ANTHROP\C

Appendix to “AI for Scientific Discovery: How Researchers Use Claude”

November 5, 2026

A. Usage sample

Our analysis is based on two samples of Claude session transcripts that involve scientific work from Chat and Cowork (henceforth labeled Claude for simplicity) and Claude Code. In a supplemental analysis, we also study a sample of sessions on Claude Science. The sample covers accounts on consumer plans (Free, Pro, and Max plans) from August 2 to September 30, 2026. For Claude Code, we include only sessions in which a person works with Claude interactively and exclude programmatic usage.

All analyses in this report are conducted through a privacy-preserving analysis pipeline. All classifications are done exclusively by classifiers that use language models: each session's transcript is shown to a language model together with a classification prompt, and the model assigns a label from a closed list of options. All numbers in this report are reported only for cells that pass a minimum number of accounts.

The sample is built in three steps in the same way for each product. First, we sample a set of session transcripts from each of the Claude products analyzed in this study. Second, we use a classifier to determine whether a session transcript involves scientific or research work, judged by the activity in the transcript. This yields 148,734 session transcripts for Claude and 86,884 for Claude Code. Third, a classifier assigns each of the session transcripts to one subfield and one task according to the SciNet taxonomy described in the next section. The classifier successfully identifies a subfield and task for 94% of scientific Claude and Claude Code session transcripts. The remaining transcripts are too unspecific to be placed and are left out.

Our analysis covers session transcripts on core research tasks, which are tasks performed by scientists to produce new research findings and papers (see the description of the SciNet taxonomy in the next section for details). The remaining tasks, which we label non-research tasks, such as teaching or service, are not tied to producing new findings or research papers. The sample of session transcripts covering core research tasks, which we use as our main analysis sample, contains 128,914 sessions on Claude and 78,135 on Claude Code.

Because we sample a constant number (as opposed to a constant share) of session transcripts per hour and product, the share of usage sampled varies

across products and time. All analyses therefore weight each session transcript by the inverse of its sampling probability: the total number of session transcripts in its product and hour, divided by the number of sampled transcripts in that same product and hour.

While our analysis focuses on core research tasks, non-research tasks can take up a substantial part of a researcher's time budget. In a survey of academic researchers, the estimates of Schneider et al. (2020) suggest that such tasks take up approximately two-thirds of researcher time, while SciNet estimates 63.4%. 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 have higher teaching and service workloads. As a result, only 7.7% of sessions on Claude and 4.2% on Claude Code are mapped to non-research work, which is substantially lower than academic scientists' time allocation on these tasks.

B. SciNet taxonomy

SciNet is a dataset describing researcher tasks across all domains of science. It comprises a hierarchy of 6 domains, 34 fields within those domains, 318 subfields, 7,262 tasks, and 90,583 task steps (Alamos and Althoff 2026). This hierarchy structures our analysis, helping us map sessions that researchers have with Claude to the work that their process of scientific discovery requires. Beyond the hierarchical field and task taxonomy, SciNet contains rich detail on various aspects of scientific work, including time researchers spend on each task and whether regulatory restrictions affect the use of AI in a given task.

SciNet is built by combining large language models with established survey designs from O*NET, a widely studied taxonomy of labor market tasks built from human expert interviews. SciNet is validated using data from thousands of academic papers, research protocols from wet-lab sciences, human expert feedback, and human ratings for scientific tasks on O*NET. Using large language models to construct SciNet allows for a level of detail that would be prohibitively expensive or logistically infeasible with human expert input alone.

SciNet is licensed under CC BY 4.0 (link). We aggregated and combined SciNet data with our own measures, as described in this Appendix.

Details. SciNet starts out by building on a taxonomy of 26 fields as defined by OpenAlex, a database of global research outputs including papers and preprints. It then uses human expert input to merge and split those into 34 refined fields and group them into six domains. SciNet then uses large language models to propose typical subfields for each field and sorts OpenAlex's research topics into them.

Having established a field hierarchy that links to OpenAlex's academic output, SciNet then defines tasks through language models and human review. It begins with a large language model proposing and the authors curating 30 universal tasks, shared by researchers across all fields, and 69 domain-specific tasks. It then asks a large language model to suggest tasks specific to subfields, which refines domain tasks. Around one-fifth are later judged as trivial and therefore removed. All task descriptions closely follow O*NET's task description style. SciNet then validates all tasks by drawing around 100 papers per subfield and recording which task each paper performs. For work that is not covered by existing tasks, new tasks are proposed. New tasks are kept only if at least two papers propose them and at least 5% of papers perform them. Finally, all tasks are broken into steps, for which various characteristics are computed (see the section below). These steps are validated using public protocols from protocols.io, where wet-lab scientists record rich information on their experiments' steps and researcher time allocated to them.

Example

- **Domain:** Life Sciences
- **Field:** Biology
- **Subfield:** Evolutionary Biology
- **Task:** Conduct phylogenetic analyses to infer evolutionary relationships among taxa.
- **Step:** Run a maximum likelihood tree search in IQ-TREE (or RAxML) under the selected partitioned models, specifying an appropriate outgroup to root the tree.

Task characteristics

In addition to the taxonomy that links specific researcher tasks to a given subfield, SciNet contains information on characteristics of those tasks and their steps. When a task is relevant to multiple subfields, these characteristics are usually specific to a task-subfield combination.

