

ANTHROP\C

AI for Scientific Discovery: How Researchers Use Claude

Published
November 5, 2026

Authors
Johannes Hermle,* Benjamin Jones,*
Lukas Althoff,* Ryan Heller, Thomas Lyttelton,
Carolyn Stein, Nathan Wilmers,
Maxim Massenkoff, Peter McCrory

Acknowledgments

Cristian Alamos, Bailey Bova, Tom Chivers, Lauren Cottrell, Eleanor Dorfman, Adam Farina, Deep Ganguli, Rebecca Hiscott, Chad Jones, Anneke Jong, Aisling Keenan, Anton Korinek, Eva Lyubich, Paruul Maheshwary, Miles McCain, Teddy McGuire, Ankur Rathi, Santi Ruiz, Szymon Sacher, David Saunders, Monika Tuchowska, Nick Turner, Heather Whitney, Kim Withee, Jack Clark.

*Lead authors of the report

Summary

- **What we measure.** AI could speed up scientific discovery, but how much depends on what AI can do and how scientists use it. Using privacy-preserving methods, we study 207,000 Claude and Claude Code research sessions to understand how Claude is used across scientific fields and research tasks. We are conducting this research to help the public better understand the state of AI for science today and assess the impact that AI may have on scientific discovery in the future.
- **Where scientists use Claude.** Relative to the number of researchers in each field, Claude is used more in the formal sciences, such as computer science, mathematics, and statistics, and less in health and lab-based physical sciences. Scientists primarily use Claude to analyze data and to write. Data analysis alone accounts for about one-third of sessions on Claude and nearly half of sessions on Claude Code. Scientists use it less for idea generation and data gathering, even though, in an accompanying survey, they rank those tasks as the most important for scientific discovery. That said, specialized AI tools designed for scientific work could narrow the gap between the research tasks scientists use AI for and the ones they consider most important. On Claude Science, for instance, idea generation makes up about twice the share of use that it does on Claude.
- **How Claude use tracks AI capability.** We create a measure of how well AI can, in theory, perform each research task. The tasks AI is best suited to (the top 20% on this measure) get two to four times as much Claude use as the time researchers spend on them would predict. The tasks AI is least suited to (the bottom 20%), which include most physical work, get very little. We observe the same pattern across fields: fields where AI can perform many tasks well, such as computer science and statistics, use Claude more.
- **What this could mean for the speed of scientific discovery.** Using a simple model, we explore scenarios for how much AI could speed up research. If existing AI capabilities are widely adopted, they could double research output in some fields. If AI could carry out all non-physical research work, output in fields such as computer science and statistics could rise as much as 15-fold, while output in bench sciences such as chemistry and biology, which are held back by physical bottlenecks, would rise three- or fourfold. In our survey of researchers who actively use AI, about 40% expect AI models to carry out full research projects on their own by 2030.

Introduction

Science has been described as an "endless frontier" (Bush 1945). But how fast we explore that frontier has historically been limited by a scarce resource: the time and talent of human researchers. Artificial intelligence could ease that limit by amplifying what researchers can do, transforming both the process of discovery and the pace of exploration (Wang et al. 2023; Amodei 2024). Few applications of AI could matter more for human welfare. Given the highly cognitive nature of research compared with most sectors of the economy, AI's effects may be particularly large and early in this domain.

Recent examples show AI playing a growing role in scientific discovery. In structural biology, AlphaFold predicted the three-dimensional shapes of over 200 million proteins (Jumper et al. 2021; Varadi et al. 2024), and its creators shared the 2024 Nobel Prize in Chemistry. In materials science, GNoME predicted 380,000 stable new crystals (Merchant et al. 2023). In mathematics, researchers have used AI systems to help resolve decades-old Erdős problems, produce a potential proof for a Millennium Prize problem, and make progress on other major problems, from open conjectures to verifying landmark proofs (Anthropic 2026a, 2026b; OpenAI 2026a, 2026b; Tsoukalas et al. 2026).

Beyond these developments, most scientific progress involves many specialized, interlinked research tasks, which can vary substantially across fields. How much AI accelerates scientific discovery depends not only on AI's capabilities, but also on how it is adopted across these tasks. In this report, we use privacy-preserving infrastructure to study the scientific use of Claude, Claude Code, and Claude Science. We document how usage varies across fields and tasks, identify what explains these usage patterns, and outline scenarios for how AI could accelerate research.

Claude use in science varies by field. Relative to their number of researchers, formal sciences such as computer science, mathematics, and statistics use Claude more heavily, while health and lab-based physical sciences use it less. Across research tasks, scientists use Claude more for analyzing data and writing, and less for gathering data or generating ideas and hypotheses. This is consistent with self-reported AI use from an accompanying survey of scientists.

To understand these differences, we develop a measure of theoretical AI capability. Claude use is higher in tasks and fields that score higher on this measure. These also tend to be more cognitive in nature and better suited to reinforcement learning, in which AI learns from feedback on repeated attempts. In short, scientists appear to use AI where it works best.

Based on these findings, we outline scenarios for how AI could accelerate scientific progress. With current capabilities, research output in some fields could as much as double, primarily from task-level augmentation. Researchers in our survey expect much greater capabilities to come. In a hypothetical scenario where AI fully automates all non-physical work, highly cognitive fields such as computer science or statistics could see up to a 15-fold increase in output. Because the fields with the largest potential for acceleration include those that produce AI itself, these gains could create an additional feedback loop, in which faster AI research produces more capable AI, which in turn could further accelerate research.

While faster scientific discovery could bring substantial benefits to humanity, AI-driven research also carries risks and uncertainties. Less human oversight could lead to the pursuit of dangerous or unethical research. AI could also reshape researchers' careers, making research work more or less rewarding, and alter the sense of purpose associated with the human pursuit of knowledge.

How we conducted this research

Our goal with this research is to measure the different ways scientists use Claude while protecting the details of their work. To do this, we use privacy-preserving infrastructure to analyze the scientific use of Claude on consumer plans. A language model, used here as a "classifier," reads each session transcript and labels it with a fixed set of categories, such as the field and the research task it involves, according to a predefined taxonomy (SciNet taxonomy by Alamos and Althoff 2026). The resulting labels, such as "chemistry" or "data cleaning," are the only information from these conversations retained and analyzed by the research team for this report. No classifier employed in this study extracts and stores open-ended information or details on research content, such as research datasets, hypotheses, experimental methods, early findings, or unpublished papers. Results are reported only in

aggregate and only for cells that meet a minimum number of distinct accounts. All classifiers introduced in this study are listed in the Appendix.

Our measurement involves two steps. A first classifier identifies whether a session transcript involves scientific work. A second classifier then assigns each of these transcripts to a field, subfield, and task in the SciNet taxonomy. SciNet contains 7,262 tasks, spanning 34 fields and 318 subfields. Some tasks in the taxonomy, such as drafting manuscripts, are universal to all fields. Others are specific to individual fields or subfields. For instance, in the taxonomy, the task of predicting optical absorption and emission spectra is specific to materials science. Together, the tasks cover the full process of producing a new research finding.

Our main sample consists of 207,049 sessions on consumer plans from August 2 to September 30, 2026, which were assigned a scientific field and a research task: 128,914 from Chat and Cowork, two general-purpose interfaces for conversation and knowledge work, hereafter referred to as "Claude," and 78,135 from Claude Code, an agentic coding tool that works directly with a user's files, writing and running code. We also study a supplemental sample of Claude Science sessions. For details on sample construction, see the Appendix.

Combining usage data and a structured taxonomy of research tasks allows us to study which fields adopt Claude and for which parts of the research process. This research contributes to a growing body of work attempting to measure the adoption of AI technology for scientific discovery (Kobak et al. 2025; Liang et al. 2025; Hao et al. 2026; Codreanu et al. 2026).

How Claude use varies across scientific fields

We begin by describing how Claude use varies across scientific fields. Figure 1 compares a field's share of sessions against the share of global researchers in that field, giving us a sense of relative use per scientific field.[1] On Claude, the three largest fields by session share are computer science (11.9%), engineering (8.4%), and medicine (7.7%), which are also among the largest fields by global researcher counts.[2] Overall, fields with a larger share of researchers also tend to account for a larger share of Claude sessions (log-log correlation of 0.79).

Claude use across scientific fields

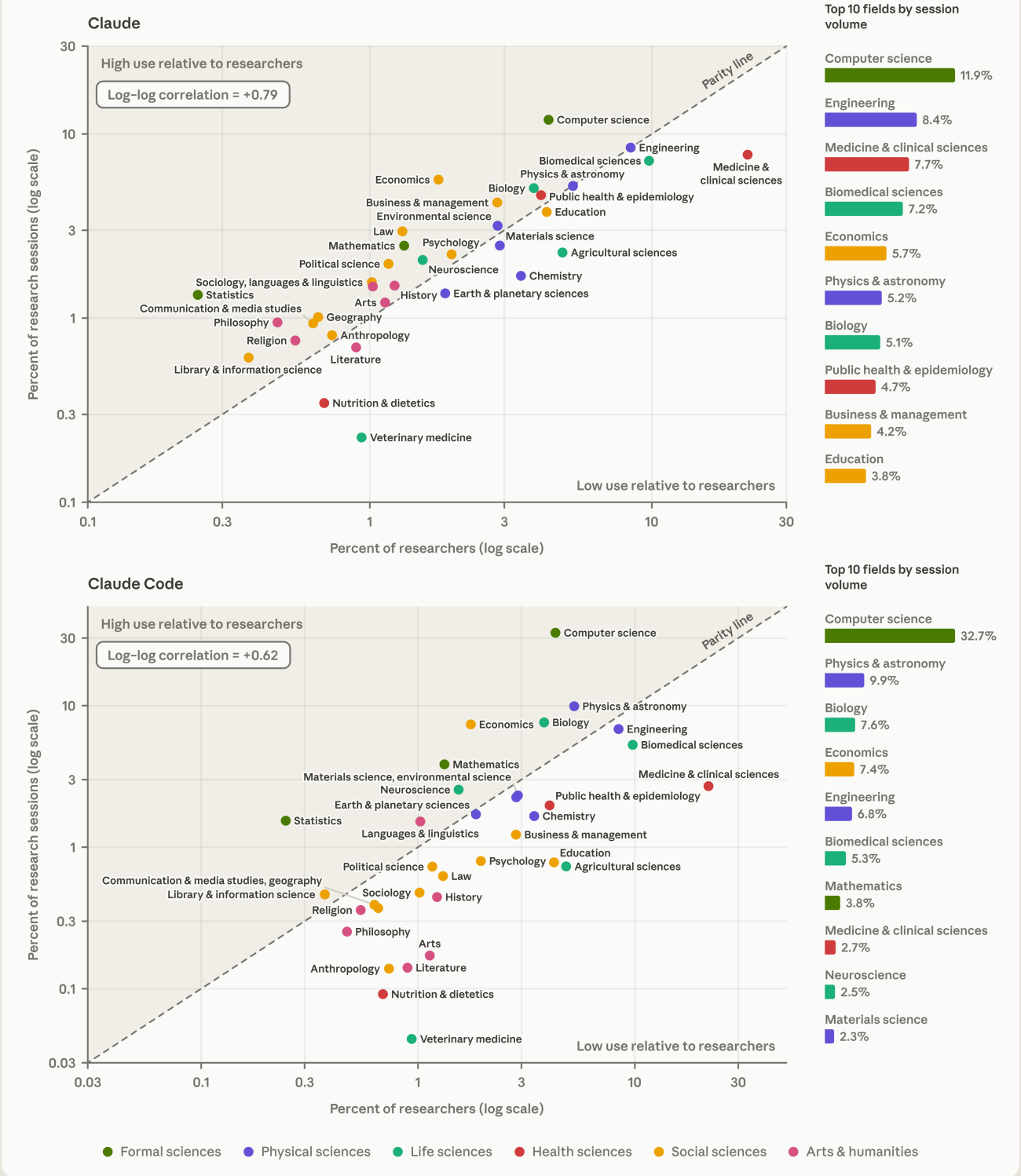


Figure 1. Claude use across scientific fields. Each dot represents a field. The horizontal axis shows the field's share of researchers, and the vertical axis shows the field's share of research sessions, both on log scales. The top panel shows sessions from Claude; the bottom panel shows sessions from Claude Code.

