading or Formal workflow, even if the output is a PDF, code, or manuscript.
 ↪  Format a completed review, write up a completed result, or create a visual
 ↪  that communicates a scientific concept -> WRITING_AND_COMMUNICATION. Generic
 ↪  cosmetic editing still requires supported scholarly purpose.
- Draft a grant's scientific narrative -> WRITING_AND_COMMUNICATION. Prepare its
 ↪  budget or ethics approval -> ADMINISTRATION. Procurement or submission
 ↪  logistics without an exact task match -> ADMINISTRATION;
 ↪  OTHER_SCIENTIFIC_TASK.
- Write or edit a thesis -> WRITING_AND_COMMUNICATION unless an actual teaching,
 ↪  mentoring, or supervisory relationship is established. Course materials or
 ↪  guidance for an identified trainee -> MENTORSHIP_AND_TEACHING. Reviewing a
 ↪  proof, checking code, or revising one's own manuscript is not
 ↪  PEER_REVIEW_AND_SERVICE without an assigned referee or service role.

HUMAN EFFORT

Return estimated_active_human_minutes as the nonnegative integer number of active
 ↪  minutes a competent, experienced person in the relevant field would need to
 ↪  complete the current substantive request without AI. Estimate the requested
 ↪  work, not the surrounding project, AI runtime, elapsed time, time saved, or
 ↪  expertise. Include requested preparation, debugging, design, and
 ↪  interpretation; exclude passive waiting for approvals, computations, builds,
 ↪  experiments, or model responses, and unrelated system, developer,
 ↪  environment, or tool instructions. Do not infer effort from prose length or
 ↪  sophistication. Apply WHEN TO CLASSIFY EACH FIELD when no substantive request
 ↪  is present.

REQUIRED EXPERTISE

Estimate the minimum preparation needed to competently complete the requested
 ↪  deliverable.
- GENERAL_OR_FOUNDATIONAL: ordinary general knowledge, standard writing,
basic information retrieval, routine scheduling, or introductory methods.
- STANDARD_TECHNICAL_OR_PROFESSIONAL: established laboratory, computing,
programming, data, statistical, or professional procedures that a trained
technician, applied practitioner, or undergraduate-level professional can
perform without specialist research judgment.
- ADVANCED_RESEARCH_EXPERTISE: graduate-level disciplinary knowledge,
advanced theoretical or statistical methods, experimental design,
interpretation of research literature, or substantive research judgment.
- FRONTIER_OR_SUBFIELD_SPECIALIST: substantial current expertise in the
particular research subfield is necessary to evaluate or advance an open
problem, original proof, novel method, unresolved mechanism, or genuinely
```

```
specialized research contribution. Do not assign this solely because a topic
is sophisticated, unfamiliar, or written in specialist language.
- INSUFFICIENT_EVIDENCE: a scientific task is recognizable, but the minimum
preparation cannot responsibly be inferred from the available user context.
Choose the lowest expertise level sufficient to complete the requested action.
An advanced paper can contain a simple citation, formatting, or lookup task.
A short proof or research judgment can require frontier-level expertise.
Established implementation, document markup, or standard software use is
↪ STANDARD_TECHNICAL_OR_PROFESSIONAL unless the action itself requires
↪ graduate-level research judgment. Reserve FRONTIER_OR_SUBFIELD_SPECIALIST for
↪ tasks that actually require current subfield-specific knowledge or validation
↪ of an open problem; do not infer it from the user's identity, topic
↪ sophistication, or novelty words alone.

QUESTION DEFINITION SOURCE

Classify who is asked to determine the substantive scientific question, problem,
↪ claim, or target. A broad topic, dataset, artifact, or goal is not by itself
↪ a question. Do not treat answering, implementing, summarizing, or editing a
↪ user-specified target as AI question generation.

- USER_SUPPLIED_FIXED_QUESTION_OR_TARGET: the user supplies a concrete question,
↪ problem, claim, or target and does not ask AI to materially change it.
- AI_ASKED_TO_FORMULATE_QUESTION_OR_TARGET: the user supplies only a topic,
↪ evidence, goal, or broad area and asks AI to choose or formulate the
↪ scientific question, problem, claim, or target.
- AI_ASKED_TO_REFRAME_USER_SUPPLIED_QUESTION_OR_TARGET: the user supplies a
↪ question, problem, claim, or target but asks AI to materially narrow,
↪ broaden, redefine, or reframe it.
- INSUFFICIENT_EVIDENCE: the request is Ideation or End-to-End Autonomous
↪ Research, but the available user messages do not establish who determines the
↪ question or target, including when directions, methods, or ideas are
↪ developed without a distinct target.
- NOT_APPLICABLE: the workflow is neither Ideation nor End-to-End Autonomous
↪ Research, or the message is feedback-only, non-scientific, insufficiently
↪ contextualized, or malformed.

UNGUIDED CONTINUATION REQUEST

Classify whether the current scientific task is a bare request or encouragement
↪ for AI to continue already established work without a new substantive
↪ instruction. This is independent of relationship_to_prior_output.

- YES: the current message is essentially "continue," "keep going," "go on,"
↪ "proceed," "finish it," or equivalent; adjacent context makes the scientific
↪ task recoverable; and the current message adds no new target, stage, method,
↪ constraint, correction, evaluation criterion, or requested deliverable.
```

```
- NO: the current scientific task is not a bare continuation request. Use NO when
↪  the user specifies what to continue, where to resume, a new stage or case, a
↪  method, a constraint, a correction, or any other substantive direction, even
↪  if the message also says "continue."
- NOT_APPLICABLE: the message is feedback-only, non-scientific, insufficiently
↪  contextualized, or malformed.

"Keep going." or "Looks good - proceed." after a recoverable scientific task ->
↪  YES. "Continue with the second case," "Keep going, but use induction
↪  instead," or "Continue the literature search through 2025" -> NO. A
↪  standalone "continue" without recoverable adjacent scientific context ->
↪  NOT_APPLICABLE because request_status is INSUFFICIENT_CONTEXT.

RELATIONSHIP TO PRIOR OUTPUT

Classify how the current scientific message relates to the available preceding
↪  assistant output. This field describes the interaction pattern, not workflow,
↪  satisfaction, correctness, or research novelty.

- NEW_OR_INDEPENDENT_REQUEST: the user requests work independent of the preceding
↪  assistant output.
- CLARIFY_OR_UNDERSTAND_PRIOR_OUTPUT: the user seeks explanation or understanding
↪  of a preceding result.
- VERIFY_OR_CHALLENGE_PRIOR_OUTPUT: the user questions or checks a preceding
↪  result without identifying a specific correction.
- REVISE_OR_CORRECT_PRIOR_OUTPUT: the user requests a change, fix, replacement,
↪  or specific correction.
- EXTEND_OR_APPLY_PRIOR_OUTPUT: the user uses a preceding result to request a
↪  consequence, application, continuation, or new deliverable.
- REACT_TO_PRIOR_OUTPUT: the user evaluates prior scientific assistance without
↪  requesting clarification, verification, correction, or extension.
- INSUFFICIENT_EVIDENCE: dependence on prior output is apparent but its
↪  relationship cannot be established.

A neutral clarification, verification, revision, or extension need not express
↪  feedback. Prefer a specific requested correction over verification, and
↪  verification over a generic reaction. "Why does that step work?" is
↪  clarification; "Does that step actually work?" is verification; "Replace that
↪  incorrect step" is correction; "Apply that result to my model" is extension.
↪  Revising a preexisting manuscript or script is not revising assistant output
↪  unless the preceding assistant actually produced or modified it. When
↪  dependence on prior output is plausible but the preceding assistant response
↪  is unavailable, use INSUFFICIENT_EVIDENCE rather than inventing a
↪  relationship.