The task characteristics we use are:

- **Hierarchical position.** SciNet distinguishes universal tasks that are shared by all fields, tasks specific to a domain, tasks specific to fields, and tasks specific to subfields.
    - Universal tasks (0.4% of tasks, e.g., "Apply statistical methods to test hypotheses and estimate effect sizes")
    - Domain-specific tasks (1.8% of tasks, e.g., "Formulate hypotheses about biological mechanisms, pathways, or evolutionary processes")
    - Field-specific tasks (4.4% of tasks, e.g., "Identify proteins from mass spectrometry data by matching spectra to reference sequences" in the field of biology)
    - Subfield-specific tasks (93.3% of tasks, e.g., "Analyze differences in microbial growth rates, gene expression, or metabolic outputs between treatment groups" in the subfield of microbiology)
- **Task category.** Tasks are grouped into categories:
    - Reading and knowledge acquisition (2.0% of tasks)
    - Ideation and hypothesis generation (7.4% of tasks)
    - Data gathering (40.8% of tasks)
    - Data analysis (42.9% of tasks)
    - Theoretical analysis (2.5% of tasks)
    - Design and development (2.2% of tasks)
    - Writing and communication (0.3% of tasks)
    - Peer review and service (0.1% of tasks)
    - Mentorship and teaching (0.2% of tasks)
    - Administration (1.6% of tasks)

    Format: Categorical variable specific to each task. Assigned by LLM.

- **Research versus non-research tasks.** Core research tasks are those that are done to generate new scientific findings and that typically result in an academic paper, including ideation, data gathering, data analysis, theoretical analysis, and writing. Non-research tasks are those that are done by researchers

for goals not directly related to their own academic output, such as teaching, mentorship, and service. In our analysis, we focus on core research tasks.

Format: Binary variable specific to task × subfield combinations. 23,543 research task × subfield combinations, 3,260 non-research task × subfield combinations. Assigned by LLM.

- **Researcher hours per task.** SciNet estimates the hours the average researcher in a given subfield spends on a given research task over the course of a year. Time estimates are generated first at the granular step level and then aggregated to the task level. With the time per task instance in hand, assessing overall researcher time spent on a task per year requires additional information on the annual number of task instances. SciNet estimates this in two steps: the number of instances per paper, and the annual number of papers per researcher. The latter estimate is derived from OpenAlex, a database on hundreds of millions of research papers that allows computing per-researcher counts for each subfield. It assigns fractional authorship to each coauthor. Total hours spent on each task are then the hours per instance multiplied by instances per paper and papers per year.

  Format: Hours specific to task × subfield combinations. Computed as a combination of paper counts by subfield on OpenAlex (2015–2019) and LLM-based time estimates validated via (1) time recorded on research protocols for wet-lab sciences from protocols.io and (2) the number of papers that contain a task, classified using an LLM and full text from 31,500 open-access papers on OpenAlex (2000–2020) as a proxy for task frequency.

- **Regulatory restrictions.** This measure captures the degree to which researchers face regulatory barriers in using AI assistance for a given task. It is scored on a scale from 0 to 4 and is based on Parshall and Lopez-Luzuriaga (2026). The scale ranks regulatory barriers from least to most binding, with 0 for none, 1 for social or market norms, 2 for professional standards or liability, 3 for statute, and 4 for tasks that require a personally accountable human, such as giving sworn testimony or consent. Each task's steps are rated separately, and a task is assigned the average value across all steps (weighting by steps' time share).

  Format: Score from 0 to 4 specific to each task (sometimes we rank the underlying values to range from 0 to 100 instead). Assigned by LLM.

- **Suitability for reinforcement learning.** This measure captures the extent to which a task is amenable to reinforcement learning (RL). RL environments have been an important part of recent progress in AI capabilities. In an RL environment, an AI model attempts a problem, receives feedback on its attempt, and uses that feedback to improve. Problems tend to be suitable for RL when they (1) are verifiable, (2) are simulatable, and (3) have high

feedback density, three key dimensions in Tomei and Klein Teeselink's (2026) rubric for RL feasibility. Verifiability ranges from low (e.g., requiring contested human expert judgment) to high (fully automatic check). Simulatability ranges from low (requiring real people and real stakes) to high (fully digital and easy to replicate). Feedback density ranges from rare (delayed and expensive signals) to continuous (immediate and cheap). Each dimension is rated by an LLM at the step level and then aggregated to the task level by averaging across the relevant steps (weighting by steps' time share). In a principal component analysis (PCA), the three dimensions are then summarized into an RL index that explains 79.5% of the variance across those variables.

Format: Scores for each dimension from 1 to 10 specific to each task, aggregated into a principal component with mean 0 and standard deviation 1 across tasks (sometimes we rank the underlying values to range from 0 to 100 instead). Assigned by LLM.

In addition to task characteristics from SciNet, we use additional characteristics created with assistance from large language models:

- **Theoretical AI capability and AI capability scores.** The AI capability score considers how well current AI capabilities can perform each task, relative to other tasks, through pairwise comparisons of the underlying SciNet research steps. We construct this measure as follows: a judge model is shown two SciNet steps, each with its task and the associated subfield, field, or domain it belongs to. Given the two steps, the judge chooses the step for which AI's typical output without human help would come closer to what an experienced researcher in that field would accept without substantial correction. AI is defined as an agent on a computer that uses the best publicly available general and specialist AI systems, communicates with people only in writing, and has no physical or laboratory integration available. Steps are paired over eight rounds, two rounds with random pairings and then six rounds with steps of similar rating. A Bradley–Terry model fitted to the 317,560 comparisons gives each step an ordered rating. The judge can also state that AI cannot do a step at all. The 16% of steps that are judged this way in both of their first two comparisons are placed below all others. Because the comparisons identify the order of the ratings better than the distances between them, we express each step's rating as its percentile among all rated steps. A given task takes the mean percentile of its steps, and our task measure is the percentile rank of that mean across all tasks (100 = most capable). If the same task has a different set of steps in different fields, that task will have a separate rank and percentile for each field or subfield in which it appears. For example, the universal SciNet task "Acquire or access primary data sources relevant to research questions" has very different steps in different fields.