However, Claude use per field cannot be explained by the number of researchers alone. In Figure 1, fields above the 45-degree line show higher usage relative to their size, while fields below the line show lower usage. Relative usage is especially high for the formal sciences, including computer science, mathematics, and statistics. Usage is also high for the social sciences. By contrast, physical sciences such as chemistry, where work often happens in a lab, tend to show lower relative usage. So does medicine, the third-largest field by session volume, where researchers spend much of their time with patients.

On Claude Code, usage is less strongly related to field size (log-log correlation with researcher counts of 0.62) and concentrates in more quantitative fields. Despite making up only 4.3% of the world's researchers, computer science alone accounts for 32.7% of Claude Code sessions.[3] Other formal sciences, such as mathematics and statistics, also show high relative usage. Relative usage is low in the arts and humanities and in the social sciences, with the exception of economics, a comparatively quantitative field.

We observe similar patterns when comparing usage across 318 subfields within those 34 fields (Figure 2).[4] This more granular subfield view reveals that artificial intelligence and machine learning stands out, with session shares of 5.0% for Claude and 15.2% for Claude Code, despite making up only 1.2% of the world's researchers. More generally, regardless of the broad scientific domain, computational subfields are among the strongest adopters. These subfields include computational biology, computational linguistics, and computational neuroscience. Robotics also shows high relative usage, likely because it builds heavily on computational methods despite its focus on physical systems.

Claude use across scientific subfields



Figure 2. Claude use across scientific subfields. Each dot represents a subfield. The horizontal axis shows the subfield's share of researchers, and the vertical axis shows the subfield's share of research sessions, both on log scales. The top panel shows sessions from Claude; the bottom panel shows sessions from Claude Code. Subfields with too few accounts are not shown for privacy reasons.

How Claude use varies across research tasks

Beyond understanding which fields use Claude, we also want to understand what scientists use it for. Research involves many different tasks, from generating hypotheses and analyzing data to writing up findings in a research paper. To better discern how Claude supports scientific discovery, we study how usage varies across these tasks. To do so, we assign each session to one of the 7,262 tasks in SciNet. These tasks can be universal to all fields, such as drafting manuscripts, or specific to fields or subfields, such as performing differential gene expression analysis to compare tumor subtypes, which is associated with cancer biology. The taxonomy distinguishes between research tasks, which researchers perform to produce new scientific findings, and non-research tasks, such as teaching, peer review, and service to the department or field.[5] We limit our analysis to research tasks spanning eight categories: (1) data analysis, (2) writing and communication, (3) data gathering, (4) reading and knowledge acquisition, (5) theoretical analysis, (6) design and development (of software, materials, and creative works), (7) ideation and hypothesis generation, and (8) other, such as administering research.

For each task category, we calculate its share of Claude and Claude Code sessions separately. To assess whether usage in a category is high or low, we consider the share of their time that researchers typically spend on that category, as estimated by SciNet. These time estimates are based on an LLM classifier, validated using both wet-lab research protocols that record researcher time spent per task and the frequency of tasks across thousands of academic papers (see the Appendix). This provides a benchmark to assess whether the share of Claude use across different task categories follows or deviates from how researchers spend their time.

Usage by task category. Data analysis tasks have the highest usage in absolute terms and are overrepresented relative to researcher time. Data analysis accounts for 33.7% of sessions on Claude and 48.1% on Claude Code, compared with 27.6% of researcher time. Writing and communication tasks also show high usage, especially on Claude (32.0%, compared with 16.0% of researcher time).[6] Data gathering tasks have low relative usage, accounting for 28.6% of researcher time but only 11.5% of sessions on Claude and 11.9% on Claude

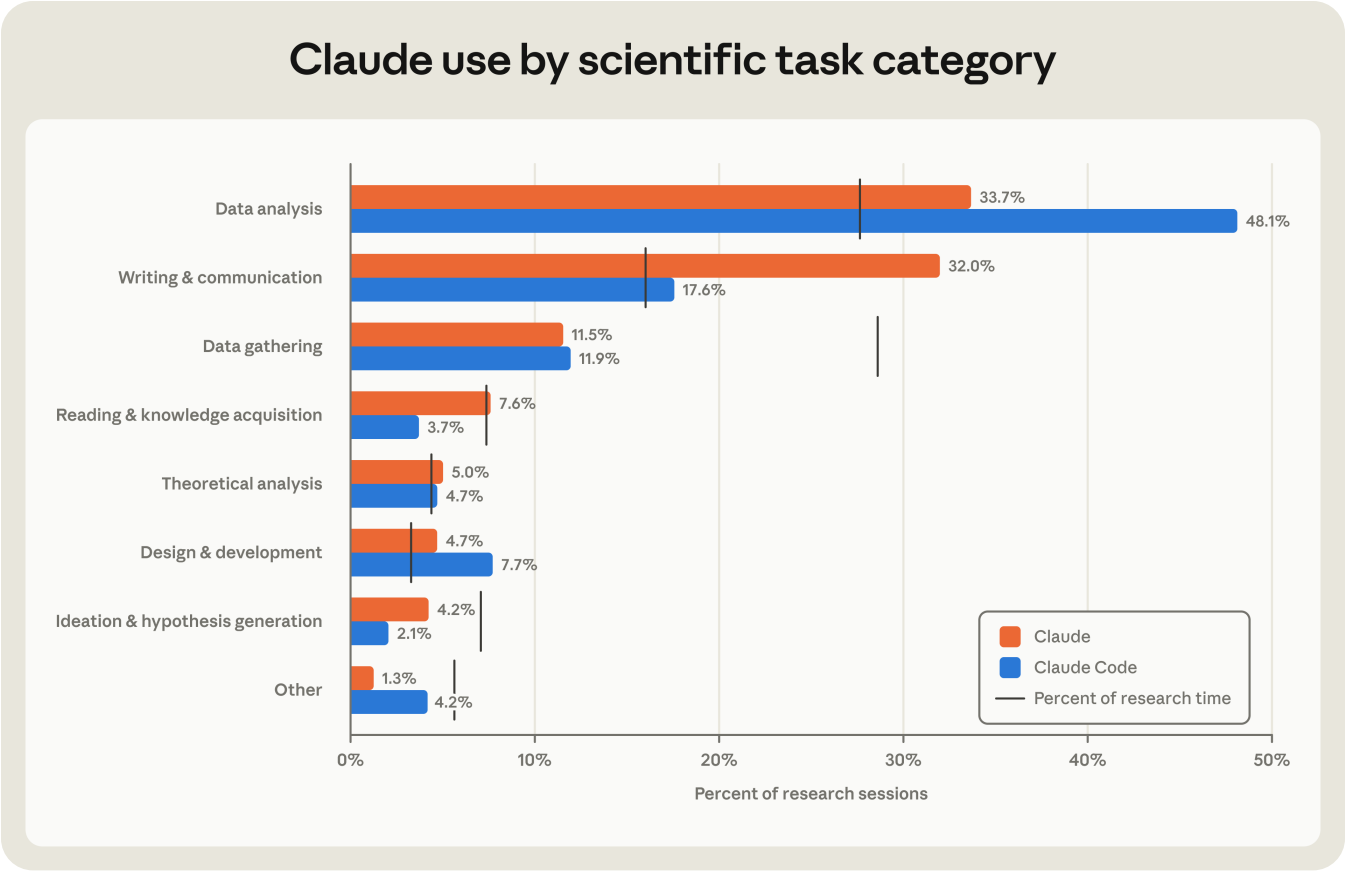


Figure 3. Claude use by scientific task category. Each row represents one of the eight task categories. Orange bars represent the share of Claude research sessions dedicated to each category; blue bars represent the same for Claude Code research sessions. The black lines represent the estimated share of research time calculated from SciNet (see the Appendix).

Code. The underrepresentation of data gathering is particularly strong for tasks with physical barriers, such as the collection of clinical or biochemical data. Data gathering tasks that can rely on more digital processes, such as retrieving archival documents or text, are less underrepresented, while purely digital data collection tasks, such as collecting online data, can be overrepresented. Ideation and hypothesis generation tasks also have low usage, accounting for 4.2% of sessions on Claude and 2.1% on Claude Code, compared with 7.1% of researcher time.

Comparing observed usage to survey evidence. Along with our analysis of Claude and Claude Code sessions, we conducted a survey of 528 scientists, whose self-reported AI usage lines up with the patterns we observed in those sessions. The sample consists of researchers primarily in the life and health sciences who are active adopters of AI.[7] Consistent with the Claude usage patterns shown in Figure 3, survey participants' self-reported AI usage is highest for data analysis and writing and communication, and lower for

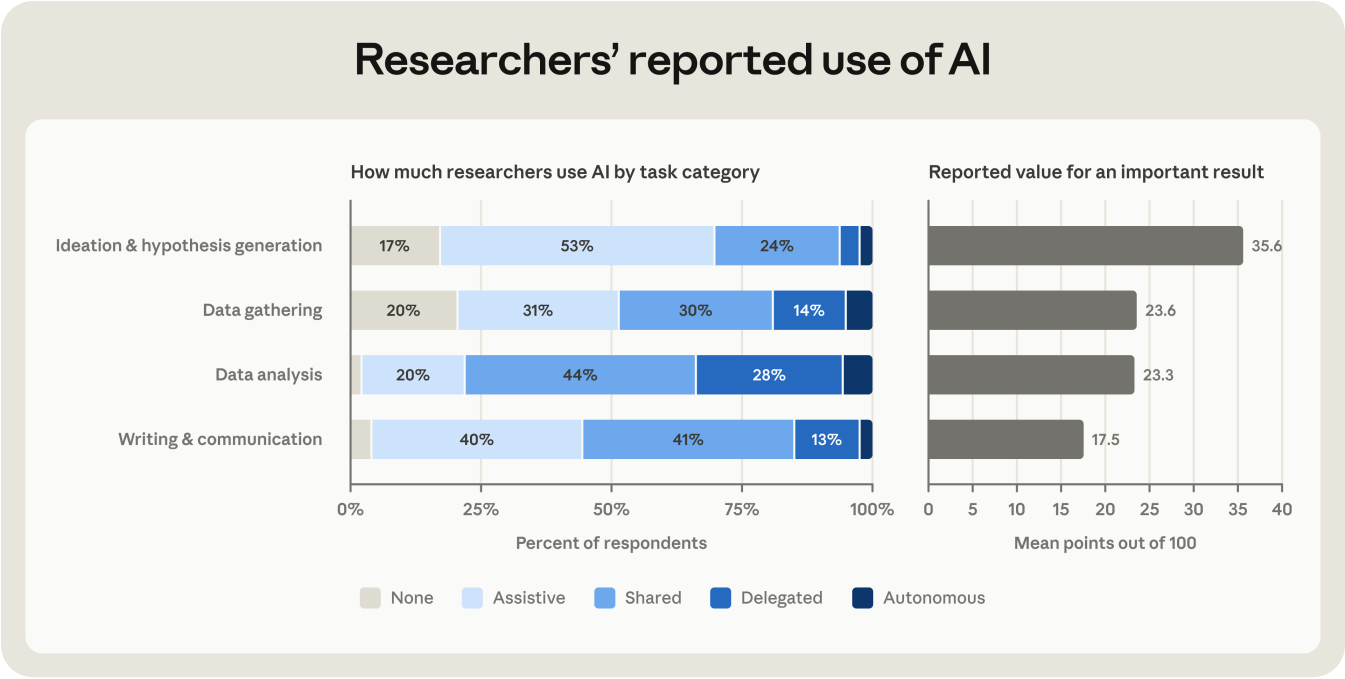


Figure 4. Researchers' reported use of AI. The left panel shows the percentage of survey respondents at each level of AI use, from none to autonomous, for four research task categories. The right panel shows how the respondents rate the importance of each category for producing an important research result (allocating 100 points across the four categories).

ideation and hypothesis generation and data gathering. Participants also report more shared or delegated use of AI for data analysis and writing and communication (Figure 4).