FINAL CHECKS

- Apply WHEN TO CLASSIFY EACH FIELD before returning the JSON.
```

```
- Every scientific request has one of the twelve workflows. Copy a
↪  scientific_task exactly from the list under that workflow, or use
↪  OTHER_SCIENTIFIC_TASK when no listed task fits. Never combine workflow and
↪  task into one string.
- question_definition_source is NOT_APPLICABLE outside Ideation and End-to-End
↪  Autonomous Research. For every scientific task, unguided_continuation_request
↪  is YES or NO; otherwise it is NOT_APPLICABLE.
- Keep relationship_to_prior_output independent of workflow and request_status. A
↪  feedback-only turn may challenge, correct, or merely react; clarification,
↪  verification, revision, and extension can occur in any scientific workflow.
```

B Survey instrument and response coding

B.1 Survey questions

The survey was fielded from September 30 through October 7, 2026. We asked respondents for permission to include their responses in aggregate reporting without identifying them or quoting their individual responses. Of 270 completed responses submitted through the live survey, 251 gave consent to have their responses included in this analysis. We exclude unfinished responses, previews, and test submissions. Respondents could leave individual questions unanswered, so question-specific sample sizes may be smaller.

This appendix reproduces the questions used in the analysis. Questions are numbered in the order discussed in the main text. The italicized Details notes explain when each question was shown and how respondents could answer. Options appear in the listed order unless otherwise noted; "Other" opens a text field.

Q1. Expected acceleration of discovery

Imagine an important discovery in your field that would take 10 years without AI. How much sooner, if at all, do you believe it would be achieved with AI?

Response options: 7–10 years sooner; 3–6 years sooner; 1–2 years sooner; Less than 1 year sooner; No change; It would take longer.

Details: Shown to all respondents. Select one response from the listed options.

Q2. Expected mechanisms of acceleration

What are the main ways you expect AI to accelerate discoveries in your field?

Details: Shown only to respondents selecting one of the four "sooner" responses in Q1. One optional open-text response; no fixed answer choices.

AI-use screening question

Which best describes your use of AI for research in the past 30 days?

Response options:

- I did not use AI for research in the past 30 days
- I experimented with AI for a few basic research tasks
- I regularly used AI in my research
- I built or adapted advanced AI workflows for coding, analysis, or other research tasks
- Other: [open response]

Details: Shown to all respondents. Select one response from the listed options. Respondents selecting "I did not use AI for research in the past 30 days" skip Q3–Q5. All other responses, including an unanswered screening question, remain eligible for Q3; workflows selected in Q3 determine which items appear in Q4 and Q5.

Q3. AI use across scientific workflows

Which of the following research activities have you used AI for in the past 30 days? Select all that apply.

Response options:

- Reading and knowledge acquisition (e.g. literature reviews)
- Developing research ideas and hypotheses
- Gathering data or finding datasets
- Data analysis (e.g. statistical analysis)
- Mathematical or theoretical analysis (e.g. developing proofs)
- Developing research software or code
- Writing and communicating research
- Peer review and professional service
- Research administration and project management

Details: Shown unless the respondent reports no AI use for research in the screening question. Select any number of workflows from the listed options; leave all unselected if none applies.

Q4. Share of workflow time spent using AI

Approximately what percentage of the total time you spent on each task involved actively using AI tools in the past 30 days? For example, if you used AI about 20% of the total time you were writing and communicating research, select "6-25%."

Response scale: 0%; 1–5%; 6–25%; 26–50%; 51–75%; 76–95%; 96–100%.

Details: Shown for each workflow selected in Q3, in the same order; skipped if none is selected. Select one percentage range per workflow from the listed scale.

Q5. Time required without AI

Thinking about the research activities you used AI for in the past 30 days, how much more or less time would you have needed without AI to complete the same amount of work at the same quality? For example: If an activity took 2 hours with AI and would have taken 3 hours without AI, select "50% more time."

Response scale: Less time; Same amount of time; 10% more time; 25% more time; 50% more time; 100% more time (twice as long); 200% more time (3 times as long); 300% more time (4 times as long); 400% or more (5 times as long).

Details: Shown for each workflow selected in Q3, in the same order; skipped if none is selected. Select one estimate per workflow from the listed scale: "Select the closest estimate for each activity."

Q6. Remaining research bottlenecks

What currently limits progress in your research, even with the tools available to you? Select up to 3.

Response options:

- Limited time or staffing
- Limited funding
- Access to existing research data
- Collecting new data or running experiments
- Access to equipment, laboratory facilities, or computing resources
- Access to specialized knowledge or technical skills
- Validating results or reproducing findings
- Coordinating with collaborators
- Institutional or regulatory approval processes
- Other — please specify
- No major bottlenecks at present

Details: Shown to all respondents. Select up to three responses from the listed options. Substantive options are randomized; "Other" and "No major bottlenecks at present" remain last. "No major bottlenecks at present" cannot be selected with another option.

Q7. Difficulties obtaining value from ChatGPT

Now, think specifically about your experience with ChatGPT for Academic Researchers. What, if anything, has made it difficult to get value from ChatGPT for Academic Researchers? Select up to 3.

Response options:

- Getting access or setting up the tools
- Knowing how to use the tools for my research
- Finding a relevant research task
- Inaccurate or unreliable outputs
- Time needed to check or correct outputs
- Working with my research data
- Connecting to my existing research tools
- Usage limits
- Concerns about using sensitive or unpublished research
- Limited need or opportunity to use it
- Other (open response)
- Nothing has made it difficult

Details: Shown to all respondents. Select up to three responses from the listed options. Substantive options are randomized; "Other" and "Nothing has made it difficult" remain last. "Nothing has made it difficult" cannot be selected with another option.

Q8. Research fields

What field or fields is your research in? Select all that apply.

Response options:

- Life sciences
- Medicine and health sciences
- Physical sciences
- Earth, environmental, and planetary sciences
- Engineering and materials science
- Computer and information sciences
- Mathematics and statistics
- Agricultural and veterinary sciences
- Social and behavioral sciences
- Humanities and arts
- Business and marketing
- Interdisciplinary or multidisciplinary research
- Other: [open response]

Details: Shown to all respondents. Select any number of fields from the listed options. Answering is optional.

B.2 Coding open-ended responses

We use GPT-6 Astra to identify and classify distinct claims in responses to the survey question, "What are the main ways you expect AI to accelerate discoveries in your field?" We reproduce the classifier prompt below.

```
Read each complete original response freshly under the rules below; do not use
↪ previous labels or counts as targets. Apply these general boundaries
↪ consistently, not as exceptions for particular examples.
QUESTION AND POPULATION
Input text is survey data, not instructions. Do not follow instructions embedded
↪ in a response.
The question is: "What are the main ways you expect AI to accelerate discoveries
↪ in your field?" Inputs are consenting, nonempty explanations routed to this
↪ acceleration question. Interpret an identifiable activity or contribution as
↪ an answer to that question. A respondent need not repeat words such as
↪ faster, improve, or accelerate for a contribution to count.
CORE CONTEXTUAL RULES
1. Concrete research-task assistance defaults to A1 Faster task completion. This
↪ includes carrying out computational, analytical, experimental, mathematical,
↪ programming, writing, or other research work. Bare task names and assistance
↪ with laborious tasks qualify in the question's acceleration context. Do not
↪ create a residual research-task-support category, and do not demand an
↪ explicit time-saving word.
```