The theoretical capability measure is similar in objective to the exposure measure of Eloundou et al. (2024). Both use an LLM judge to assess how well AI can perform specific tasks. Eloundou et al. (2024) used GPT-4 to place each O*NET task in one of three categories, based on whether a language model could cut the time it takes by at least half. We use pairwise comparisons rather than categorical ratings for two reasons. First, there is no given absolute scale for how well AI can perform a given research step. Where no such scale exists, the literature suggests that comparative judgments are typically more accurate than absolute judgments for both human judges (Mantiuk, Tomaszewska, and Mantiuk 2012; Shah et al. 2014) and LLM judges (Liu et al. 2024). Second, the AI capability score ranks tasks far more finely. The AI capability score correlates with the broad task exposure labels of Eloundou et al. (2024) when we apply their rubric to SciNet tasks, but the AI capability score distinguishes between the tasks that their rubric places within the same category. For example, 58% of tasks are labeled as "directly exposed" according to the Eloundou et al. (2024) rubric but with no further distinction between them.

Format: Percentile AI rank from 0 to 100 for each task (computed separately for each field or subfield where the task's steps differ). Assigned by LLM.

- **Cognitive, physical, and interpersonal work.** A language model is shown each SciNet task step, with its task and associated subfield, field, and domain, and assigns the step to exactly one of three categories. The categories are defined as material things worked on with one's own body (physical), information and ideas (cognitive), or people, through any medium (interpersonal). A step with any physical component is classified as physical. Only steps without any physical component are classified as cognitive or interpersonal, whichever dominates. At the task level, the measure is the share of its steps in each category, with each step counting equally.

  Format: Percentage of a task's steps in each category, summing to 100% (computed separately for each field and subfield where a given task's steps differ). Assigned by LLM.

Researcher counts

SciNet measures the number of active researchers in a given field and subfield using data on hundreds of millions of working papers and publications from OpenAlex. First, the corpus on OpenAlex is restricted to active authors with at least five papers in total—at least one before 2020 and at least one since 2020. Second, SciNet randomly samples 746,517 researchers (sampled separately in each country) and classifies each researcher's subfield based on the titles,

journals, and years of up to 40 of their papers. The main classifier is Claude Haiku 5.5, validated with 1,500 classifications from Sonnet 5.5, which is given each paper's abstract as additional information (85% agreement on field, 74% agreement on subfield). Lastly, the number of researchers is estimated as a subfield's share of researchers determined in the previous step, times the total number of researchers in a given country, and summed over countries.

For further information on SciNet, see https://github.com/lukasalthoff/scinet.

C. Classifying Claude sessions

We use a privacy-preserving analysis pipeline in which model-based classifiers are provided with a session transcript and a prompt and asked to assign labels such as a field and task from a closed-form response rubric. Results are reported only in aggregate. For a given cell, estimates are only reported if they pass a minimum number of accounts.

Classifier overview

We use five classifiers to characterize scientific Claude sessions:

- **Scientific work screener:** An initial classifier that identifies whether a given session transcript covers scientific use. Only transcripts that involve scientific use are reviewed and classified by all other downstream classifiers.
- **Scientific field and task classifier:** Identifies the domain, field, and subfield of a given scientific transcript. Then, the classifier identifies the task being worked on among the tasks of the subfield picked in the previous step. The classifier can pick any of the universal, domain-specific, field-specific, or subfield-specific tasks relevant to the particular subfield. The classifier also returns a confidence level between 1 and 5 for the chosen subfield and task.
- **Claude contribution classifier:** Rates the contribution of Claude relative to the task picked by the scientific field and task classifier. Contribution can vary from "none" to "generated" (Claude completed the task's work itself).
- **Memory-based role and career stage classifier:** Identifies only the scientific role and career stage using only explicit statements in the memory block. This classifier is only used for Claude, and not other surfaces.
- **Memory-based scientific field classifier:** Identifies only the scientific field of the researcher using only explicit statements in the memory block. This classifier is only used for Claude, and not other surfaces.

In addition, we use data from additional classifiers that we introduced in Appel et al. (2026) and Massenkoff et al. (2026), for which more detail is provided below under "Supplementary classifiers."

Preprocessing. Before a session transcript is passed to a classifier, it is reduced to the parts the classifier should read and shortened to a fixed maximum length. For the screener, the field and task classifier, and the contribution classifier, we remove the system prompt. Each turn is limited to 10,000 characters, and the transcript as a whole is limited to 100,000 characters. If it is longer, we keep whole turns from both ends, alternating between the first and the last remaining turn. The two memory-based classifiers see no part of the conversation. For these, we extract only the memory block, up to 100,000 characters, and a transcript without a memory block is recorded as such.

Classifier prompts

Scientific work screener

Model: Sonnet 5

```
Human: Consider the following conversation:

<conversation>
{TRANSCRIPT}
</conversation>

Classify whether this conversation is SCIENTIFIC OR SCHOLARLY
RESEARCH WORK: the user is doing, or getting help with, a task that
belongs to the work of a scientist or scholar, in any of the fields
listed below.

Research work is the set of tasks by which new scientific or
scholarly knowledge is produced, and the tasks that directly serve
producing it: framing questions and hypotheses; literature review;
designing studies, experiments, trials, surveys, corpora, and
instruments; ethics approvals; collecting data in the lab, field,
clinic, or archive; cleaning, coding, analyzing, modeling, and
simulating data; proofs and derivations; close reading, source
criticism, doctrinal and interpretive analysis; building devices,
```

materials, software, or creative works to answer a research question; writing papers, dissertations, research grant proposals, conference talks, and responses to reviewers; peer review, grant review, and editorial work; supervising doctoral students and postdocs; coordinating collaborators; managing a project's budget, personnel, equipment, and computing; and learning the methods a specific study needs. Routine work counts when it belongs to a study, however mundane the task.