The survey also asked which categories matter most for producing important research results. Respondents rank ideation and hypothesis generation first and data gathering second. Notably, these categories are underrepresented in self-reported usage as well as in Claude sessions. The reasons for this underrepresentation likely differ. Much of data gathering is physical, while ideation is not. For ideation, scientists may prefer generating ideas themselves, or they may perceive AI's capabilities as limited. Our data may also undercount ideation, since it tends to be woven into many other research tasks and is difficult to classify. But AI may also be genuinely weaker at high-level ideation, such as choosing which research question to pursue and designing a study to answer it.

Biological and biomedical use of Claude Science. While Claude and Claude Code are general-purpose interfaces, specialized AI tools for the sciences are increasingly available. To study whether such tools support more specialized and demanding research, we use our privacy-preserving analysis

infrastructure to compare biology and biomedical work on Claude Science with that on Claude and Claude Code.

Claude Science connects Claude to many research tools and databases that are otherwise fragmented and acts to integrate and accelerate scientific workflows. Scientists must often use different platforms, each with unique data structures or query languages, to combine distinct tools and databases in their research. For example, biologists studying whether existing drugs can act on a specific protein must typically consult separate databases regarding the protein's function, its structure, compounds already tested against it, and related drug trials or papers. Claude Science combines these into one workflow.

Use of Claude Science is concentrated in the life sciences. The top two fields are biomedical sciences and biology, which together comprise over 50% of sessions. Within these fields, we assess which tasks are most overrepresented relative to tasks performed by Claude.

We find that researchers use Claude Science for more specialized tasks than Claude (Figure 5). For example, the task "Propose mechanistic models for protein–protein or protein–ligand interactions" is 5.6x more likely to appear on Claude Science than on Claude. This task involves explaining, at the level of individual molecular contacts, how a protein binds to another protein or to a small molecule such as a drug. Claude Science is well positioned for this task, as it can gather protein structures and binding data from multiple connected databases and combine them into a testable model of the interaction. By contrast, Claude Science is used relatively less for general tasks such as drafting manuscripts or coordinating with collaborators. Moreover, we find that ideation and hypothesis generation accounts for 8% of Claude Science sessions, roughly on par with the share of researcher time these tasks take, suggesting that specialized AI tools for science can facilitate more complex and creative work.

Focusing on biomedical sciences and biology, and using measures from our economic primitives framework introduced in prior work (Appel et al. 2026), we find that tasks on Claude Science are longer, more sophisticated, and performed more autonomously than those on Claude and Claude Code (Figure 6). This is consistent with specialized AI tools being used for more demanding tasks.[8] Sessions on Claude Science contain work that would take nearly twice as many human hours to complete without AI as work on Claude and Claude Code. The content of these sessions is also more advanced. Over 80% of Claude Science

use is estimated to require doctoral-level training (20 years of education), compared with about 50–70% on other interfaces. Over 70% of Claude Science sessions are performed with a high degree of AI autonomy, far higher than on Claude (23.7%) and somewhat higher than on Claude Code (64.1%).

Claude Science use across fields and tasks

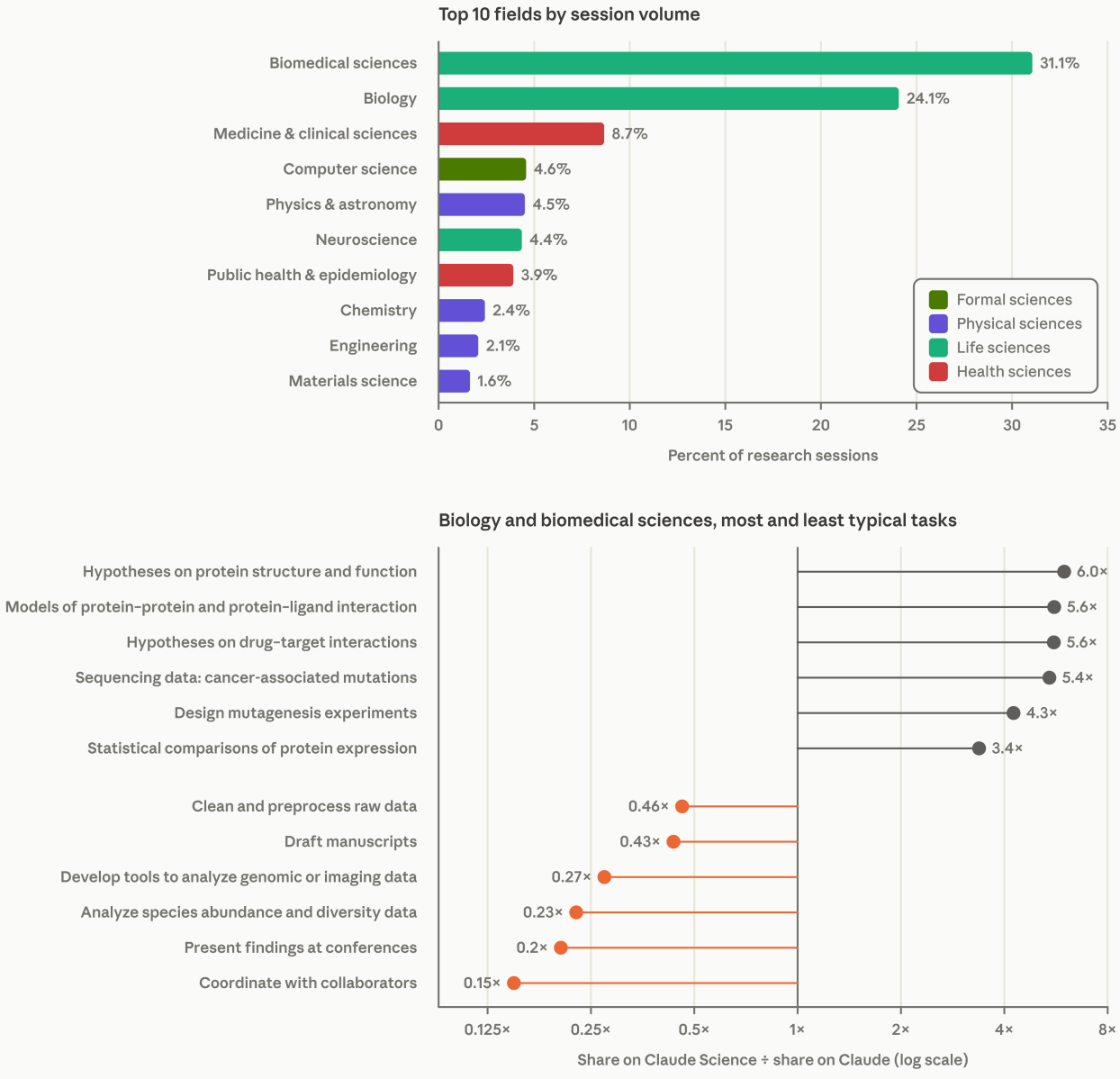


Figure 5. Claude Science use across fields and tasks. The top panel shows the top 10 fields by their share of all Claude Science research sessions. The bottom panel shows the most and least typical tasks on Claude Science relative to Claude within biology and biomedical sciences specifically, calculated as follows: (task's share of Claude Science research sessions in these fields) / (its share of Claude research sessions in these fields).

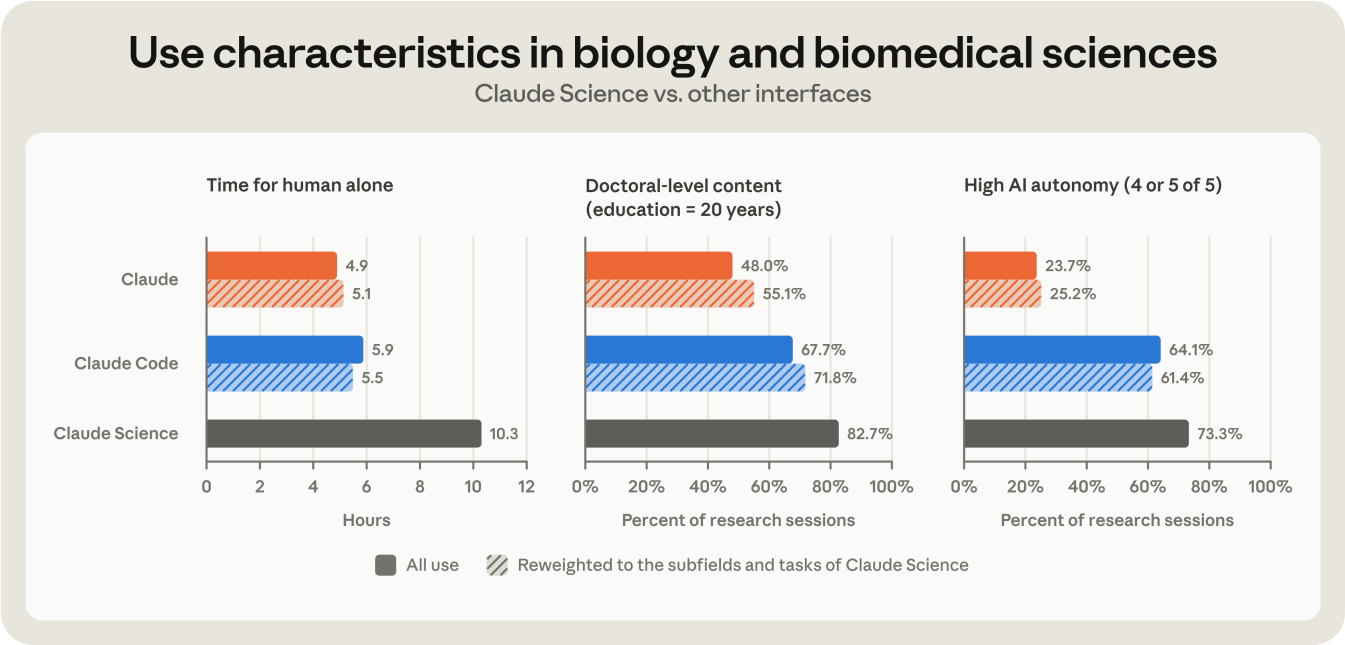


Figure 6. Use characteristics in biology and biomedical sciences. Comparing human task time, doctoral-level content, and level of autonomy for tasks completed on Claude, Claude Code, and Claude Science in biology and biomedical sciences. The hatched orange and blue bars reweight the subfields and tasks performed by Claude and Claude Code to match those performed by Claude Science.

Outputs of Claude use. Finally, the artifacts produced with Claude show clear differences across fields (Figure 7).[9] Written artifacts such as documents, reports, or academic papers are more common in the social sciences and humanities, relative to the natural and formal sciences. Conversely, technical artifacts such as code, software outputs, and data are more frequent in the natural and formal sciences. The health sciences overall are more similar to the social sciences in terms of artifacts produced. Technical artifacts are about twice as common on Claude Code as on Claude, but the broad pattern across fields is similar (see the Appendix).