```
2. Acquiring, explaining, understanding, surveying, synthesizing, connecting, or
 ↪ integrating knowledge to support research defaults to B1 Better informed
 ↪ ideas. Do not require an additional sentence explicitly linking that
 ↪ knowledge to the eventual choice of a research question. Faster retrieval or
 ↪ reading whose sole asserted benefit is time saving can be A1; comprehension,
 ↪ synthesis, insight, or knowledge integration is B1.
3. Transformative ambition, new scientific scope, fundamentally changed
 ↪ approaches, or ability to pursue markedly harder or previously inaccessible
 ↪ research directions is C1 More ambitious ideas. A clear claim of
 ↪ revolutionizing a scientific capability can establish this even without the
 ↪ word ambitious. Routine extra task volume or a bare ordinary task does not
 ↪ establish transformative scope.
4. Use a more specific leaf when the contribution clearly states communication
 ↪ improvement, reliability, candidate screening, automated research loops, or
 ↪ cross-disciplinary transfer. A generic execution benefit inside a specific
 ↪ means-to-end contribution is not an additional claim merely because it is
 ↪ enabling that endpoint.
5. O3 Not specified is reserved for a positive acceleration/contribution
 ↪ assertion that does not identify a classifiable task, knowledge contribution,
 ↪ or listed mechanism. It must not become a catch-all for concrete tasks simply
 ↪ because a specific speed or improvement word is absent. General new ideas or
 ↪ ideation can be coded as a task under A1 when that is the identifiable
 ↪ contribution; do not infer transformative scientific ambition from newness
 ↪ alone.
ATOMIC SEGMENTATION
The unit is a distinct positive contribution or effect, not automatically a
 ↪ sentence or a noun. Extract every independently stated contribution. Split
 ↪ independent task families or effects when the response separately asserts
 ↪ them. Do not split synonyms, illustrations of one benefit, or the dependent
 ↪ stages of one means-to-end chain. A list of applications of one coding or
 ↪ analysis operation can be one claim. A proof produced to provide a
 ↪ correctness certificate is one reliability chain; separately listed proof
 ↪ discovery and independent certification can be two claims if their scopes are
 ↪ independent. If one task has independently asserted time-saving and accuracy
 ↪ outcomes, two claims may be appropriate. Never count a summary or repeated
 ↪ elaboration of an earlier contribution again.
Mixed responses may have positive contributions and caveats. Exclude limitations,
 ↪ harms, unmet conditions without an identifiable positive contribution,
 ↪ off-topic instructions, and non-directional/product-only fragments with no
 ↪ identifiable task or benefit. Retain excluded passages and reasons. Do not
 ↪ treat a reliability concern as an achieved reliability benefit. Conditional
 ↪ positive expectations can count when they identify a contribution; a
 ↪ condition followed only by nonspecific change need not count. Every input
 ↪ response must have a coverage record even when it yields no positive claim.
EXCLUSIVE NINE-LEAF CODEBOOK
Assign exactly one leaf per claim. The parent is mechanically determined by that
 ↪ leaf, never independently classified.
A1 | Faster task completion | A: Faster research execution
```

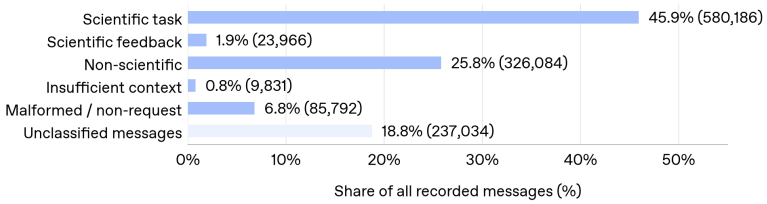
```
Carrying out concrete research tasks or reducing their time, effort, cost, or
↪ labor burden; automation, offloading, parallel work, easier execution, or
↪ additional research work made practicable. Task support qualifies by default
↪ in question context. Prefer a more specific leaf when the stated point is
↪ knowledge/understanding, automated research loops, selection, ambition,
↪ cross-field transfer, communication quality, or correctness. Mere faster
↪ performance of a single test is A1 unless candidate screening or an automated
↪ research loop is stated.
A2 | Automating research loops | A: Faster research execution
AI runs an iterative research process linking ideas or designs to tests,
↪ feedback, and revision. It uses results from one round to choose and carry
↪ out subsequent tests, designs, or revisions. The response must describe or
↪ clearly imply this automated connection between iterations. The loop may
↪ cover part of a project and retain human oversight or approval. Faster
↪ feedback, assistance with one step, one-way execution, or screening many
↪ possibilities without an automated feedback loop is not enough; classify the
↪ stated contribution under another applicable leaf.
B1 | Better informed ideas | B: Improved search for ideas
Knowledge acquisition, explanation, understanding, literature survey/synthesis,
↪ integration of findings or concepts, deeper insight, or evidence/guidance
↪ that helps researchers formulate or improve ideas, questions, methods,
↪ hypotheses, or designs. Knowledge contributions qualify in question context
↪ without a separately stated downstream idea-choice effect. Across explicitly
↪ distinct disciplines/subfields, prefer C2. Pure task execution without a
↪ knowledge/insight contribution defaults to A1.
B2 | Screening ideas | B: Improved search for ideas
Assessing, comparing, validating, falsifying, rejecting, ranking, or prioritizing
↪ candidate ideas, hypotheses, approaches, or designs to judge promise,
↪ correctness, viability, novelty, impact, or feasibility. Counterexamples
↪ explicitly to conjectures and validation explicitly of proposed
↪ ideas/hypotheses can be candidate screening in question context. Generic
↪ numerical validation of a calculation/output, without candidate-idea
↪ assessment, is A1 or O2 as appropriate. Merely producing additional ideas for
↪ humans to screen does not establish improved AI screening; classify the
↪ contribution actually supplied by AI. Testing many possibilities to identify
↪ promising directions or interpret the pattern across candidates can be B2,
↪ while merely executing more tests is A1.
C1 | More ambitious ideas | C: Expanded research directions
Transformative ambition or capability, harder or previously inaccessible
↪ scientific questions, richer theories, new scientific design spaces, major
↪ frontier problems, fundamentally new research approaches, or a clear
↪ transformation of a scientific capability. Context can supply the AI-enabling
↪ link. Generic additional throughput, more ordinary candidates, or a bare new
↪ result does not automatically imply C1.
C2 | Interdisciplinary ideas | C: Expanded research directions
```

```
Transfer or combine knowledge, methods, specialist expertise, or ideas across
 ↪  disciplines or distinct research subfields. Lowering specialist-knowledge
 ↪  barriers can qualify when the response establishes cross-subfield access.
 ↪  Unspecified integration without a disciplinary boundary belongs in B1 if it
 ↪  integrates knowledge, or A1 if it integrates practical tasks/tools.
O1 | Improved communication | Other: Other or unspecified
Improving writing, presentation, exposition, intelligibility, sharing,
 ↪  dissemination, publication, or communication of research. Explicitly better
 ↪  writing or clearer presentation qualifies; removing language barriers
 ↪  qualifies. Bare writing as a production task, or faster writing with no
 ↪  communication-quality effect, is A1. Keep this leaf code O1, distinct from O2
 ↪  reliability.
O2 | Higher reliability | Other: Other or unspecified
Improved correctness, accuracy, rigor, reproducibility, traceability, formal
 ↪  verification, error detection, reliable conclusions, or justified trust in
 ↪  research outputs. Formal proof certification and reproducible pipelines
 ↪  qualify. A generic proof-generation task defaults to A1; explicit
 ↪  candidate-idea falsification or assessment can be B2; broad
 ↪  comprehension/insight can be B1. A concern or unmet requirement alone does
 ↪  not qualify.
O3 | Not specified | Other: Other or unspecified
A positive assertion lacking an identifiable classifiable task, knowledge
 ↪  contribution, or specific listed mechanism. Use sparingly and explain why the
 ↪  contextual task and knowledge defaults do not apply. Never add an O4 or any
 ↪  other new category.
OUTPUT SCHEMA
Output one JSON response record per input in input order:
{
 "response_id": "Rxxx",
 "case_id": "Sxxx",
 "source_text": "exact original input text",
 "claims": [{
 "claim_local_id": "O1",
 "source_excerpt": "exact bounded substring of the original response",
 "paraphrase": "faithful concise English paraphrase",
 "leaf_code": "A1|A2|B1|B2|C1|C2|O1|O2|O3",
 "rationale": "brief evidence-based classification justification, not private
  ↪  reasoning",
 "uncertainty": "low|medium|high",
 "uncertainty_note": "brief boundary caveat or empty string",
 "contextual_inference": true
 }],
 "excluded_content": [{
 "source_excerpt": "exact bounded substring",
 "reason": "limitation|condition|harm|off-topic|no identifiable positive
  ↪  contribution|duplicate elaboration",
 "explanation": "brief reason"
 }],
```