Fields, all counting equally: Chemistry; Earth & Planetary Sciences; Engineering; Environmental Science; Materials Science; Physics & Astronomy; Computer Science; Mathematics; Statistics; Agricultural Sciences; Biology; Biomedical Sciences; Neuroscience; Veterinary Medicine; Medicine & Clinical Sciences; Nutrition & Dietetics; Public Health & Epidemiology; Anthropology; Business & Management; Communication & Media Studies; Economics; Education; Geography; Law; Library & Information Science; Political Science; Psychology; Sociology; Arts; History; Languages & Linguistics; Literature; Philosophy; Religion. Empirical, computational, formal, interpretive, doctrinal, normative, and design-based research all count, in academic, industrial, government, clinical, or independent settings.

EVIDENCE. The user's own messages must show the work is research: a research output they are producing or evaluating (paper, manuscript, preprint, doctoral thesis or dissertation, research grant proposal, conference talk or poster, peer review, research dataset or research software); a study they are running (experiment, trial, survey, fieldwork, participants, specimens, ethics application, hypothesis under test, corpus or archive assembled for a study); or a research role (lab or research group, doctoral candidate or postdoc, supervisor of doctoral students or postdocs, reviewer or editor). A stated role is not enough on its own for routine tasks: a script, a budget, or an email counts only when the study or output it serves is shown. A scientific topic, technical depth, academic tone, or the word "research" is not evidence, and the AI proposing a hypothesis, design, or citations does not make the conversation research. The conversation may be a fragment: judge from what is shown and do not assume a research context that is not shown. If the conversation mixes activities, answer research_work when any substantial part of the user's activity is research work with such evidence.

NOT research work, even when a paper, study, experiment, or research role is mentioned (the majority of conversations):
- Coursework at any level: homework, exam preparation, class tutorials, and any "research paper", lab report, literature

```
  review, or thesis for a course. Student work counts only at
  doctoral level: a doctoral student's own dissertation research or
  their work on their research group's studies. Students below
  doctoral level do not count, whether on a bachelor's or master's
  thesis or as research assistants.
- The same activity as another occupation's work: clinical care;
  legal advice; software engineering, data engineering, and model
  development for a product, service, or internal tool, however
  experimental the process, and testing or comparing tools for a
  product or out of interest; business, financial, and operational
  analytics; product A/B tests; market, user, UX, equity, and policy
  research for a client, employer, or product decision; journalism;
  consulting; a company's "lab" or "R&D" team building a product
  (industrial R&D counts only when it is producing a study, paper,
  or research dataset).
- Using research results outside a study: following a guideline
  for a patient, writing a hospital protocol or patient handout from
  existing guidelines, using a published model in a product, reading
  papers to make a personal or business decision.
- Teaching other than supervising doctoral students and
  postdocs: school and undergraduate instruction, teaching
  assistantships, grading, tutoring, advising undergraduate or
  master's projects, preparing lecture courses.
- Learning with no study in view, and a literature review or
  reference list with no study or output in view: a stated role
  alone is not enough for either.
- Curiosity, fact-checking, news, science communication for lay
  audiences, questions about research as a career; personal
  genealogy, devotional study, self-experimentation, and hobby
  observation; consumer health questions from a patient's or carer's
  perspective, even when studies are cited; non-research grants
  (nonprofit, community, arts, program, business); fiction and
  roleplay with research settings.

EXAMPLES, in pairs (research_work / not_research_work):
- "Fitting a Lorentzian to our spectrometer data for the paper;
  the fit won't converge" / "Clean this CSV and run a t-test
  comparing group A and group B" (no study shown)
- "For my dissertation I'm reconstructing a timeline from parish
  burial registers" / "I need a 10-page research paper on the causes
  of World War I for my history class"
- "Revising my article on Hamlet's delay for a Shakespeare
  journal; reviewer 2 says I ignore the First Quarto" / "My paper
  argues Hamlet's delay is moral reasoning; sharpen my thesis
  statement" (no output, study, or role shown)
```

```
- "Our IRB approved the gig-worker study; refine the
  semi-structured interview guide" / "I'm a UX researcher; write the
  interview guide for our onboarding study"
- "Proving the bound in Lemma 3 is tight for our paper; the
  construction fails for odd n" / "Prove that the sum of the first n
  odd numbers is n squared"
- "We built a microfluidic chip to test whether channel geometry
  affects cell sorting; design the validation experiment" / "Build
  me a React dashboard for our sales team's weekly numbers"
- "Discovery biology at a pharma company; writing up our
  kinase-inhibitor mechanism study for journal submission" / "ML
  research team at a fintech; tune our fraud model's threshold for
  next quarter's release"
- "Postdoc in ecology; I need to learn mixed models to analyze
  my transect data" / "PhD student in sociology; explain what a
  mixed-effects model is"
- "My PhD student's chapter uses difference-in-differences; what
  should she check before her committee?" / "I'm supervising three
  undergraduate capstone projects; what rubric should I use?"
- "Our trial's data monitoring board meets Friday; draft the
  interim safety summary from these adverse-event tables" / "My
  patient's HbA1c is 8.2 on metformin; add a second agent?"
- "For our cohort study, compute age-standardized incidence from
  these registry counts" / "A trial showed statins cut stroke risk;
  should my father be on one?"
- "We randomized 400 villages for our cash-transfer study; check
  balance on baseline covariates" / "Checkout-page experiment:
  treatment converted 4.1% vs 3.8%; ship it?"
- "Our RNA-seq pipeline fails at alignment; here is the log from
  our knockout study" / "Our clinical lab's variant-calling pipeline
  flags too many uncertain variants for patient reports"
- "Writing a law-review article on preemption; reconcile these
  circuit cases" / "Draft a memo for our client on whether the state
  statute is preempted"
- "Draft the budget justification for our NSF proposal" / "Draft
  the needs statement for our food bank's community-foundation
  grant"
- "Reviewing a manuscript for a sociology journal; does the
  mediation analysis support the causal claim?" / "Summarize this
  intermittent-fasting paper so I can decide whether to try it"

The default is not_research_work. Answer research_work only when
there is clear evidence of scientific or scholarly research work as
defined above. When in doubt, answer not_research_work.
```

```
Output your answer in <classification> tags. Write ONLY
research_work or not_research_work.

Assistant: <classification>
```