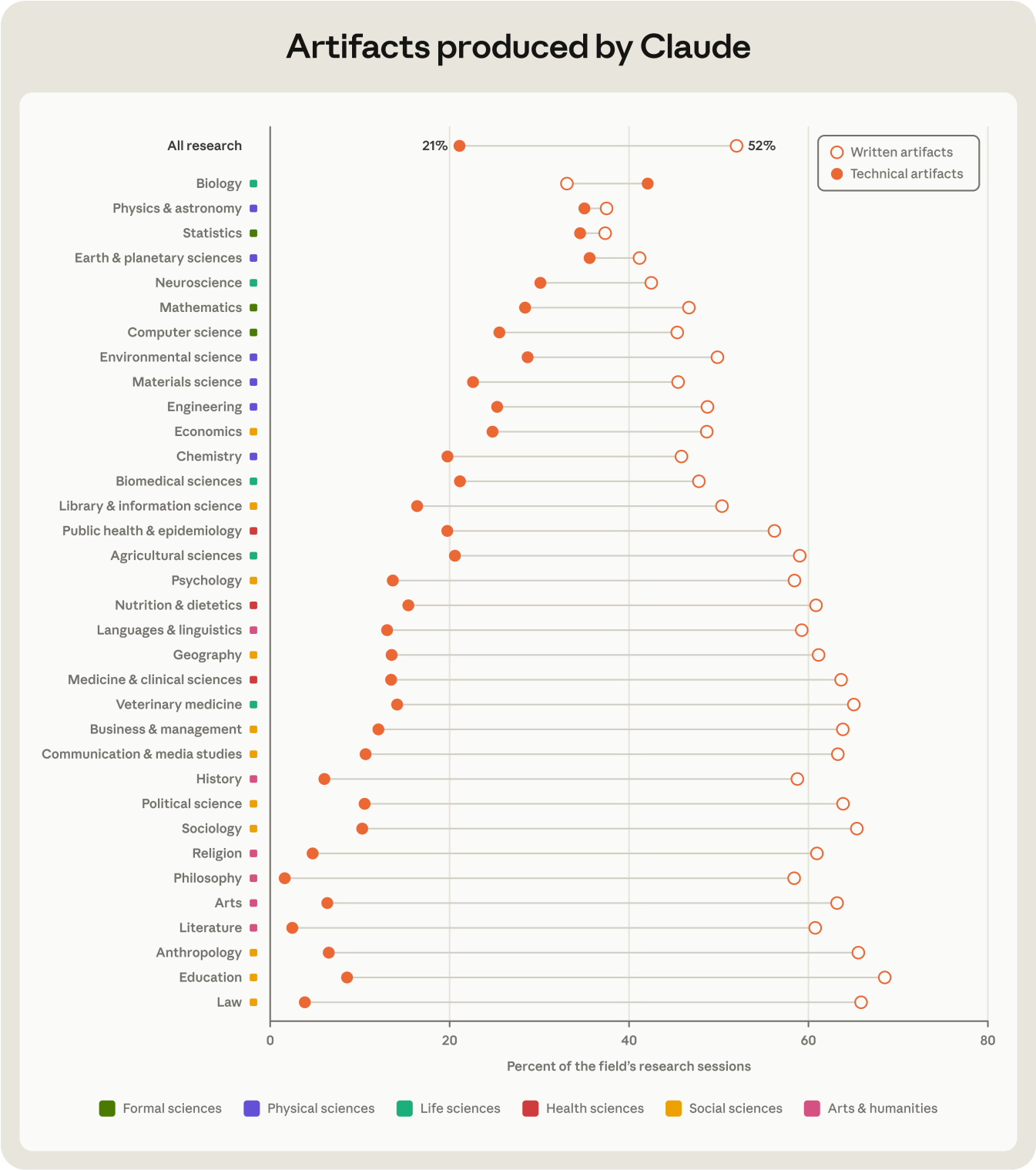


Figure 7. Artifacts produced by Claude. The share of sessions on Claude that produce written artifacts and technical artifacts by field. Fields are ordered by the difference between the share of written artifacts and technical artifacts.

How use relates to capabilities

Does AI usage track AI capabilities? Field size explains some of the variation in Claude use, but not all of it. For example, researchers in computer science use Claude far more than the field's size predicts, while those in nutrition use Claude far less. Similarly, where researchers spend their own time does not fully predict where they use Claude the most. Some broad task categories, such as data analysis and writing, have a much larger share of usage than of research time, while others, such as data gathering and ideation and hypothesis generation, have a much smaller share. Here, we test whether these differences in usage reflect what AI can actually do.

To measure task-specific AI capabilities, we first construct capability scores using an LLM classifier. For each task, SciNet provides a list of individual steps that a scientist needs to perform to complete the task. For instance, in chemistry, one task involves purifying reaction mixtures, and one of that task's steps involves packing a chromatography column. The classifier is shown pairs of steps across the taxonomy and asked which one AI can perform better. We aggregate these comparisons using a Bradley–Terry model—the same model behind the Elo ratings that rank chess players based on head-to-head results. A step's score rises when it wins comparisons and falls when it loses them. We then average each task's steps to obtain a task-level score (see the Appendix for details). We label this measure *theoretical AI capability*. We refer to a task's specific score as its *AI capability score*.

To compare usage across tasks, we ask whether Claude use is higher or lower for a task, relative to the time researchers spend on that task. For example, a task that takes 5% of researchers' time but accounts for 10% of Claude sessions would represent twice as much use as expected, a ratio of 2. A ratio of 1 indicates that Claude use is in line with researcher time spent on the task. A ratio below 1 indicates that Claude is used less than expected based on the amount of researcher time the task takes.

Figure 8 shows that this use ratio rises sharply with a task's AI capability score. For tasks in the lowest 20% of capability scores, like purifying reaction mixtures and conducting participant observation, Claude is barely used. At median capability scores, the ratio is roughly 1, so usage is in line with researcher time. For tasks in the top 20% of capability scores, like visualizing data and applying

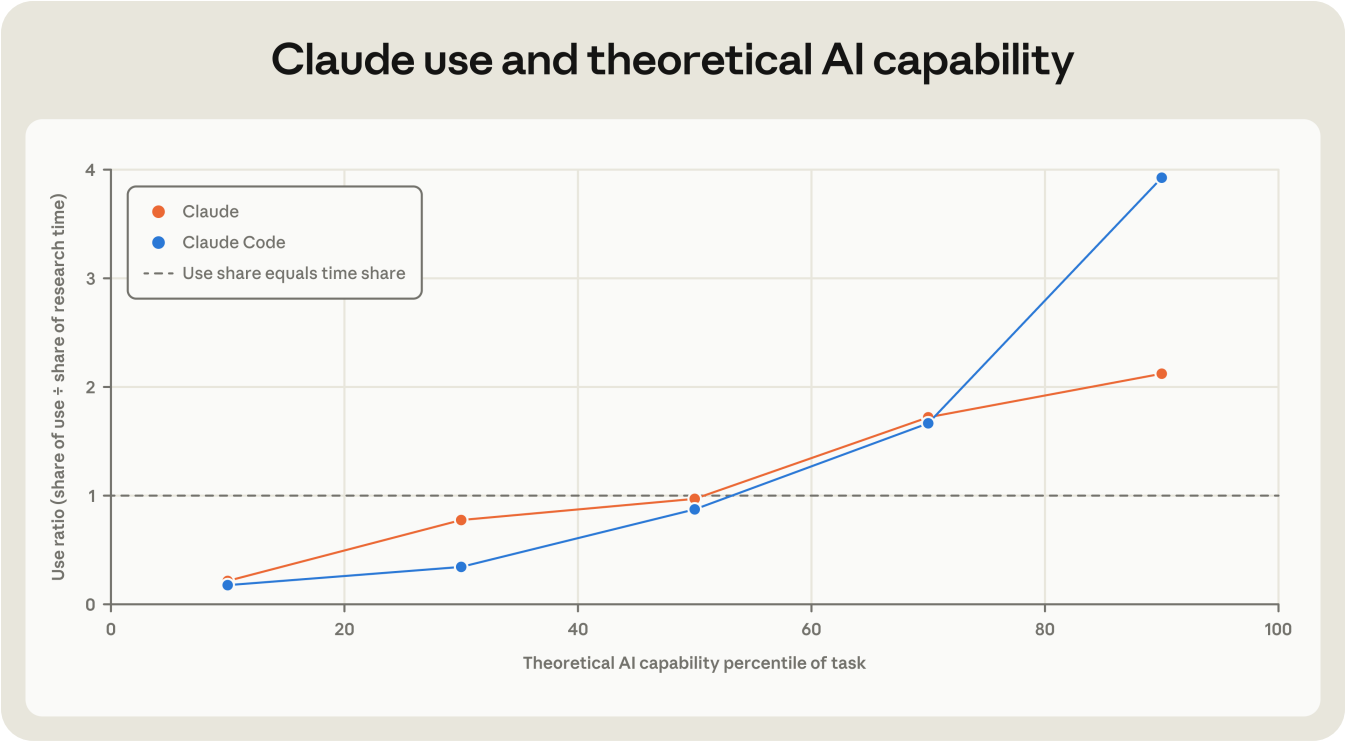


Figure 8. Claude use and theoretical AI capability. The vertical axis represents the use ratio for the tasks in each capability quintile, calculated as the share of research sessions devoted to those tasks divided by their share of research time estimated from SciNet. The horizontal axis represents the task's AI capability score (percentile). The orange dots represent the average use ratio for tasks in each AI capability quintile for Claude; the blue dots represent the same for Claude Code.

statistical tests, usage is about twice as high on Claude and about four times as high on Claude Code, compared with the typical time researchers spend on those tasks. These patterns suggest that scientists' use of Claude closely tracks AI capability scores. The steeper gradient for Claude Code reflects its focus on computational work, which has higher capability scores.

Theoretical AI capabilities can differ across tasks for many reasons. To understand these differences, we relate each task's AI capability score to three groups of LLM-derived task characteristics: its suitability for reinforcement learning (RL); whether the work is predominantly cognitive, interpersonal, or physical; and whether regulatory restrictions limit AI use (Figure 9; see the Appendix for definitions and data sources).

Consider RL suitability. In RL, an AI model tries a task repeatedly and is rewarded when it succeeds, allowing the model to learn how to perform that task. Tasks are particularly suitable for RL when success is easy to verify, when many examples can be simulated in a computing environment, and when feedback can be delivered quickly (Lambert et al. 2024; Guo et al. 2025). We use task-specific

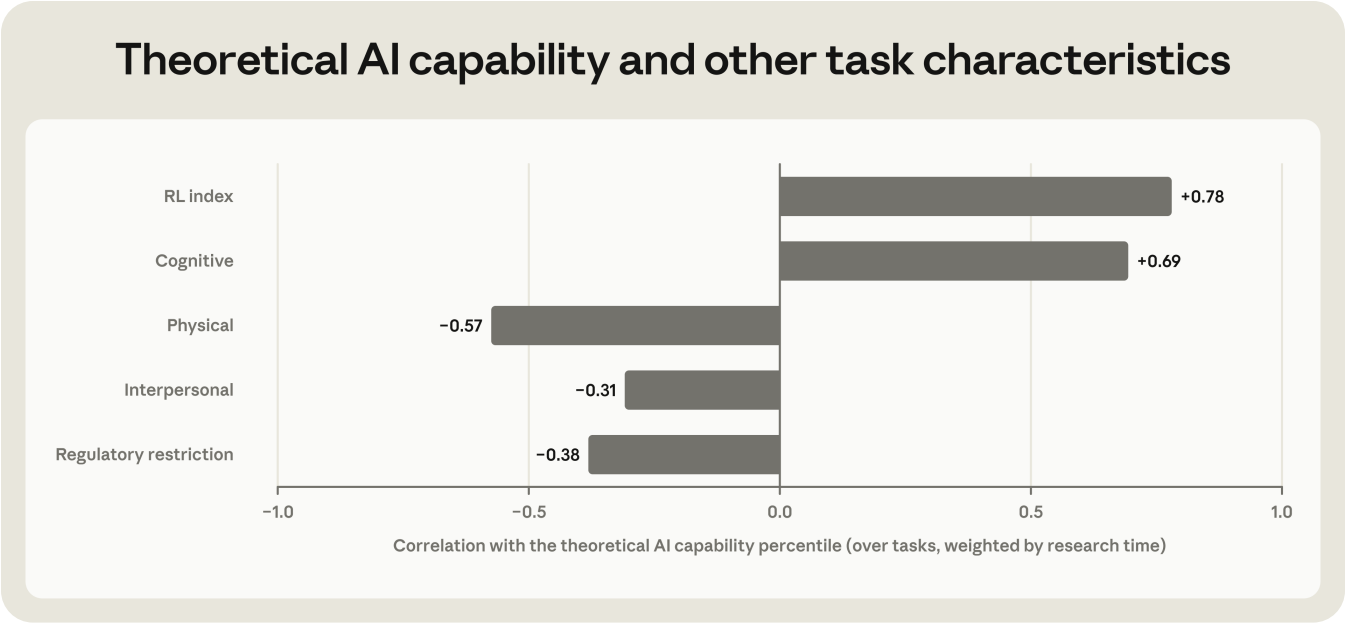


Figure 9. Theoretical AI capability and other task characteristics. Correlations between a research task's AI capability score percentile and other characteristics of the task.

measures from SciNet for these three RL-related components and summarize them in a single index. A task's capability score is highly correlated with its RL index score, which was classified separately (correlation of 0.78). This suggests that AI is most capable at the research tasks where RL is easiest to apply.