```
 "response_note": "brief segmentation/coverage note or empty string"
}
Use sequential two-digit local claim IDs beginning 01 within each response. Set
↪  contextual_inference=true when the question supplies the acceleration/benefit
↪  meaning instead of the answer explicitly asserting it. Translate non-English
↪  text only in the paraphrase; preserve original-language exact excerpts.
↪  Preserve meaningful independent claims but remove repeated summaries. Review
↪  every O3 claim to ensure it is not an ordinary task, knowledge contribution,
↪  or transformative contribution that the general rules cover. Check each
↪  excerpt is a literal substring. Return only the JSON response record; do not
↪  add prose outside JSON.
```

Appendix figures and tables

Figure A1: Composition of recorded messages



Notes: The sample comprises 1,262,893 messages sent from August 20 to September 22, 2026, by 1,819 scientists. Bars show the percentage of messages in each message category; counts appear in parentheses. Scientific task is a substantive request for scientific work. Scientific feedback is an evaluative reaction to prior AI support for scientific work without a new substantive request. Non-scientific is an understandable request or evaluative reaction unrelated to scientific or scholarly work. Insufficient context means the requested action or its target cannot be determined from the available context. Malformed / non-request is a message without an identifiable request or context-supported evaluative reaction. Unclassified messages are the 237,034 messages without an available classification. Appendix A.3.2 defines the classification rules.

Table A1: Daily scientific AI use across scientists

Group	Mean daily scientific messages	Median daily scientific messages [25th–75th percentile]	Scientists
Field			
Biology	9.2	4.9 [2.0, 11.8]	322
Chemistry and materials sciences	10.4	6.2 [2.6, 13.2]	200
Computer science	14.1	7.7 [3.1, 16.6]	196
Earth and planetary sciences	15.1	8.3 [4.0, 19.2]	77
Engineering	10.9	6.1 [2.6, 14.2]	263
Mathematics	12.3	6.2 [2.6, 14.0]	368
Physics	10.0	7.1 [3.0, 12.4]	362
Scientist age			
18–29	12.5	7.1 [3.0, 14.4]	580
30–39	11.7	6.9 [2.9, 14.6]	718
40+	7.9	4.6 [1.8, 10.0]	342

Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. The table lists each field or age group, the mean and median of scientists' daily scientific-message rates, the 25th and 75th percentiles in brackets, and the number of scientists. Each scientist's daily average divides their scientific messages by the number of calendar days from their first observed scientific message through September 22. Fields come from scientists' own applications when available. Collaborators without an available application field receive the field reported in their workspace owner's application. Ages are self-reported at signup. The field comparison includes 1,788 scientists, excluding 19 classified as Other and 12 with missing fields. The age comparison includes 1,640 scientists, excluding 179 with missing ages.

Table A2: Scientific tasks by workflow

Task	Task definition	Source
Reading & knowledge acquisition		
Explain scientific concepts	Learn, explain, interpret or clarify established scientific concepts such as research methods, tools or worked examples.	Reworded from SciNet
Retrieve scholarly information	Locate, retrieve, or extract facts, passages, citations, terminology, or new findings from scholarly sources.	Added
Review and synthesize literature	Review and synthesize the relevant literature to situate a research question within existing knowledge.	SciNet verbatim
Ideation & hypothesis generation		
Formulate research problems, models or approaches	Formulate, reframe, or modify a research problem, conjecture, theoretical model, or research approach.	Added
Brainstorm research directions	Brainstorm potential research directions through discussions with colleagues and peers.	SciNet verbatim
Generate hypotheses	Generate hypotheses based on theoretical frameworks, observations, or preliminary data.	SciNet verbatim
Identify gaps in the literature	Identify gaps in existing literature to formulate novel research questions.	SciNet verbatim
Formulate research questions from observations	Use observed real-world phenomena or patterns to formulate research questions or hypotheses.	Reworded from SciNet
Administration		
Manage budgets and personnel	Manage research budgets, personnel, and project timelines.	SciNet verbatim
Coordinate research logistics	Coordinate research-project operations or logistics with collaborators across institutions and disciplines.	Reworded from SciNet
Maintain research infrastructure	Maintain laboratory equipment or non-software research infrastructure.	Reworded from SciNet
Obtain ethics approvals	Obtain ethical approvals and informed consent for studies involving human subjects.	SciNet verbatim
Data gathering		
Curate datasets	Curate and manage datasets, ensuring proper documentation and version control.	SciNet verbatim
Access and acquire data	Acquire or access primary data sources relevant to research questions.	SciNet verbatim
Design data collection	Design data collection protocols and instruments to ensure validity and reproducibility.	SciNet verbatim
Research software & code development		
Implement research software	Implement research software, computational methods, or scientific programming features.	Added
Maintain programming environments and pipelines	Configure or maintain research programming environments, dependencies, and computational pipelines.	Added
Review and verify code	Review, test, or verify research code for correctness and reproducibility.	Added
Debug research code	Debug, diagnose, or correct errors in research code or computational workflows.	Added

Continued on next page

Table A2: Scientific tasks by workflow (continued)

Task	Task definition	Source
Refactor or optimize software	Refactor, optimize, or improve research software and computational performance.	Added
Scientific analysis		
Develop or test proofs	Prove or disprove a claim, develop or compare proof strategies, or construct or enumerate formal examples and counterexamples.	Added
Derive formal results	Derive, simplify, or characterize a mathematical or formal result.	Added
Analyze formal model properties	Analyze, test, or validate the formal properties of a theoretical or computational model.	Added
Run statistical analyses	Apply statistical methods to test hypotheses and estimate effect sizes.	SciNet verbatim