Scientific field and task classifier

Model: Sonnet 5

Note: the order of domain, field and subfield in step 1 as well as the order of tasks in step 2 is shuffled across transcripts. The None option is always appended as the last option.

```
STEP 1:

You are analyzing a conversation between a user (a person doing
scientific research) and a chatbot. Your task is to determine which
research discipline the work in the conversation belongs to, using
the taxonomy below.

Treat the conversation as a single unit and classify its single most
prominent research activity. Pick exactly one path — if several
activities appear, choose the one most central to what the user is
trying to accomplish.

Step 1: In 1-2 sentences, summarize this conversation and the
research activity in it.

Step 2: Pick the discipline of the work the chatbot is helping with,
at the most specific level the conversation supports. Judge by the
research question being pursued, not by the tools used. If the field
of research is unclear, do not pick from the taxonomy: answer
{"level": "None", "name": "unclear_field"} as your single pick.

Levels:
- [Domain] (6 options)
- [Field] (34 options)
- [Subfield] (318 options)

Confidence guide:
```

```
For each picked item, assign a confidence score from 1 to 5 using
the following rubric. The rubric asks a single question: how certain
am I that this is what the user's research activity actually is in
this conversation?
5 - Certain. Explicit, direct textual evidence.
4 - High confidence. Clearly supported, inferable with only
    minor interpretation.
3 - Moderate confidence. Plausible, but the evidence is
    partial, implicit, or open to alternative readings. A reasonable
    reviewer could either agree or disagree.
2 - Low confidence. Possible, but concluding so requires
    substantial interpretation.
1 - Very low confidence. Implied by context at best, with no
    clear textual support. Picks at this level are near-noise —
    exclude unless nothing stronger exists.

[Full taxonomy tree: an indented list of the [Domain], [Field] and
[Subfield] options]

The conversation to classify:
{conversation}

The conversation to classify is quoted above.

Output JSON only:
<answer>{"summary": "<1-2 sentences>", "picks": [{"level":
"Domain|Field|Subfield|None", "name": "<exact name>", "confidence":
1-5}], "notes": "<optional>"}</answer>

STEP 2:

The conversation to classify:
{conversation}

You are analyzing the conversation quoted above, between a user (a
person doing scientific research) and a chatbot. You previously
determined the subfield of the research based on the following
summary:
Summary: {summary}
{subfield}

Your task now is to pick the research task that the chatbot is
helping the user with, from that subfield's task menu.
```

```
Pick exactly one task — the one most central to what the user is
trying to accomplish. Pick "None" if the chatbot is not helping with
any task in the menu.

Task options for [Subfield] {subfield}:
- {task}
- {task}
- …
- None

Confidence guide:
For each picked item, assign a confidence score from 1 to 5 using
the following rubric. The rubric asks a single question: how certain
am I that this is what the user's research activity actually is in
this conversation?
5 - Certain. Explicit, direct textual evidence.
4 - High confidence. Clearly supported, inferable with only
    minor interpretation.
3 - Moderate confidence. Plausible, but the evidence is
    partial, implicit, or open to alternative readings. A reasonable
    reviewer could either agree or disagree.
2 - Low confidence. Possible, but concluding so requires
    substantial interpretation.
1 - Very low confidence. Implied by context at best, with no
    clear textual support. Picks at this level are near-noise —
    exclude unless nothing stronger exists.

Output JSON only:
<answer>{"leaf_picks": [{"subfield": "<exact subfield name>",
"name": "<exact task text>", "confidence": 1-5}]}</answer>
```

Claude contribution classifier

Model: Sonnet 5

Note: this classifier is chained to the scientific field and task classifier.

```
The conversation to classify:
{conversation}
```

```
You are analyzing the conversation quoted above, between a user (a
person doing scientific research) and a chatbot. You previously
determined the subfield of the research and the research task based
on the following summary:
Summary: {summary}
Subfield: {subfield}
Task: {chosen_task}

Classify the assistant's contribution to this task in this
conversation:

- none: the assistant did not contribute to any work within this
  task's scope.
- form_work: the assistant did form work on the user's own
  material for this task, without adding content of its own. For
  text: spell and grammar checks, rewording, translation,
  reformatting, or restating it. The same applies to the task's
  other artifacts: reformatting or tidying the user's code without
  changing its behavior, restyling figures or tables, or converting
  files between formats is form_work.
- advised: the assistant contributed guidance without producing
  the task's work itself — it diagnosed problems, evaluated whether
  a claim, proof, model, or analysis is sound, answered technical
  questions, or recommended methods, designs, or literature, and the
  user does the work.
- co_produced: the assistant produced parts of the task's work,
  and the user did other parts — the work within the task's scope
  was shared between them.
- generated: the assistant did essentially all of this task's
  work itself. The task's inputs — the data, materials, or results
  it works on — may come from the user; what matters is who did the
  work within the task's scope. If the user contributed only inputs,
  direction, and acceptance, answer generated.

RULES.
- The task description defines the task's scope: the work that
  makes up this task.
- Every level is judged within this task's scope only.
  Contributions to other tasks — however substantial — do not count
  here: if all of the assistant's contributions fall outside this
  task's scope, answer none.
- Text that describes this task's work is not itself this task's
  work: if the assistant only edited or translated such text, answer
  none here — that contribution belongs to the writing task. For
  writing and communication tasks the text IS the work, so editing
  it counts as normal.
```

```
- Judge what the assistant actually contributed in its
  responses, not what the user asked for. If the conversation
  touches the task at several depths, answer with the deepest
  contribution the assistant actually made (none < form_work <
  advised < co_produced < generated).
- The user supplying the task's inputs does not make the work
  theirs: when the work is within this task's scope, a full analysis
  of the user's data, or a complete draft written from the user's
  results, is generated. The user doing part of this task's own work
  — writing sections the assistant extends, code the assistant
  completes — makes it co_produced.
- For learning and literature tasks: teaching or explaining
  concepts and methods in response to the user's questions is
  advised; producing a standalone synthesis, review, or tutorial
  from a topic is generated; reformatting the user's own notes or
  summaries is form_work.

Output JSON only:
<answer>{"task_contribution": {"level":
"none|form_work|advised|co_produced|generated"}}</answer>
```