Other important task characteristics concern the nature of the work. Some research tasks are fundamentally cognitive, while others require physical actions or interpersonal contact. Physical tasks in general are beyond the reach of purely cognitive machines. There can also be regulatory restrictions that limit AI use. AI capability scores are higher for tasks characterized by cognitive work (correlation of 0.69). By contrast, research tasks that involve more physical work have lower AI capability scores (−0.57). Scores also tend to be lower for research activities that include interpersonal work (−0.31) or that involve regulatory restrictions on AI use (−0.38).

The same task characteristics also predict how much researchers use Claude for a given task (Figure 10). Claude is barely used for tasks in the bottom fifth of RL suitability. For tasks in the top fifth, Claude use is disproportionately high relative to the time researchers spend on them: twice as high on Claude, and four times as high on Claude Code. This is similar to the association between Claude use and AI capability scores overall, where the top 20% of tasks also draw two to four times the expected use.

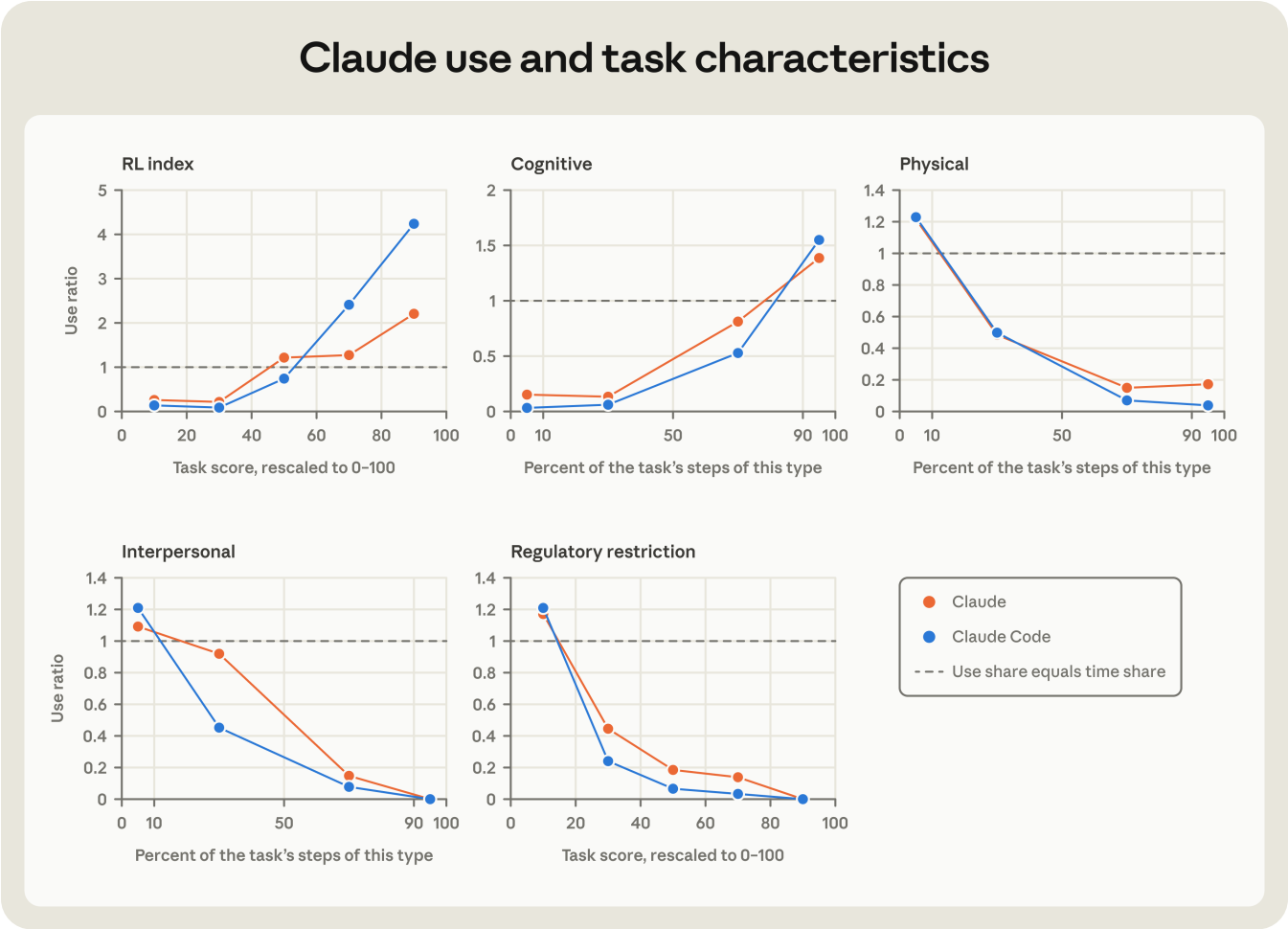


Figure 10. Claude use and task characteristics. Each panel considers a different task characteristic. The vertical axes represent the use ratio for groups of tasks ("bins") with different scores on that characteristic. The horizontal axes represent a measure of the characteristic's intensity. Orange dots represent the average use ratio for tasks in each bin on Claude; blue dots represent the same for Claude Code. Dots at zero represent bins hidden for privacy because too few accounts are responsible for that use.

Claude use is also disproportionately high in predominantly cognitive tasks. By contrast, Claude use is low for physical and interpersonal tasks. Claude use is also lower for tasks where regulatory barriers limit AI assistance. Overall, across the spectrum of highly specialized research tasks, usage and capability measures are strongly related.

Different scientific fields rely on different mixes of research activities. Does the correlation between usage and capability scores for narrow research tasks help explain why some fields use Claude more than others? Our data shows it does: a subfield's Claude use per researcher tends to rise with the average AI capability score across its tasks (correlations of 0.45 for Claude and 0.57 for Claude Code), as shown in Figure 11.

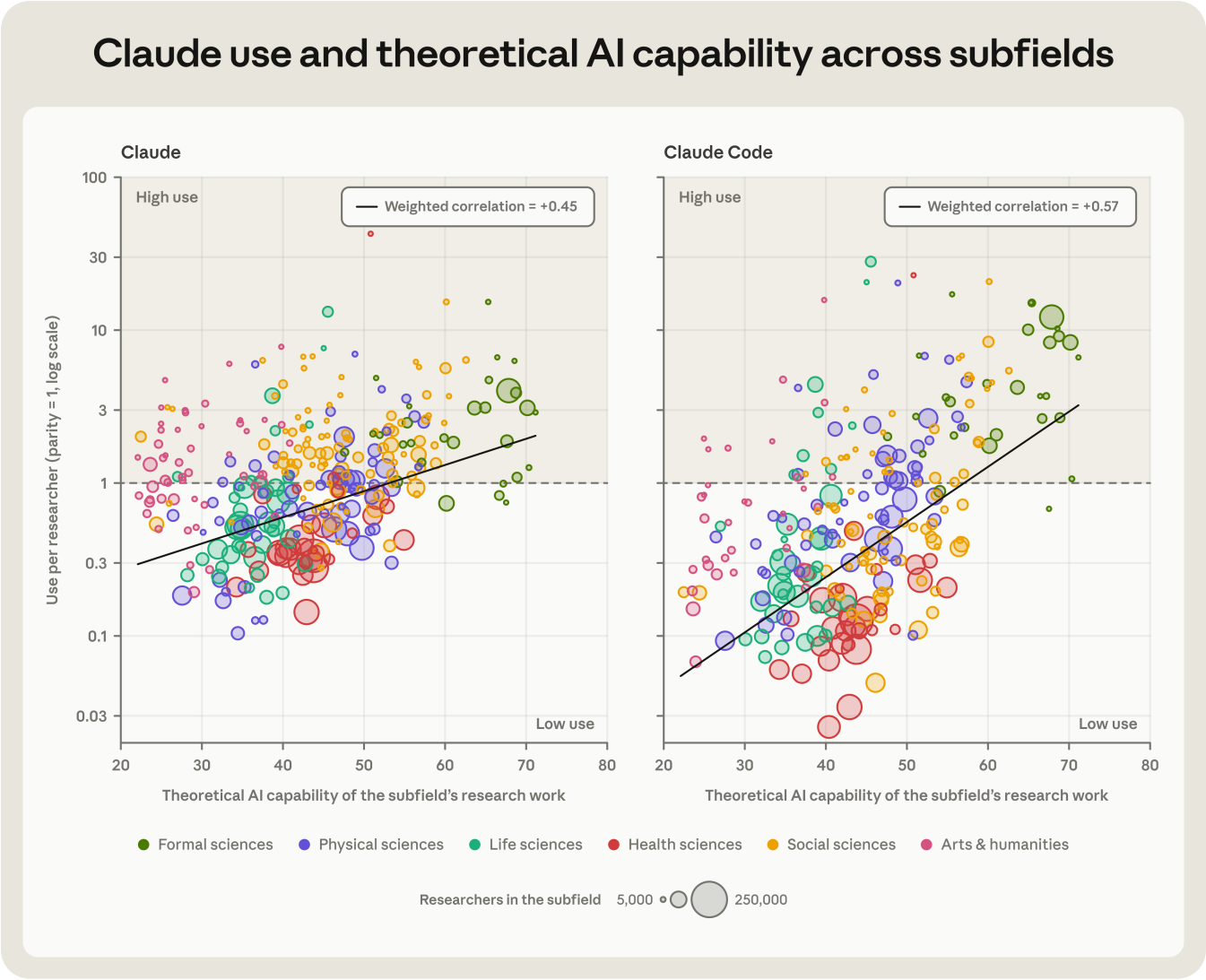


Figure 11. Claude use and theoretical AI capability across subfields. Both panels show each subfield as a dot, sized by the number of researchers in the subfield. The vertical axis shows usage per researcher on a log scale. Usage per researcher is computed as the subfield's share of research sessions divided by the subfield's share of total researchers. The horizontal axis shows the subfield's AI capability score, computed by aggregating across all of the subfield's research tasks weighted by researcher hours. The black line shows the best fit, and each panel reports the correlation weighted by subfield size. The left panel shows the results for research sessions on Claude; the right panel shows the results for research sessions on Claude Code.

There is also substantial variation across domains, which suggests that Claude use is affected by factors beyond what AI can do. Health sciences researchers use Claude relatively little compared with the AI capability scores of their tasks, possibly because regulation of medical data and research limits the use of AI or makes it harder to integrate. Researchers in the formal sciences, social sciences, and the arts and humanities use Claude more than the AI capability scores would predict. Much of the work in these fields is done through text and code, which likely facilitates the adoption of AI tools. Claude use in the physical

sciences and life sciences generally tracks AI capability scores more closely, though individual subfields can deviate.

Theoretical AI capability also predicts how much researchers allow Claude to work autonomously. Figure 12 shows that autonomous use correlates with the theoretical AI capability of the subfield. In other words, not only do subfields with more theoretical AI capability use AI more, but they also use it in a more autonomous manner.