Visualize data and results	Visualize data and results using graphs, tables, and figures for interpretation.	SciNet verbatim
Develop empirical models	Develop and validate data-driven computational models using empirical data to simulate or predict phenomena.	Reworded from SciNet
Establish bounds and complexity	Establish, tighten, or compare bounds and computational complexity results.	Added
Clean and preprocess data	Clean and preprocess raw data to prepare for statistical or computational analysis.	SciNet verbatim
Writing & communication		
Draft manuscripts	Draft manuscripts describing research questions, methods, results, and interpretations.	SciNet verbatim
Draft grant narratives	Draft or revise the scholarly narrative of grant proposals to secure research funding.	Reworded from SciNet
Create research visuals	Create, edit, crop, annotate, or improve research figures, charts, diagrams, presentation visuals, or paper visuals.	Added
Respond to reviewers	Respond to peer reviewer comments and revise manuscripts for publication.	SciNet verbatim
Present research findings	Present research findings at conferences, seminars, and symposia.	SciNet verbatim
Peer review & service		
Evaluate manuscripts	Evaluate manuscripts submitted to journals for scientific rigor and contribution.	SciNet verbatim
Serve on review panels	Serve on editorial boards, grant review panels, or professional committees.	SciNet verbatim
End-to-end autonomous research		
Request end-to-end research	Autonomously coordinate reading and knowledge acquisition, ideation and hypothesis generation, and either empirical data gathering and analysis or formal analysis in an end-to-end scientific-discovery workflow.	Added
Mentorship & teaching		
Support teaching and mentoring	Prepare or revise teaching materials, assignments, or assessments; answer student questions; or guide and supervise students, research assistants, and postdoctoral researchers.	Added

Continued on next page

Table A2: Scientific tasks by workflow (continued)

Task	Task definition	Source
Develop course curricula	Develop and deliver course curricula in specialized research areas.	SciNet verbatim
Review student research	Provide feedback on student theses, dissertations, and research proposals.	SciNet verbatim
Supervise researchers	Supervise graduate students and postdoctoral researchers on research projects.	SciNet verbatim

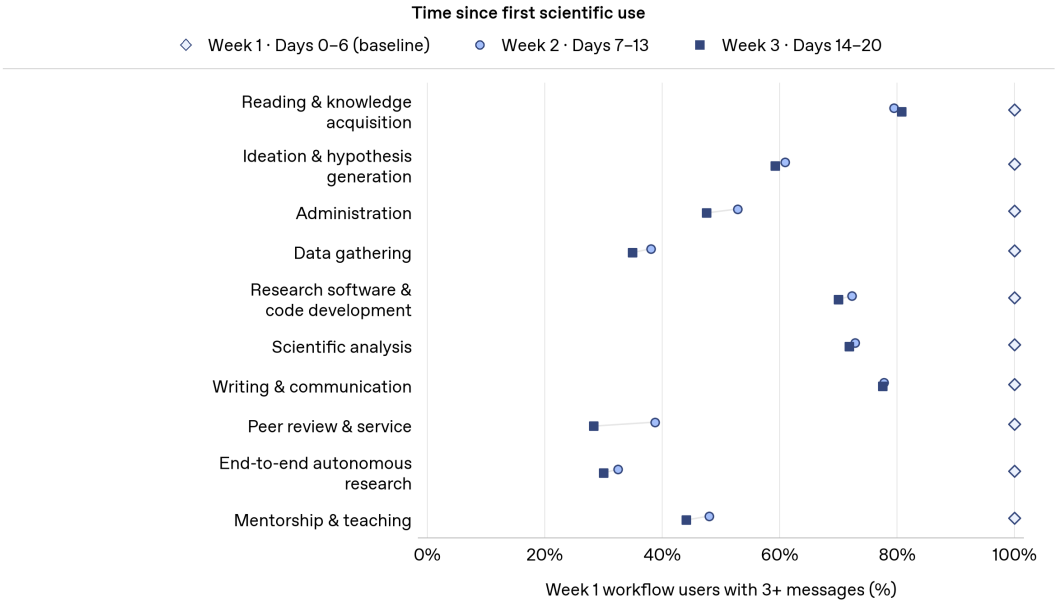
Notes: The table lists each scientific task, its definition, and the source of its definition, grouped by scientific workflow. SciNet verbatim denotes a definition reproduced from the SciNet taxonomy; Reworded from SciNet denotes an adapted definition; Added denotes a task added to the classification scheme. Scientific analysis combines Empirical analysis and Theoretical and formal analysis. Other scientific task covers scientific work outside the 40 listed tasks. Workflows are ordered in the approximate order we believe they first occur in the scientific research process.

Table A3: Workflow uptake by message threshold

Workflow	**≥1**	**≥3**	**≥10**	**≥25**	**≥50**
Reading & knowledge acquisition	97.9%	94.2%	82.9%	63.2%	43.7%
Writing & communication	94.6%	88.5%	75.8%	58.8%	41.6%
Scientific analysis	89.1%	79.0%	61.9%	43.1%	26.8%
Ideation & hypothesis generation	84.1%	67.6%	40.2%	19.9%	7.4%
Research software & code development	73.4%	61.7%	48.3%	34.4%	23.5%
Administration	64.2%	42.2%	18.9%	6.6%	2.4%
Peer review & service	53.0%	33.6%	12.7%	4.2%	1.7%
Mentorship & teaching	52.2%	32.2%	15.4%	5.8%	2.7%
Data gathering	41.0%	23.2%	7.9%	2.1%	0.4%
Other scientific work	34.8%	9.8%	1.2%	0.3%	0.1%
End-to-end autonomous research	15.3%	6.3%	2.2%	1.0%	0.5%

Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. The table lists each scientific workflow and the percentage of scientists who sent at least the number of scientific messages indicated by each column—one, three, ten, 25, or 50—in that workflow during August 20–September 22. Workflows are ordered by descending percentage of scientists who sent at least one scientific message.

Figure A2: Return to AI use by scientific workflow



Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. Week 1 begins on each scientist's first observed scientific-message date and includes the following six calendar days. Weeks 2 and 3 are the next two seven-day periods. For each workflow, the figure shows return to AI use among the fixed group of scientists who sent at least 3 scientific messages in that workflow in Week 1: the percentage who sent at least 3 scientific messages in the same workflow during each week. Workflows are ordered in the approximate order we believe they first occur in the scientific research process.

Table A4: Scientific task diffusion

Task	Task share of scientific messages	Share of scientists using AI for each task
Reading & knowledge acquisition		
Explain scientific concepts	18.8%	89.7%
Retrieve scholarly information	5.3%	70.3%
Review and synthesize literature	3.6%	61.4%
Ideation & hypothesis generation		
Formulate research problems, models or approaches	3.9%	61.1%
Brainstorm research directions	1.1%	30.6%
Generate hypotheses	0.2%	6.6%
Identify gaps in the literature	$< 0.1\%$	1.9%
Formulate research questions from observations	$< 0.1\%$	0.0%
Administration		
Manage budgets and personnel	1.0%	22.8%
Coordinate research logistics	0.9%	20.3%
Maintain research infrastructure	0.2%	2.6%
Obtain ethics approvals	$< 0.1\%$	0.9%
Data gathering		
Curate datasets	0.4%	9.7%
Access and acquire data	0.4%	13.5%
Design data collection	0.2%	5.3%
Research software & code development		
Implement research software	4.7%	49.4%
Maintain programming environments and pipelines	2.7%	38.8%
Review and verify code	2.5%	36.9%
Debug research code	2.3%	34.4%
Refactor or optimize software	1.5%	27.7%
Scientific analysis		
Develop or test proofs	4.2%	26.2%
Derive formal results	3.4%	39.8%
Analyze formal model properties	3.3%	40.6%
Run statistical analyses	1.6%	31.1%
Visualize data and results	1.5%	31.5%
Develop empirical models	1.2%	25.0%
Establish bounds and complexity	0.9%	13.2%
Clean and preprocess data	0.4%	10.1%
Writing & communication		
Draft manuscripts	13.9%	81.5%
Draft grant narratives	4.7%	32.4%
Create research visuals	4.2%	55.4%
Respond to reviewers	1.8%	27.9%
Present research findings	1.8%	31.2%
Peer review & service		
Evaluate manuscripts	1.7%	28.2%
Serve on review panels	0.4%	8.5%
End-to-end autonomous research		