Memory-based scientific field classifier

Model: Sonnet 5

```
Human: Consider the following memory record that claude.ai stores
about a user:

<memory>
{MEMORY}
</memory>

Classify, from the memory record above and nothing else, the
research discipline of the user, if the user is a scientist or
scholar: a person whose stated occupation or position is scientific
or scholarly research work.

The memory record is what claude.ai has stored about this user
across past conversations; it is not a conversation.

ANSWERS.
```

```
- A field code (F01-F34) when the memory establishes a research
  occupation and settles the field; the domain-only code (D1-D6)
  when it settles the domain but not the field.
- discipline_not_researcher: the memory explicitly states an
  occupation or status that is not research work: software engineer,
  company data scientist or analyst, physician or nurse without a
  research title, resident or clinical fellow, lawyer, school
  teacher, teaching-only instructor, UX, market, or user researcher,
  and the like; any student below doctoral level, even with a thesis
  or research-assistant job; retired from a non-research occupation.
- discipline_unclear: the occupation is unclear ("works at a
  university", "works in R&D at a company", "PhD" with no current
  role), or the user is a researcher whose discipline the memory
  does not settle ("postdoc", "works in STEM").
- discipline_no_memory: the record is the line "[transcript
  elided: no memory blocks present]", or contains no statement about
  the user at all.

EVIDENCE. Answer only from what the memory explicitly states about
the user's own occupation, position, degree program, or own research
output (their doctoral dissertation, their own papers, preprints, or
research grants); never infer from other facts, however suggestive.
A research occupation is a research position or title (professor,
lecturer, principal investigator, research scientist or fellow,
staff scientist, postdoc, doctoral student — not positions marked
teaching-only), or a research discipline stated as the user's
profession ("physicist", "professor of economics", "PhD student in
linguistics"). A workplace alone never settles it, whether
university, institute, or company lab. Topics the user asks about,
interests, hobbies, courses, stored preferences, and statements
about other people do not count. Academic, industrial, government,
clinical, and independent settings all count.

For the discipline, choose the field whose researchers would carry
out the user's stated research, not the field of the tools they use:
a biologist who codes is F14, a computer scientist working on
protein structure is F07, an economist using machine learning is
F27. When the memory names a specialty, answer its parent field: an
organic chemist is F18, an astrophysicist is F23, a machine-learning
researcher is F07, an econometrician is F27, an archaeologist is
F24, an epidemiologist is F12, a cardiologist doing clinical
research is F10. Engineering of any kind is F20. Clinical and
medical research is F10; laboratory biomedical research without
patients is F15. If the memory names two or more disciplines for the
user, take the one named first. Answer at the most specific level
the memory settles: a named field always beats its domain code, and
```

```
the domain code is only for memories that name the domain but no
field. Prefer a domain-only code over guessing a field: "social
scientist" is D6, "life scientist" without more is D4.

<disciplines>
D1 Arts & Humanities (domain only: field not stated)
- F01 Arts: creation, analysis, and history of visual, sonic,
  spatial, and performed works
- F02 History: past human events and societies examined through
  primary sources and material evidence
- F03 Languages & Linguistics: language structure, use, history,
  acquisition, and processing
- F04 Literature: written verbal art interpreted through close
  reading, context, and theory
- F05 Philosophy: rational inquiry into morality, knowledge,
  reality, and value
- F06 Religion: sacred texts, doctrines, belief systems, and
  religious communities

D2 Formal Sciences (domain only: field not stated)
- F07 Computer Science: algorithms, software systems, machine
  learning, and computational theory
- F08 Mathematics: abstract structures and logical relationships
  studied through proof
- F09 Statistics: collection, analysis, and interpretation of
  data under uncertainty

D3 Health Sciences (domain only: field not stated)
- F10 Medicine & Clinical Sciences: diagnosis, treatment, and
  mechanistic study of disease across clinical specialties
- F11 Nutrition & Dietetics: nutrient metabolism, dietary
  patterns, and food environments
- F12 Public Health & Epidemiology: population-level disease
  distribution, determinants, and prevention

D4 Life Sciences (domain only: field not stated)
- F13 Agricultural Sciences: crop, livestock, aquatic, and
  forest production systems; soil and food systems
- F14 Biology: organismal structure, function, diversity, and
  evolution from cells to ecosystems
- F15 Biomedical Sciences: molecular and cellular mechanisms of
  disease; drug development
- F16 Neuroscience: neural bases of behavior, cognition, and
  disease at any level
- F17 Veterinary Medicine: disease and treatment in non-human
  animals
```

```
D5 Physical Sciences (domain only: field not stated)
- F18 Chemistry: composition, structure, properties, and
  transformation of matter
- F19 Earth & Planetary Sciences: composition, structure, and
  history of Earth and other planetary bodies
- F20 Engineering: design, building, and optimization of
  systems, structures, and devices
- F21 Environmental Science: pollutants, hazards, ecosystems,
  and sustainability at the human-Earth interface
- F22 Materials Science: properties, processing, and performance
  of materials
- F23 Physics & Astronomy: fundamental laws of matter, energy,
  space, and time

D6 Social Sciences (domain only: field not stated)
- F24 Anthropology: human biological variation, cultural
  diversity, and archaeological remains
- F25 Business & Management: firm strategy, marketing, finance,
  operations, and management
- F26 Communication & Media Studies: mass media, journalism,
  digital platforms, and public discourse
- F27 Economics: resource allocation, markets, incentives, and
  distribution
- F28 Education: teaching, learning, and schooling as an object
  of study
- F29 Geography: spatial patterns of human activity and physical
  landscapes
- F30 Law: legal rules, institutions, and normative frameworks
  as an object of study
- F31 Library & Information Science: organization, access, and
  preservation of recorded knowledge
- F32 Political Science: power, governance, institutions,
  elections, and political thought
- F33 Psychology: mental processes, behavior, and individual
  differences
- F34 Sociology: social structures, institutions, and collective
  behavior
</disciplines>

EXAMPLES (memory statement -> answer):
- "User is a professor of economics at a state university." ->
  F27
- "User is a PhD student in computational linguistics." -> F03
- "User is a postdoc in immunology working on T-cell signaling."
  -> F15
```

```
- "User is a computer scientist whose research is on protein
  structure prediction." -> F07
- "User is a social scientist at a think tank; publishes on
  migration." -> D6
- "User is a clinician-scientist in cardiology running a trial."
  -> F10
- "User works as a research scientist at Los Alamos National
  Laboratory." -> discipline_unclear
- "User works at the Max Planck Institute for Brain Research."
  (workplace only) -> discipline_unclear
- "User is a senior software engineer at a fintech; interested
  in physics." -> discipline_not_researcher
- "User is a master's student in data science writing a thesis
  on churn prediction." -> discipline_not_researcher
- "User is a UX researcher at a SaaS company." ->
  discipline_not_researcher
- "Prefers concise answers; uses British spelling." ->
  discipline_unclear
- "[transcript elided: no memory blocks present]" ->
  discipline_no_memory

Output your answer in <classification> tags. Write ONLY one code
from the list (F01-F34 or D1-D6), discipline_not_researcher,
discipline_unclear, or discipline_no_memory.

Assistant: <classification>
```