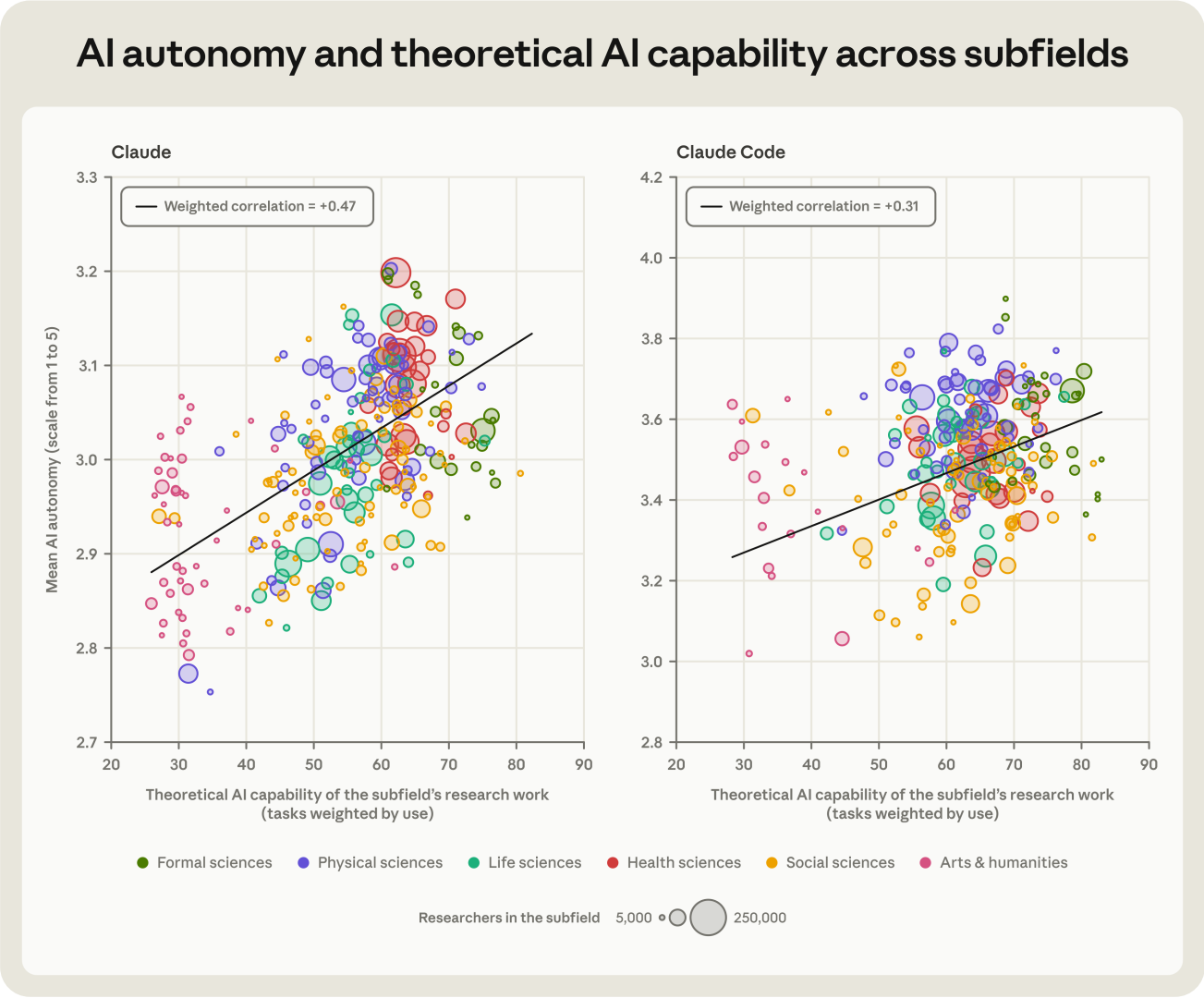


Figure 12. AI autonomy and theoretical AI capability across subfields. Both panels show each subfield as a dot, sized by the number of researchers in the subfield. The left panel represents research sessions on Claude; the right panel represents research sessions on Claude Code. The vertical axis shows the mean autonomy score for sessions in the subfield, weighted by session volume. The horizontal axis shows the mean theoretical capability of the subfield's tasks, also weighted by usage. Tasks that are not covered by any session are excluded from the analysis, since they cannot be scored on autonomy. For legibility, outlier subfields have been dropped, but they are included in the line of best fit and the correlation calculation. Correlations are weighted by subfield size.

Research acceleration scenarios with AI

The pace of scientific and technological progress depends on good ideas and the means to pursue them. Historically, both have been limited by human capacity. There are only so many researchers, and each has only so much time to formulate a hypothesis, design the right tests, run the analyses, and synthesize the findings.

AI could ease these constraints in two ways. First, it could take on work that researchers now do, allowing them to pursue more ideas. Second, it could contribute ideas that researchers might not otherwise identify. The second may ultimately matter most, but it is much harder to measure, so our analysis focuses on the first.

Approach

We consider three scenarios of scientific acceleration through AI. Scenario 1 assumes wide adoption of current AI capabilities, as estimated from our survey and the AI capability scores. Scenarios 2 and 3 are hypothetical futures in which AI fully automates 75% or 100% of non-physical research steps—the individual actions that make up each research task in SciNet—starting with the steps that have the highest AI capability scores. In all three scenarios, researchers continue to perform physical steps. These are illustrative scenarios, not forecasts. They show how physical work, theoretical AI capability, and the cost of AI could shape the acceleration of scientific research, and which fields could accelerate the most.

For each scenario, we report two measures. The first is the *AI share*: the fraction of researcher time spent on a study that AI could take on. Given how researchers currently allocate their time, it measures how much human research time AI could free up, and it follows directly from our analyses above. Researchers can redirect that freed-up time to other work, increasing research output. Our second measure, the *acceleration*, estimates this gain: the increase in research output for a fixed research budget, relative to the research output without AI. It tells us how many more research results the same budget could produce. Freed-up time does not translate one-for-one into more output, both

because researchers redirect their effort toward the tasks where AI helps less, and because AI itself has costs.

We build the AI share for Scenario 1 from our survey, which records five levels of AI use: none, assistive, shared, delegated, and autonomous. We assume these levels correspond to AI taking on 0%, 25%, 50%, 75%, and 100% of a researcher's time for a given task step. Within each task category, we rank task steps by AI capability score. We then assign a level of AI use to each task step, in the same proportions reported in our survey, with the highest levels going to the highest-scoring steps. For ideation and hypothesis generation, for example, 17% of responses report no AI use, and 53% report assistive use. The 17% of ideation task steps with the lowest AI capability scores are therefore assigned no AI share, the next 53% are assigned 25%, and so on up the scale. Because respondents are likely to be heavy AI adopters, we treat their reported use as a proxy for what AI can do today. Finally, we aggregate researcher time freed up by AI from steps to fields (see the Appendix).

To compute the acceleration, we use an ideas production function, a standard economic model of how research inputs, such as researcher time, are turned into research outputs. In this model, a paper requires many research tasks (Jones 2026). The tasks are complementary, such that strength in one cannot fully make up for weakness in another. As AI makes some tasks cheaper, researchers shift their effort toward tasks where AI helps less, thereby easing the bottlenecks that limit the overall gain. For example, researchers could spend more time coming up with new questions or running more physical experiments to expand empirical evidence. We calibrate how much these bottlenecks limit the overall gain using Ahmadpoor and Jones (2019). For AI's cost, we assume that an hour of AI agent time costs $30, compared with $95 for an hour of a researcher's time, and that AI runs for 0.2 hours for each researcher hour it saves. These are illustrative assumptions: the AI cost uses today's public list prices. The smaller these costs, the more research can be done within a given budget and the larger the acceleration, as we discuss below. The Appendix describes the model and assumptions.

Scenario 1: Current capabilities

With current capabilities, if adopted widely, AI could take on 27–52% of the researcher time needed to produce a new study, with a median of 38% across

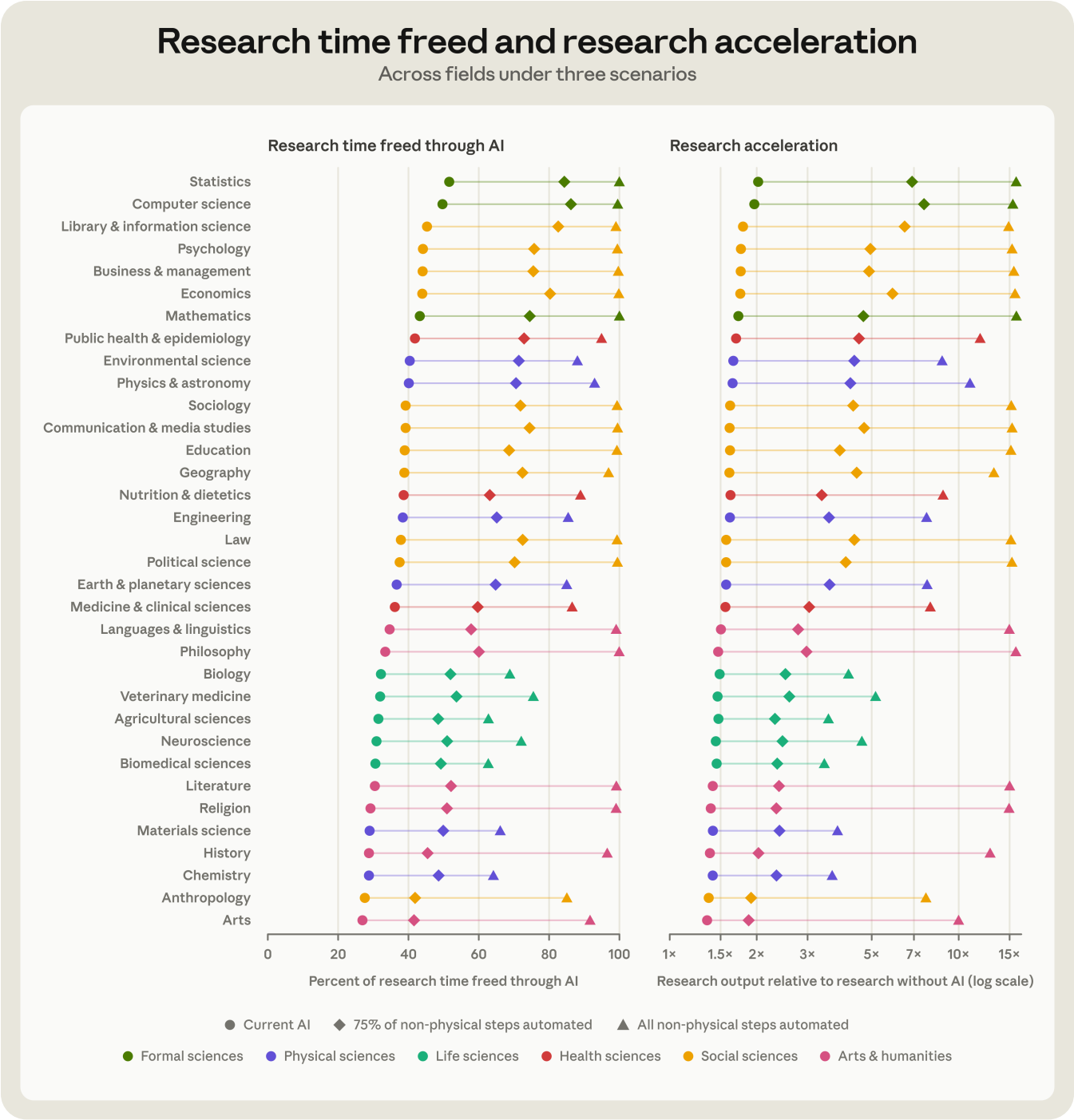


Figure 13. Research time freed up by AI and research acceleration. Every row is a field. The left panel shows fields ranked by the estimated share of researcher time that AI could take on. Circles show estimates for current capabilities; diamonds show the estimated share of time freed up if 75% of non-physical steps are automated; triangles show the estimated share of time freed up if 100% of non-physical steps are automated. The right panel shows the research acceleration given the research budget of a field. Its horizontal axis shows research output with AI as a multiple of the output without AI (1x = no change) under our three scenarios.

fields (Figure 13, left panel). These shares imply research accelerations ranging from 1.35x in the arts, 1.36x in anthropology, and 1.41x in chemistry to 1.96x in computer science and 2.02x in statistics (Figure 13, right panel). Accelerations

remain limited because AI work today is mostly assistive or shared rather than autonomous, so researchers still spend substantial time on nearly every step. Even the most AI-ready fields, whose work is almost entirely cognitive, reach only about a 2x acceleration. These estimates assume that AI diffuses widely and that researchers apply it to every step where it is useful today.

Two factors explain most of the differences across fields. The first is physical work, which AI cannot take on without physical integrations. In lab and field sciences such as the agricultural sciences, biomedical sciences, chemistry, and materials science, 38–41% of researcher time goes to physical steps, compared with 4% in the median field. The physical share alone explains about one-third of the variation in the AI share. The second factor is the distribution of AI capability scores within cognitive work. The arts and history spend little time on physical steps (10% and 4%, respectively), yet their cognitive steps have low AI capability scores, perhaps because qualitative and interpretive work is less amenable to reinforcement learning. Computer science and statistics, by contrast, have high scores and large AI shares, driven by data analysis, the type of work with the highest AI capability scores across fields.

Fields with a higher use ratio on Claude also tend to have larger AI shares (Figure 14). The AI share is calculated without any Claude use data, so this relationship is further evidence that AI use tracks theoretical AI capability.