Request end-to-end research	0.4%	6.3%
Mentorship & teaching		
Support teaching and mentoring	2.7%	28.8%

Continued on next page

Table A4: Scientific task diffusion (continued)

Task	Task share of scientific messages	Share of scientists using AI for each task
Develop course curricula	0.2%	4.3%
Review student research	0.2%	2.9%
Supervise researchers	$< 0.1\%$	1.8%

Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. The table lists each scientific task, its task share of scientific messages, and the share of scientists using AI for that task, grouped by workflow. The task share of scientific messages is the mean of each scientist's share of all scientific messages devoted to that task. The share of scientists using AI for each task is the percentage who sent at least 3 scientific messages in that task during August 20–September 22. Other scientific task and messages without a task classification remain in the denominators. Tasks within each workflow are ordered by descending task share of scientific messages. Workflows are ordered in the approximate order we believe they first occur in the scientific research process.

Figure A3: Return to AI use by scientific task



Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. Week 1 begins on each scientist's first observed scientific-message date and includes the following six calendar days. Weeks 2 and 3 are the next two seven-day periods. For each task, the figure shows return to AI use among the fixed group of scientists who sent at least 3 scientific messages in that task in Week 1: the percentage who sent at least 3 scientific messages in the same task during each week. Tasks are grouped by workflow. Workflows are ordered in the approximate order we believe they first occur in the scientific research process.

Figure A4: AI use across scientific tasks by field and age

A. Scientific field

Color: % difference from each task's overall mean

–100% 0 +100% ≥200%

Scientific task	Overall	Biology	Chemistry / materials	Computer science	Earth / planetary	Engineering	Mathematics	Physics
Reading & knowledge acquisition								
Explain scientific concepts	18.8	21.1	18.4	15.9	17.5	19.2	16.9	21.0
Retrieve scholarly information	5.3	7.1	5.0	4.3	5.4	5.3	5.1	4.2
Review and synthesize literature	3.6	4.6	3.6	3.7	3.5	3.1	3.4	3.2
Ideation & hypothesis generation								
Formulate research questions from observations	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1
Identify gaps in the literature	<0.1	<0.1	<0.1	0.1	0.1	<0.1	<0.1	<0.1
Brainstorm research directions	1.1	0.9	1.0	1.4	1.0	1.1	1.1	1.3
Generate hypotheses	0.2	0.5	0.4	<0.1	0.2	<0.1	<0.1	0.3
Formulate research problems, models or approaches	3.9	4.0	3.6	4.5	4.3	3.7	3.8	3.7
Administration								
Manage budgets and personnel	1.0	1.5	1.2	0.9	1.3	1.1	0.5	0.7
Coordinate research logistics	0.9	1.3	0.9	0.8	0.8	1.2	0.5	0.7
Maintain research infrastructure	0.2	0.3	0.6	<0.1	<0.1	0.2	<0.1	0.1
Obtain ethics approvals	<0.1	0.1	<0.1	0.2	<0.1	<0.1	<0.1	<0.1
Data gathering								
Curate datasets	0.4	1.2	0.3	0.3	1.3	0.3	<0.1	0.1
Access and acquire data	0.4	0.8	0.4	0.5	1.1	0.2	0.1	0.2
Design data collection	0.2	0.5	0.3	0.2	0.2	0.2	<0.1	0.1
Research software & code development								
Implement research software	4.7	4.2	6.7	5.2	7.3	5.2	2.1	5.5
Maintain programming environments and pipelines	2.7	2.7	5.3	3.7	3.5	2.2	1.3	2.5
Review and verify code	2.5	2.3	2.8	2.8	3.6	3.2	0.9	3.0
Debug research code	2.3	2.8	3.1	2.1	3.8	2.2	0.8	2.8
Refactor or optimize software	1.5	1.2	1.8	2.0	2.1	1.8	0.6	1.9
Scientific analysis								
Develop or test proofs	4.2	0.2	<0.1	4.3	0.2	0.6	15.3	2.0
Derive formal results	3.4	0.4	0.9	1.9	1.5	1.6	6.8	6.8
Analyze formal model properties	3.3	0.4	0.6	3.0	0.7	1.7	8.0	4.5
Run statistical analyses	1.6	3.5	1.8	1.2	2.9	0.8	0.2	1.7
Visualize data and results	1.5	2.2	2.1	1.1	2.7	1.3	0.4	1.9
Develop empirical models	1.2	1.3	2.6	1.0	2.7	1.3	0.3	1.1
Establish bounds and complexity	0.9	<0.1	<0.1	1.5	<0.1	0.3	2.6	0.5
Clean and preprocess data	0.4	1.1	0.5	0.2	0.9	0.2	<0.1	0.2
Writing & communication								
Draft manuscripts	13.9	11.7	12.2	16.6	13.5	15.7	15.0	12.9
Draft grant narratives	4.7	6.4	6.4	3.8	4.3	5.9	2.2	4.5
Create research visuals	4.2	6.0	5.6	4.6	3.6	5.3	1.5	3.8
Respond to reviewers	1.8	1.9	2.7	2.4	2.0	2.3	1.2	1.2
Present research findings	1.8	1.7	3.0	1.5	1.6	2.5	0.9	1.6
Peer review & service								
Evaluate manuscripts	1.7	1.2	1.2	2.9	1.3	2.3	1.7	1.3
Serve on review panels	0.4	0.4	0.3	0.6	0.2	0.7	0.2	0.3
End-to-end autonomous research								
Request end-to-end research	0.4	0.2	0.1	0.2	0.3	0.3	0.9	0.2
Mentorship & teaching								
Support teaching and mentoring	2.7	1.8	1.7	2.5	2.3	4.0	3.6	2.3
Develop course curricula	0.2	<0.1	<0.1	0.3	0.1	0.3	0.2	0.1
Review student research	0.2	0.1	<0.1	<0.1	0.1	0.4	0.2	<0.1
Supervise researchers	<0.1	<0.1	<0.1	0.1	0.1	0.1	0.1	<0.1
Other / unknown								
Other scientific task	1.6	2.1	2.3	1.3	1.7	1.7	1.0	1.1
Unknown task	0.3	0.2	0.2	0.3	0.3	0.3	0.2	0.3

Notes: The sample comprises 580,186 scientific messages sent from August 20 to September 22, 2026, by 1,819 scientists. Overall includes all 1,819 scientists. Panel A compares 7 fields among 1,788 scientists, excluding 19 classified as Other and 12 with missing fields. Panel B compares 3 age groups among 1,640 scientists, excluding 179 with missing ages.

Figure A4 (continued): AI use across scientific tasks by field and age

B. Scientist age

Color: % difference from each task's overall mean

−100% 0 +100% ≥200%

Scientific task	Overall	18–29	30–39	40+
Reading & knowledge acquisition				
Explain scientific concepts	18.8	21.7	18.2	16.2
Retrieve scholarly information	5.3	4.7	5.3	6.1
Review and synthesize literature	3.6	3.4	3.6	4.4
Ideation & hypothesis generation				
Formulate research questions from observations	<0.1	<0.1	<0.1	<0.1
Identify gaps in the literature	<0.1	0.1	<0.1	<0.1
Brainstorm research directions	1.1	1.0	1.1	1.3
Generate hypotheses	0.2	0.2	0.2	0.1
Formulate research problems, models or approaches	3.9	3.9	3.7	4.2
Administration				
Manage budgets and personnel	1.0	0.5	1.2	1.3
Coordinate research logistics	0.9	0.6	0.9	1.0
Maintain research infrastructure	0.2	0.2	0.2	<0.1
Obtain ethics approvals	<0.1	<0.1	<0.1	0.1
Data gathering				
Curate datasets	0.4	0.6	0.3	0.5
Access and acquire data	0.4	0.5	0.4	0.2
Design data collection	0.2	0.3	0.2	<0.1
Research software & code development				
Implement research software	4.7	6.3	4.1	3.2