Memory-based role and career stage classifier

Model: Sonnet 5

```
Human: Consider the following memory record that claude.ai stores
about a user:

<memory>
{MEMORY}
</memory>

Classify, from the memory record above and nothing else, the user's
current role in or outside academia: whether they are a student
(doctoral or below), hold an academic position (faculty,
```

```
teaching-only, postdoc, or research staff), are employed outside
academia, have another stated occupation or status, or none of these
is established.

The memory record is what claude.ai has stored about this user
across past conversations; it is not a conversation.

ROLES.
- doctoral_student: enrolled in a doctoral program (PhD, DPhil,
  ScD, doctoral candidate, "doing a PhD") at a university or
  research institution. A doctoral student employed by a company or
  hosted at an institute is still doctoral_student.
- postdoc: a postdoctoral researcher, research fellow, or other
  early-career research position held after a doctorate and before a
  faculty or permanent position, at a university or research
  institution. A clinical fellow, or "fellow" without "research", is
  a physician in training and is not a postdoc; see
  employed_outside_academia.
- professor_faculty: professor of any rank (assistant,
  associate, full, emeritus), reader, principal investigator, group
  leader, or a lecturer, adjunct, or other faculty member at a
  university or research institution whom the memory does not
  describe as teaching only.
- teaching_faculty: a teaching position at a college or
  university with no research position stated: instructor, adjunct
  instructor, teaching professor, professor of practice,
  community-college faculty, or a lecturer or adjunct the memory
  describes only as teaching ("teaches undergraduate statistics",
  "teaches composition"). A professor or lecturer the memory does
  not describe as teaching only is professor_faculty, not
  teaching_faculty.
- academic_research_staff: employed at a university or research
  institution in a non-faculty research role and not enrolled as a
  student: staff scientist, research engineer, research assistant or
  associate, lab manager, research coordinator, technician.
  University administrators and other non-research university staff
  are role_unclear unless the memory states more.
- nondoctoral_student: enrolled below doctoral level: high
  school, bachelor's, master's, diploma, or a professional degree
  program (MD, JD, MBA, DDS, PharmD and the like), full- or
  part-time, with or without a thesis or a research-assistant job.
  Medical residents are not students; see employed_outside_academia.
- employed_outside_academia: employed anywhere other than a
  university or research institution: a company, a hospital or
  clinic without an academic research role, a school, a government
  body, a non-profit, self-employment, or freelance work.
```

```
  Physicians, residents, clinical fellows, and other clinicians
  count here unless the memory states a research title. Researchers
  at companies count here.
- other_role: retired, unemployed, homemaker, on leave,
  independent or unaffiliated researcher or writer, or a stated
  occupation that fits none of the roles above.
- role_unclear: the memory contains statements about the user
  but none that explicitly states a position, employment, or student
  status; or the statement fits more than one role and nothing
  settles it ("works at a university" with no position; "PhD" with
  no indication whether in progress or completed; "academic").
- role_no_memory: the record is the line "[transcript elided: no
  memory blocks present]", or contains no statement about the user
  at all (for example only a listing of file names).

EVIDENCE. Answer only from what the memory explicitly states about
the user's own position, employer, institution, or degree program.
Do not infer any of these from other facts in the record, however
suggestive; a role is assigned only when the memory states it, and
otherwise the answer is role_unclear. Do not infer a role from
topics, interests, activities, or tone: writing a dissertation does
not establish doctoral_student (master's theses are also called
dissertations), asking about tenure does not establish
professor_faculty, and analyzing lab data does not establish any
role. Statements about other people (partner, parent, child,
advisor, colleague, client) do not count. A university or research
institution means any university, college, academic medical center,
or public or non-profit research institute (national laboratories,
government research agencies, and independent research institutes
included). A company's research lab is not a research institution.

PRECEDENCE.
- If the memory states a student status below doctoral level
  together with a job, the student status decides, unless the memory
  says the study is part-time or secondary to the job.
- If the user holds several roles at the same time, the role the
  memory calls the main one decides ("mainly", "full-time", "day
  job", "primarily"). If it names no main role, an academic position
  (doctoral_student, postdoc, professor_faculty, teaching_faculty,
  academic_research_staff) decides over employment outside academia:
  a doctoral enrollment alongside a job is doctoral_student, a
  physician who is also an adjunct professor is professor_faculty, a
  professor who also consults is professor_faculty. Two academic
  positions at once: the one named first decides.
- An emeritus or retired professor is professor_faculty. Retired
  from any other occupation is other_role.
```

```
EXAMPLES (memory statement -> answer):
- "User is a third-year PhD student in chemistry at ETH Zurich."
  -> doctoral_student
- "User is doing an industrial PhD with Siemens and TU Munich."
  -> doctoral_student
- "User is a postdoc in computational neuroscience at the Allen
  Institute." -> postdoc
- "User is an associate professor of economics; teaches graduate
  econometrics." -> professor_faculty
- "User is a lecturer in economics at the University of
  Manchester." -> professor_faculty
- "User is an adjunct instructor of composition at a community
  college." -> teaching_faculty
- "User teaches undergraduate statistics at a state university;
  no research mentioned." -> teaching_faculty
- "User is a practicing physician and adjunct professor at a
  medical school." -> professor_faculty
- "User is a staff scientist at Lawrence Berkeley National
  Laboratory." -> academic_research_staff
- "User works as a research assistant in a psychology lab after
  finishing their BA." -> academic_research_staff
- "User is a master's student in public health; works part-time
  at a clinic." -> nondoctoral_student
- "User is a second-year medical student." ->
  nondoctoral_student
- "User is a senior data scientist at a bank; studying for an
  MBA in the evenings." -> employed_outside_academia
- "User is a research scientist at Google DeepMind." ->
  employed_outside_academia
- "User is a resident in internal medicine at a university
  hospital." -> employed_outside_academia
- "User is a cardiology fellow at a teaching hospital." ->
  employed_outside_academia
- "User is a high-school physics teacher." ->
  employed_outside_academia
- "User is an independent scholar writing a book on medieval
  trade." -> other_role
- "User is a retired civil engineer." -> other_role
- "User has a PhD in physics." (no current role) -> role_unclear
- "User works at the University of Toronto." (no position) ->
  role_unclear
- "User was a postdoc; now works at a startup." (startup role
  unstated, but employment outside academia stated) ->
  employed_outside_academia
```

```
- "User frequently asks about grant deadlines and lab budgets."
  -> role_unclear
- "[transcript elided: no memory blocks present]" ->
  role_no_memory

Output your answer in <classification> tags. Write ONLY one of
doctoral_student, postdoc, professor_faculty, teaching_faculty,
academic_research_staff, nondoctoral_student,
employed_outside_academia, other_role, role_unclear, or
role_no_memory.

Assistant: <classification>
```