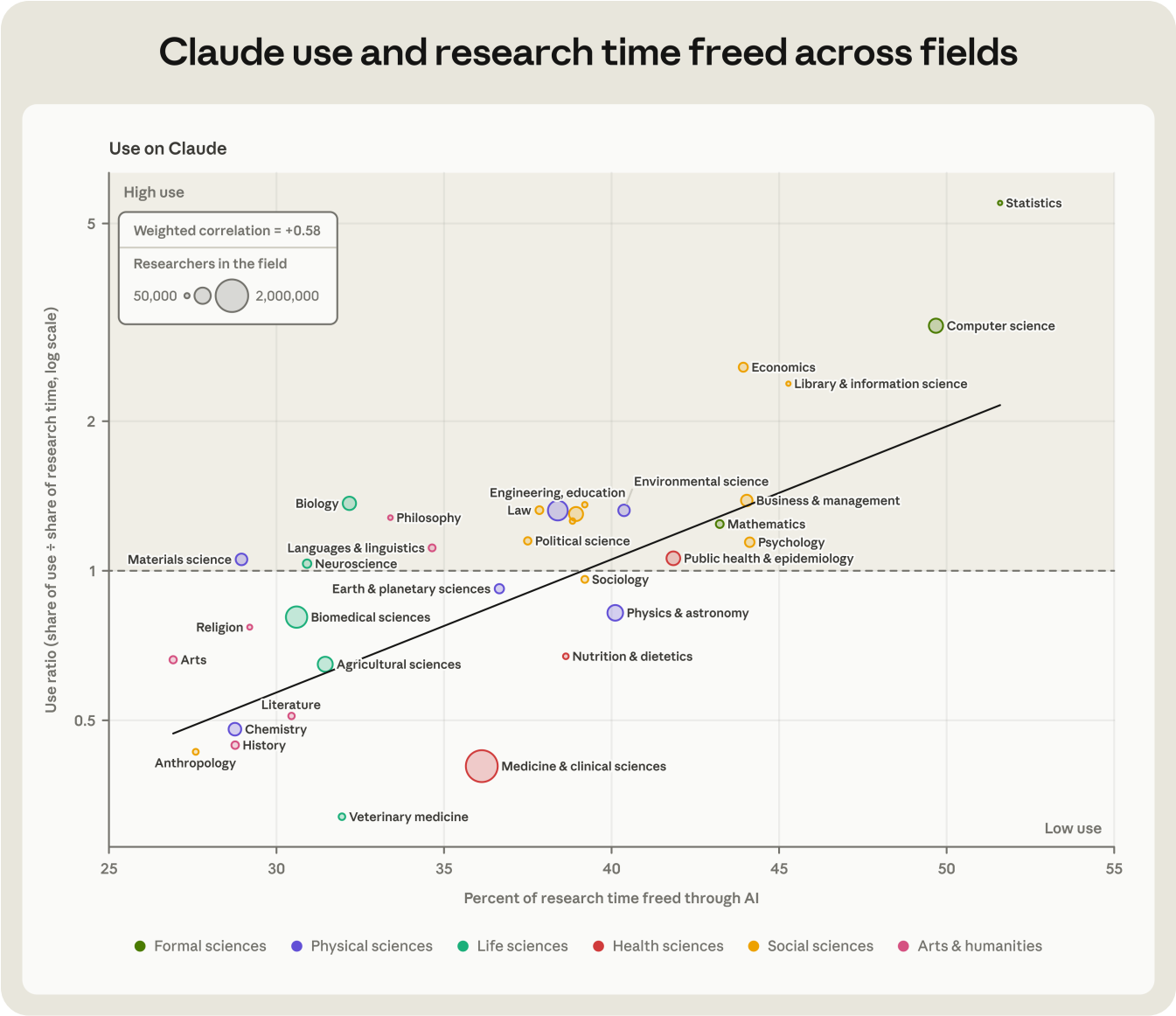


Figure 14. Claude use and research time freed up by AI across fields. The correlation between observed Claude use and the AI share for each field. Each field is a dot, sized by the number of researchers. The vertical axis shows the field's use ratio on Claude, on a log scale. The horizontal axis shows the estimated share of human research time that could be completed by AI with current capabilities.

Scenarios 2 and 3: Greater automation

In these more speculative scenarios, which depend heavily on assumptions about AI's reach and costs, greater automation raises both the AI share and the acceleration sharply, and widens the differences between fields.

In Scenario 2, where 75% of non-physical steps can be done by AI, accelerations reach about 8x in computer science, 7x in statistics and library and information science, and 6x in economics. Most other social sciences, mathematics, and physics and astronomy reach 4–5x, while bench and medical sciences reach

about 2–3x. The bench sciences and the arts and humanities gain the least, but for different reasons. In the bench sciences, much of the work is physical and must still be done by researchers. In the arts and humanities, the work is mostly cognitive but has low AI capability scores, so few of its steps are among those automated. Automated steps make up 82% of non-physical researcher time in computer science but only 28% in the arts. History, anthropology, and the arts accelerate by about 2x or less.

In Scenario 3, the arts and humanities pull ahead, while the bench sciences remain held back. Once AI automates all non-physical steps, low AI capability scores no longer limit cognitive work, and AI can take on nearly all research work in many fields (Figure 13, left panel). In this speculative scenario, fields with few physical tasks, including statistics, computer science, economics, mathematics, psychology, and most of the social sciences and the arts and humanities, reach about 15–16x. Fields with substantial lab or field work, such as biology, chemistry, materials science, agricultural sciences, and biomedical sciences, reach only 3–4x. Absent automated labs, their physical steps, representing about 40% of researcher time, remain the bottleneck.

Two lessons follow from these scenarios. First, physical bottlenecks remain powerful. Steps that AI cannot perform without physical integrations come to dominate researcher time and budgets, capping acceleration. Second, where no such bottlenecks exist, extraordinary rates of progress become possible, and AI's cost becomes the limit. Under our assumptions, a task that is fully automated by AI still costs about one-sixteenth of what it would cost a researcher to perform, so accelerations approach about 16x and stop there. Lower AI costs would raise this ceiling proportionately, such that halving the cost of AI, for example, would double the rate of progress. If costs fall sharply, including for harder tasks that AI automates later, the fields without physical bottlenecks could accelerate far beyond these estimates.

These calculations have important limits that cut in both directions. We do not consider constraints at the level of a field rather than a research team. For instance, only so many experiments can run at facilities like the Large Hadron Collider. We also count only the cost of researcher and AI time, not the costs of machinery or inputs for experiments, which can be substantial in the health, physical, and life sciences. Nor do we consider simulations to replace physical experiments, which could ease bottlenecks in some fields. Finally, we do not

consider the quality of ideas. If AI identifies better hypotheses with higher chances of success, research could accelerate further; if research taste remains uniquely human, the acceleration would be slower.

Researcher expectations

The scenarios above characterize a range of possibilities. Researchers' own expectations can provide a unique guide grounded in expertise. To gather their viewpoints, our survey of researchers in the life and health sciences asked when they thought AI, working on its own, would match the best researchers in their field for different categories of research work (Figure 15). Respondents' expectations are high in the near term: over half expect AI to match top researchers at data analysis by 2027, and 38% say that AI has already reached this level. By 2030, 94% expect AI will match the best researchers in data analysis, 89% in writing, and 72% each in ideation and data gathering.

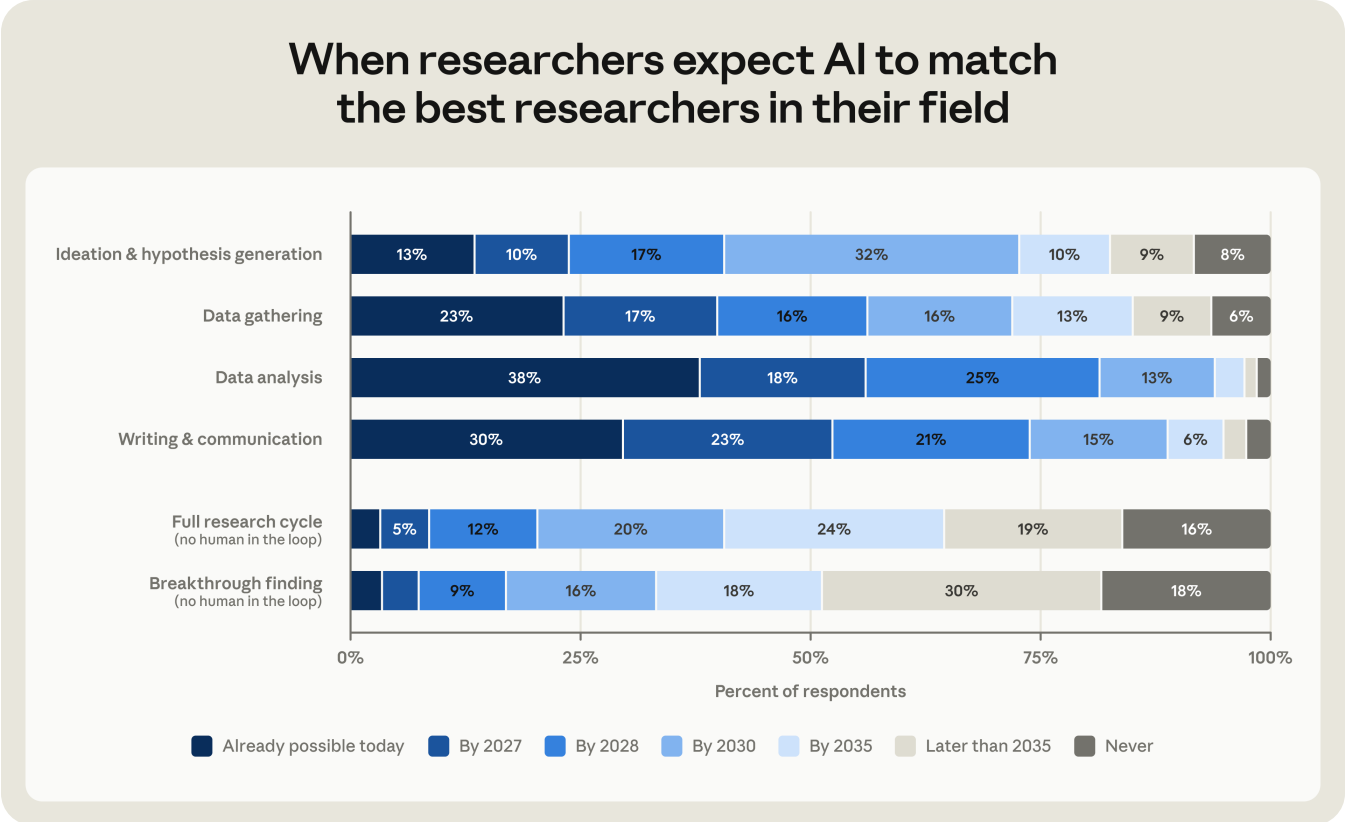


Figure 15. When researchers expect AI to match the best researchers in their field. Results come from our own survey. The top four rows correspond to when researchers expect AI, working on its own, to match the best researchers in their field in different task categories. The bottom two rows correspond to when researchers expect AI, without a human scientist in the loop, to go "from question to finding" and to produce "a finding your field would regard as a major advance" in their respective fields.

Expectations for ideation and data gathering are lower, consistent with the lower use of Claude for these tasks today.

Expectations for end-to-end AI research without humans in the loop are more conservative, but researchers still expect remarkable progress. About 40% of respondents expect AI to carry out full research projects by 2030, and about 64% by 2035, although 16% say it will never happen. Expectations that AI will autonomously produce scientific breakthroughs are more cautious still, with approximately half of researchers expecting this to occur only after 2035, if ever. Overall, many researchers expect high levels of automation in the next decade, broadly in line with our greater automation scenarios. How fast and to what extent these expectations materialize will depend not only on advances in AI as a software technology, but also on efforts to build automated labs that overcome physical barriers.

Comparing past survey evidence to realized outcomes suggests that AI progress has often been dramatically underestimated (Rosenberg et al. 2026). In a survey in mid-2022, for example, the median estimate for AI achieving gold-level performance in the International Mathematical Olympiad was 2030 among domain experts and 2035 among superforecasters (Karger et al. 2023). AI achieved this milestone in 2025, some five to 10 years ahead of those predictions. While forecasters have systematically underestimated progress in AI capabilities, there are a few exceptions, primarily in domains that require execution in the physical world, such as AI-assisted laboratory tasks in biology.