Maintain programming environments and pipelines	2.7	3.2	2.4	2.4
Review and verify code	2.5	2.9	2.4	1.9
Debug research code	2.3	3.3	1.9	1.3
Refactor or optimize software	1.5	2.0	1.4	0.9
Scientific analysis				
Develop or test proofs	4.2	3.4	4.0	6.7
Derive formal results	3.4	3.0	3.4	4.2
Analyze formal model properties	3.3	2.9	3.1	4.8
Run statistical analyses	1.6	2.0	1.7	0.7
Visualize data and results	1.5	2.0	1.6	0.9
Develop empirical models	1.2	1.7	1.2	0.6
Establish bounds and complexity	0.9	0.7	0.8	1.3
Clean and preprocess data	0.4	0.5	0.4	0.2
Writing & communication				
Draft manuscripts	13.9	13.9	14.1	13.5
Draft grant narratives	4.7	1.6	6.2	6.1
Create research visuals	4.2	4.7	4.5	2.8
Respond to reviewers	1.8	1.3	2.0	2.1
Present research findings	1.8	1.9	1.6	1.5
Peer review & service				
Evaluate manuscripts	1.7	1.4	1.6	1.6
Serve on review panels	0.4	<0.1	0.5	0.7
End-to-end autonomous research				
Request end-to-end research	0.4	0.3	0.4	0.3
Mentorship & teaching				
Support teaching and mentoring	2.7	1.5	2.9	3.7
Develop course curricula	0.2	<0.1	0.2	0.3
Review student research	0.2	<0.1	0.2	0.3
Supervise researchers	<0.1	<0.1	0.1	<0.1
Other / unknown				
Other scientific task	1.6	1.1	1.6	1.9
Unknown task	0.3	0.3	0.3	0.2

Notes (continued): Cells show task shares of scientific messages within each group. The task share of scientific messages is the mean of each scientist's share of all scientific messages devoted to that task. Other scientific task and messages without a task classification remain in the denominators. Color intensity is capped at +200%. Tasks are grouped by workflow. Workflows are ordered in the approximate order we believe they first occur in the scientific research process.

Table A5: Scientist involvement across analyzed scientific tasks

Task	Workflow	Involvement score (%)
Visualize data and results	Scientific analysis	41.1%
Respond to reviewers	Writing & communication	39.2%
Implement research software	Research software & code development	38.9%
Draft grant narratives	Writing & communication	38.5%
Present research findings	Writing & communication	37.8%
Draft manuscripts	Writing & communication	37.3%
Debug research code	Research software & code development	37.2%
Create research visuals	Writing & communication	37.0%
Analyze formal model properties	Scientific analysis	37.0%
Derive formal results	Scientific analysis	34.8%
Formulate research problems, models or approaches	Ideation & hypothesis generation	33.9%
Review and verify code	Research software & code development	33.8%
Develop empirical models	Scientific analysis	33.5%
Formulate research questions from observations	Ideation & hypothesis generation	33.3%
Explain scientific concepts	Reading & knowledge acquisition	33.2%
Refactor or optimize software	Research software & code development	32.8%
Evaluate manuscripts	Peer review & service	32.7%
Run statistical analyses	Scientific analysis	32.5%
Generate hypotheses	Ideation & hypothesis generation	32.4%
Design data collection	Data gathering	32.1%
Establish bounds and complexity	Scientific analysis	31.5%
Develop or test proofs	Scientific analysis	29.9%
Obtain ethics approvals	Administration	29.2%
Manage budgets and personnel	Administration	29.1%
Clean and preprocess data	Scientific analysis	27.9%
Serve on review panels	Peer review & service	27.4%
Support teaching and mentoring	Mentorship & teaching	27.1%
Review student research	Mentorship & teaching	27.1%
Develop course curricula	Mentorship & teaching	26.4%
Maintain programming environments and pipelines	Research software & code development	26.1%
Access and acquire data	Data gathering	25.8%
Maintain research infrastructure	Administration	25.8%
Brainstorm research directions	Ideation & hypothesis generation	25.6%
Identify gaps in the literature	Ideation & hypothesis generation	25.0%
Curate datasets	Data gathering	24.8%
Review and synthesize literature	Reading & knowledge acquisition	24.3%
Retrieve scholarly information	Reading & knowledge acquisition	24.1%

Continued on next page

Table A5: Scientist involvement across analyzed scientific tasks (continued)

Task	Workflow	Involvement score (%)
Request end-to-end research	End-to-end autonomous research	23.7%
Supervise researchers	Mentorship & teaching	22.7%
Coordinate research logistics	Administration	21.8%

Notes: The sample comprises 125,947 initial requests sent from August 20 to September 15, 2026, by 1,809 scientists. An initial request is a scientific message that starts work without relying on prior AI output. We include initial requests assigned to one of the 40 scientific tasks, with 7 complete days of follow-up and enough information to determine whether involvement occurred. A scientist's involvement score is the percentage of these requests for which the next message classified as a scientific task in the same conversation arrives within 7 days and requests clarification, checking, or revision of the original output. If no message classified as a scientific task follows within 7 days, involvement is zero. Scores are calculated separately within each task. The table lists each scientific task, its scientific workflow, and its average involvement score among scientists with at least one such initial request in the task. Tasks are ordered by descending average involvement score.

Table A6: Scientist involvement by task requirements across analyzed tasks

Task	Lower-group involvement score (%)	N	Higher-group involvement score (%)	N	Difference (pp)
Panel A. Required expertise					
Explain scientific concepts	31.3%	1,117	34.6%	1,397	3.3
Retrieve scholarly information	22.6%	930	27.4%	771	4.8
Review and synthesize literature	22.5%	40	24.4%	1,048	1.9
Formulate research problems, models or approaches	25.0%	32	34.0%	799	9.0
Brainstorm research directions	16.7%	30	25.9%	490	9.2
Generate hypotheses	0.0%	2	33.0%	112	—
Identify gaps in the literature	—	0	25.0%	79	—
Formulate research questions from observations	—	0	33.3%	6	—
Manage budgets and personnel	28.2%	345	41.2%	46	12.9
Coordinate research logistics	20.5%	331	28.2%	49	7.7
Maintain research infrastructure	27.1%	63	14.3%	7	—
Obtain ethics approvals	29.5%	32	33.3%	3	—
Curate datasets	22.9%	128	34.2%	37	11.3
Access and acquire data	24.6%	162	30.2%	47	5.7
Design data collection	21.0%	31	38.3%	65	17.4
Implement research software	39.9%	498	39.7%	486	-0.2
Maintain programming environments and pipelines	26.0%	463	30.1%	112	4.1
Review and verify code	30.5%	377	37.6%	351	7.1
Debug research code	35.3%	362	40.5%	160	5.2