Discussion

If AI helps speed up scientific discovery, it could contribute to higher standards of living and faster progress in treating disease. Our estimates suggest the potential gains are substantial. But advancements in AI's research capabilities do not automatically translate into rapid progress; achieving such acceleration will also depend on how we respond to several constraints.

First, adoption can substantially lag capabilities. Diffusion of new technologies is typically slow (Comin and Hobijn 2010), and productivity gains tend to materialize only after organizations restructure around them (David 1990; Brynjolfsson, Rock, and Syverson 2021). Individual learning processes, friction

at the organizational or system level, and constraints on AI access or infrastructure can all slow adoption.

Second, research may continue to face bottlenecks that AI does not remove. Physical experimentation, which can be costly and slow, can substantially limit progress even with extremely capable AI. We see this especially in the bench sciences in the scenarios above. Choosing what to study may be another bottleneck. Research questions largely determine a research project's value, and AI's capacity to generate good questions remains unclear. If ideation is a bottleneck, it will limit acceleration even as answering a given question becomes much cheaper. Conversely, if AI outstrips humans in formulating high-quality hypotheses, acceleration could be more substantial, especially in domains where testing hypotheses is costly.

Third, AI-enabled research could raise the quantity or quality of research output, with different implications for science. A surge in new research papers could advance a field faster but also overwhelm scientists' capacity to absorb or verify the results. On the other hand, new technologies, from microscopes and telescopes to statistical software, have historically raised the quality of scientific insights, and AI can likewise make more comprehensive analyses and more ambitious projects feasible. And as papers become easier to produce, the research community may increasingly reward work that is especially novel or high quality. Whether more efficient research leads to greater quantity or greater quality will depend on researchers' choices and the incentives set by journals, funders, and institutions.

Fourth, the way research is organized may change in ways that are hard to anticipate. As AI becomes more capable and autonomous, AI systems could take the lead in verification, refereeing, and communicating important results, which would ease the absorption problem discussed above. Further in the future, AI research agents might also push the frontiers of science in large communities of their own, independently of humans, using robotics and automated labs. Such autonomous AI research would raise questions about how humans keep track of new knowledge, whether they can understand it, and how it is generated.

Fifth, more autonomous AI systems may increase the challenge of ensuring that AI-driven research remains aligned with human goals. Keeping humans involved in key research decisions, such as which questions to pursue and how

to pursue them, may matter for safety and human wellbeing as AI agents become more powerful and creative. Discoveries themselves, or the byproducts of research, can also carry risks, particularly in biology, where the same capabilities that speed up medical advances could also produce dangerous knowledge or materials. In such cases, deliberate checkpoints in the research process may be warranted, requiring us to accept slower progress to reduce risk.

Finally, the work that researchers undertake may change substantially. One effect may be a change in researchers' job satisfaction. If AI takes over enjoyable or rewarding tasks, satisfaction may fall and fewer people may choose research careers. If AI instead takes over tasks that researchers find tedious, satisfaction may rise. Administrative work is one potential target. In the US, principal investigators on federally funded projects report spending about 44% of their research time meeting administrative requirements rather than doing research (Schneider et al. 2020), and easing this burden could be a large boon, saving time and improving overall satisfaction. Ultimately, the net effect is uncertain and may change as AI's capabilities evolve. An optimistic view is that AI will expand opportunities for discovery, which is core to what makes research careers rewarding. If science is a voyage into the "endless frontier," AI may be a new vessel, allowing researchers to travel farther across a vast expanse of research questions and discoveries.

Appendix

Available here.

---

[1] We exclude countries in which Claude is not available when calculating the number of global researchers in each field.

[2] We obtain researcher counts from SciNet. For details on the measure, see the Appendix.

[3] Transcripts are assigned to fields by the research being pursued rather than the tools used, so coding work in other computational sciences is not counted as computer science.

[4] Subfield-level usage shares are estimated less precisely than field-level shares, particularly for subfields with lower session volume, because of greater sampling variation and classification noise. We therefore focus on broad patterns rather than the position of any individual low-volume subfield.

[5] While our analysis focuses on research tasks, non-research tasks such as teaching, service, and administration can take up a substantial share of researchers' time budget. However, our scientific usage classifier intentionally leaves aside undergraduate teaching and does not select administrative or routine work if it is not serving research or a study directly. Furthermore, our sample is not restricted to scientists

in academic positions, who naturally have higher teaching and service workloads. We therefore leave the study of AI use in these workflows to future research.

[6] In terms of individual tasks, drafting manuscripts is the most frequent task on both Claude (18.6%) and Claude Code (11.1%). On Claude, the next most frequent tasks are reviewing and synthesizing literature (6.8%) and responding to peer reviewer comments (6.7%). On Claude Code, they are designing, building, and refining artifacts (4.8%) and visualizing data and results (4.6%).

[7] Respondents were recruited from applicants to Anthropic's AI for Science and related grant programs and are affiliated with research institutions in the US and Canada. Because they applied for AI research credits, they are likely more intensive AI users than the typical scientist. See the Appendix for details.

[8] The differences in tasks performed on Claude Science may partly reflect who uses it. Any paid user can access Claude Science, but adopters are likely those with more specialized or demanding use cases. We therefore describe the differences across products as patterns of use.

[9] To identify artifacts for each Claude session, we use a classifier introduced in our prior work (Massenkoff et al. 2026). The classifier identifies the most prominent output Claude produced in the exchange. We categorize the more granular responses into two buckets: written artifacts and technical artifacts. Written artifacts primarily include documents, reports, academic papers, or presentations. Technical artifacts primarily include software output, such as scripts, code fixes, and data and charts. Artifacts that are not written or technical artifacts include analyses or summaries and explanations or advice.

Suggested citation

```
@techreport{hermle2026AIScience,
author = {Hermle, Johannes and Jones, Benjamin and Althoff, Lukas and
Heller, Ryan and Lyttelton, Thomas and Stein, Carolyn and Wilmers,
Nathan and Massenkoff, Maxim and McCrory, Peter},
title = {{AI} for scientific discovery: How researchers use
{Claude}},
institution = {The Anthropic Institute},
type = {Technical Report},
year = {2026},
month = nov,
url = {https://www.anthropic.com/research/how-researchers-use-claude}
}
```

References

Ahmadpoor, Mohammad and Benjamin F. Jones, "Decoding Team and Individual Impact in Science and Invention," *Proceedings of the National Academy of Sciences*, 2019, 116 (28), 13885–13890.

Alamos, Cristian and Lukas Althoff, "SciNet: The Anatomy of Science," dataset, 2026. https://www.anatomyofscience.com/

Amodei, Dario, "Machines of Loving Grace: How AI Could Transform the World for the Better," essay, October 2024. https://darioamodei.com/essay/machines-of-loving-grace

Anthropic, "Learning More about Claude's Mathematical Capabilities," research post, August 10, 2026a. https://www.anthropic.com/research/riemann-zeta

Anthropic, "Formalizing Fermat's Last Theorem," research post, September 4, 2026b. https://www.anthropic.com/research/formalizing-fermats-last-theorem

Appel, Ruth et al., "Anthropic Economic Index Report: Economic Primitives," research post, January 15, 2026. https://www.anthropic.com/research/anthropic-economic-index-january-2026-report

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics*, 2021, 13 (1), 333–372.

Bush, Vannevar, *Science, the Endless Frontier: A Report to the President*, Washington, DC: United States Government Printing Office, 1945.

Codreanu, Mihai et al., "AI in Science: Early Insights," arXiv preprint arXiv:2609.28504, 2026. https://arxiv.org/abs/2609.28504

Comin, Diego and Bart Hobijn, "An Exploration of Technology Diffusion," *American Economic Review*, 2010, 100 (5), 2031–2059.

David, Paul A., "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox," *American Economic Review*, 1990, 80 (2), 355–361.

Eloundou, Tyna et al., "GPTs Are GPTs: Labor Market Impact Potential of LLMs," *Science*, 2024, 384 (6702), 1306–1308.

Guo, Daya et al., "DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning," *Nature*, 2025, 645 (8081), 633–638.

Hao, Qianyue et al., "Artificial Intelligence Tools Expand Scientists' Impact but Contract Science's Focus," *Nature*, 2026, 649 (8099), 1237–1243.

Jones, Benjamin F., "Artificial Intelligence in Research and Development," in Ajay Agrawal, Erik Brynjolfsson, and Anton Korinek, eds., *The Economics of Transformative AI*, Chicago: University of Chicago Press, 2026, 37–61.

Jumper, John et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature*, 2021, 596 (7873), 583–589.

Karger, Ezra et al., "Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament," Forecasting Research Institute Working Paper 1, 2023. https://forecastingresearch.org/research/existential-risk-persuasion-tournament

Kobak, Dmitry et al., "Delving into LLM-Assisted Writing in Biomedical Publications through Excess Vocabulary," *Science Advances*, 2025, 11 (27), eadt3813.

Lambert, Nathan et al., "Tülu 3: Pushing Frontiers in Open Language Model Post-Training," arXiv preprint arXiv:2411.15124, 2024. https://arxiv.org/abs/2411.15124

Liang, Weixin et al., "Quantifying Large Language Model Usage in Scientific Papers," *Nature Human Behaviour*, 2025, 9 (12), 2599–2609.

Liu, Yinhong et al., "Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators," in *Proceedings of the First Conference on Language Modeling (COLM)*, 2024.

Mantiuk, Rafał K., Anna Tomaszewska, and Radosław Mantiuk, "Comparison of Four Subjective Methods for Image Quality Assessment," *Computer Graphics Forum*, 2012, 31 (8), 2478–2491.

Massenkoff, Maxim et al., "Anthropic Economic Index Report: Cadences," research post, June 26, 2026. https://www.anthropic.com/research/economic-index-june-2026-report

Merchant, Amil et al., "Scaling Deep Learning for Materials Discovery," *Nature*, 2023, 624 (7990), 80–85.

OpenAI, "An OpenAI Model Has Disproved a Central Conjecture in Discrete Geometry," research post, May 20, 2026a. https://openai.com/index/model-disproves-discrete-geometry-conjecture/

OpenAI, "On the Navier–Stokes Millennium Prize Problem," research post, September 8, 2026b. https://openai.com/index/navier-stokes-solution/

Parshall, Daniel and Andrea Lopez-Luzuriaga, "Measuring AI's Economic Reach: A Multi-Dimensional Task Taxonomy," George Washington University Center for Economic Research Working Paper 2026-005, 2026. https://www2.gwu.edu/~forcpgm/2026-005.pdf

Robinson-Garcia, Nicolas et al., "Meta-Research: Task Specialization Across Research Careers," *eLife*, 2020, 9, e60586.

Rosenberg, Josh et al., "How Accurate Have AI Progress Forecasts Been So Far?" research post, Forecasting Research Institute, September 22, 2026. https://forecastingresearch.org/research/ai-progress-accuracy-update

Schneider, Sandra L. et al., "2018 FDP Faculty Workload Survey: Report of Primary Findings," Federal Demonstration Partnership, 2020. https://thefdp.org/wp-content/uploads/FDP-FWS-2018-Primary-Report.pdf

Shah, Nihar B. et al., "When Is It Better to Compare than to Score?" arXiv preprint arXiv:1406.6618, 2014. https://arxiv.org/abs/1406.6618

Tomei, Philip Moreira and Bouke Klein Teeselink, "What Jobs Can AI Learn? Measuring Exposure by Reinforcement Learning," arXiv preprint arXiv:2605.02598, 2026. https://arxiv.org/abs/2605.02598

Tsoukalas, George et al., "Advancing Mathematics Research with AI-Driven Formal Proof Search," *Science*, 2026, 394 (6820), 234–239.

Varadi, Mihaly et al., "AlphaFold Protein Structure Database in 2024: Providing Structure Coverage for over 214 Million Protein Sequences," *Nucleic Acids Research*, 2024, 52 (D1), D368–D375.

Wang, Hanchen et al., "Scientific Discovery in the Age of Artificial Intelligence," *Nature*, 2023, 620 (7972), 47–60.