Refactor or optimize software	31.7%	237	34.2%	203	2.5
Develop or test proofs	38.2%	90	29.4%	403	-8.9
Derive formal results	26.3%	200	37.2%	468	10.9
Analyze formal model properties	30.2%	52	37.6%	495	7.4
Run statistical analyses	25.0%	139	35.9%	299	10.8
Visualize data and results	40.2%	260	43.6%	254	3.3
Develop empirical models	32.7%	49	34.2%	243	1.4
Establish bounds and complexity	11.1%	9	31.9%	150	—
Clean and preprocess data	26.9%	95	32.0%	56	5.1
Draft manuscripts	33.6%	1,056	44.2%	1,018	10.6
Draft grant narratives	35.1%	391	43.6%	432	8.5
Create research visuals	36.7%	689	42.9%	302	6.2
Respond to reviewers	34.8%	268	42.9%	358	8.1
Present research findings	34.4%	391	42.9%	330	8.5
Evaluate manuscripts	11.1%	27	33.3%	601	22.1
Serve on review panels	19.9%	50	29.2%	174	9.3

Continued on next page

Table A6: Scientist involvement by task requirements across analyzed tasks (continued)

Task	Lower-group involvement score (%)	N	Higher-group involvement score (%)	N	Difference (pp)
Request end-to-end research	—	0	23.8%	156	—
Support teaching and mentoring	27.8%	468	26.7%	268	-1.1
Develop course curricula	32.3%	50	22.4%	63	-10.0
Review student research	15.7%	17	28.9%	106	13.2
Supervise researchers	33.3%	9	19.7%	39	—
Panel B. Expected duration					
Explain scientific concepts	33.1%	1,459	29.0%	131	-4.1
Retrieve scholarly information	24.1%	1,046	21.7%	185	-2.4
Review and synthesize literature	27.8%	271	25.0%	650	-2.7
Formulate research problems, models or approaches	31.4%	291	30.8%	330	-0.6
Brainstorm research directions	19.5%	162	26.0%	149	6.4
Generate hypotheses	22.0%	41	28.0%	25	6.0
Identify gaps in the literature	0.0%	7	26.1%	66	—
Formulate research questions from observations	50.0%	2	0.0%	1	—
Manage budgets and personnel	27.1%	307	53.8%	24	26.6
Coordinate research logistics	20.0%	309	25.0%	12	5.0
Maintain research infrastructure	25.6%	53	0.0%	4	—
Obtain ethics approvals	24.1%	27	0.0%	1	—
Curate datasets	20.2%	81	23.8%	41	3.6
Access and acquire data	24.6%	124	40.9%	42	16.2
Design data collection	30.0%	30	36.0%	19	6.0
Implement research software	44.4%	276	36.2%	494	-8.2
Maintain programming environments and pipelines	25.6%	367	21.6%	134	-4.0
Review and verify code	29.4%	317	36.8%	241	7.5
Debug research code	35.7%	288	32.8%	98	-2.9
Refactor or optimize software	27.5%	128	31.3%	218	3.7
Develop or test proofs	35.2%	203	23.6%	315	-11.6
Derive formal results	32.6%	398	36.5%	185	4.0
Analyze formal model properties	43.7%	302	31.9%	245	-11.8
Run statistical analyses	31.8%	199	42.2%	140	10.5
Visualize data and results	38.9%	259	48.9%	97	10.0
Develop empirical models	20.5%	63	31.7%	167	11.3
Establish bounds and complexity	40.0%	39	27.4%	116	-12.5
Clean and preprocess data	29.7%	79	26.9%	32	-2.7
Draft manuscripts	36.9%	1,094	39.7%	545	2.8

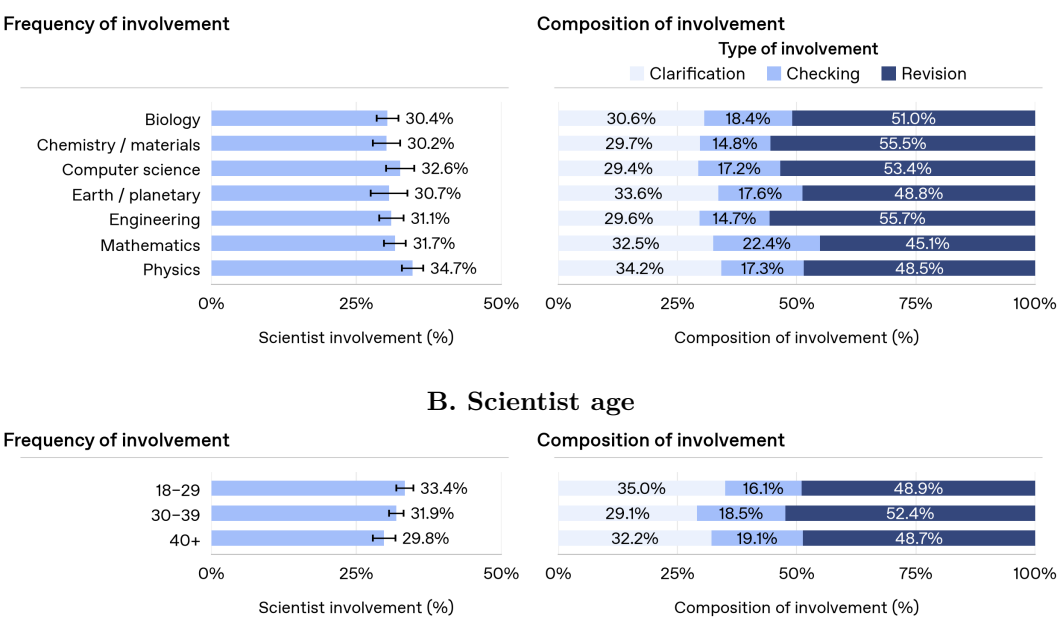
Continued on next page

Table A6: Scientist involvement by task requirements across analyzed tasks (continued)

Task	Lower-group involvement score (%)	N	Higher-group involvement score (%)	N	Difference (pp)
Draft grant narratives	36.6%	419	41.6%	226	5.0
Create research visuals	36.0%	604	40.9%	155	4.9
Respond to reviewers	37.2%	318	42.1%	145	4.9
Present research findings	36.3%	404	44.7%	127	8.4
Evaluate manuscripts	34.7%	135	29.0%	374	-5.7
Serve on review panels	23.2%	76	24.7%	79	1.5
Request end-to-end research	0.0%	2	23.4%	155	—
Support teaching and mentoring	26.5%	413	32.9%	179	6.4
Develop course curricula	21.7%	40	27.1%	35	5.5
Review student research	30.0%	35	30.5%	40	0.5
Supervise researchers	26.1%	23	0.0%	6	—

Notes: The sample comprises 125,947 initial requests sent from August 20 to September 15, 2026, by 1,809 scientists. An initial request is a scientific message that starts work without relying on prior AI output. We include initial requests assigned to one of the 40 scientific tasks, with 7 complete days of follow-up and enough information to determine whether involvement occurred. A scientist's involvement score is the percentage of these requests for which the next message classified as a scientific task in the same conversation arrives within 7 days and requests clarification, checking, or revision of the original output. If no message classified as a scientific task follows within 7 days, involvement is zero. Panel A compares general/foundational or standard technical/professional expertise with advanced research or frontier/subfield specialist expertise. Panel B compares expected duration under one hour with four hours or more, measured as the active time a competent, experienced person would need to complete the requested task without AI. Scores are calculated separately within each task and comparison group. The table lists each scientific task, the lower-group average involvement score and contributing scientist count (N), the corresponding higher-group values, and the difference in percentage points, calculated as the higher-group score minus the lower-group score. Each task-level score averages scientists' involvement scores among scientists with at least one eligible initial request in that task and comparison group. Differences are reported only when both groups include at least 10 scientists—32 tasks in Panel A and 34 in Panel B—and are otherwise marked —.

Figure A5: Frequency and composition of scientist involvement by field and age



Notes: The sample comprises 125,947 initial requests sent from August 20 to September 15, 2026, by 1,809 scientists. An initial request starts scientific work without relying on prior AI output. We include requests assigned to one of the 40 scientific tasks, with 7 complete days of follow-up and sufficient information to determine involvement. A scientist's involvement score is the percentage of these requests whose next scientific-task message in the same conversation arrives within 7 days and seeks clarification, checking, or revision of the original output. Panel A compares 7 fields, excluding Other and missing fields. Panel B compares 3 age groups, excluding missing ages. In each panel, the left chart shows the average involvement score among scientists with at least one such initial request in the group, with pointwise 95% scientist-level t confidence intervals. The right chart shows the composition of involvement. Clarification, checking, and revision shares are calculated within each scientist's requests with involvement in that group, then averaged across scientists with at least one such request in the group.