Supplementary classifiers

In addition to the classifiers introduced in this report, we use the following classifiers introduced in Appel et al. (2026) and Massenkoff et al. (2026):

1. AI autonomy (prompt text)
2. Human time to complete a task (prompt text)
3. Time to complete a task with AI help (prompt text)
4. Education needed to understand user prompt (prompt text)
5. Education needed to understand Claude's response (prompt text)
6. Artifacts produced (prompt text)

Validation of classifiers

When no ground truth is available, validation of any classifier is inherently difficult. Without a ground-truth mapping from session transcripts to scientific fields or tasks, we inspect the validity of our classifiers through multiple robustness exercises and quality metrics.

Validation on synthetic session transcripts

We assess the accuracy of our field and task classifier on 5,000 synthetic session transcripts for which we know the correct subfield and task label by

construction. Synthetic transcripts are generated as follows: for each of the 318 subfields we randomly sample about 16 tasks belonging to this subfield, drawn from its universal, domain-level, field-level, and subfield-level tasks. A first model call turns each of the 5,000 subfield-task combinations into a description of a concrete research situation. A second model call then converts the research situation into a synthetic session transcript without seeing the label. Finally, a third model call removes subfield mentions and task wording where present. The output is a synthetic session transcript with a subfield and task label.

Using the resulting synthetic session transcripts and labels, we then evaluate the accuracy of our classifier, which is 92% (domain), 83% (field), 74% (subfield), 69% (task), 63% (task × field), and 58% (task × subfield). For comparison, drawing a subfield and one of its tasks at random would be correct in 20% of cases for the domain, 3% for the field, 0.3% for the subfield, 0.5% for the task, and less than 0.1% for the two combinations. About half of the transcripts still echo some of the label's wording after the rewrite. Excluding them lowers accuracy to 64% (task) and 54% (task × subfield), while field and subfield accuracy barely change.

Comparison of results based on transcript-based classification and memory-based classification

Our main results measure use across scientific fields based on transcript classifications (i.e., based on the content of sessions). As an alternative, we could measure use based on the field of the researcher, which is typically not observed. For a subset of users, however, we can infer a field of research by having a classifier identify this information based on Claude's memory. This gives us two separate signals of the researcher's field, one based on the content of the session and one based on the memory.

As a validation check for our field classifier, we use a sample of Claude session transcripts where we have memory available. We then correlate the share of transcripts in a given field based on the transcript-based classifier with the share of transcripts in a given field based on the memory-based classifier. These shares do not need to show a correlation of 1, even if classifiers have 100% accuracy, because of interdisciplinary use: researchers in one field may conduct interdisciplinary research in another field, which would lead to

differences in the memory- and transcript-based field counts. We find a log-log correlation of 0.92.

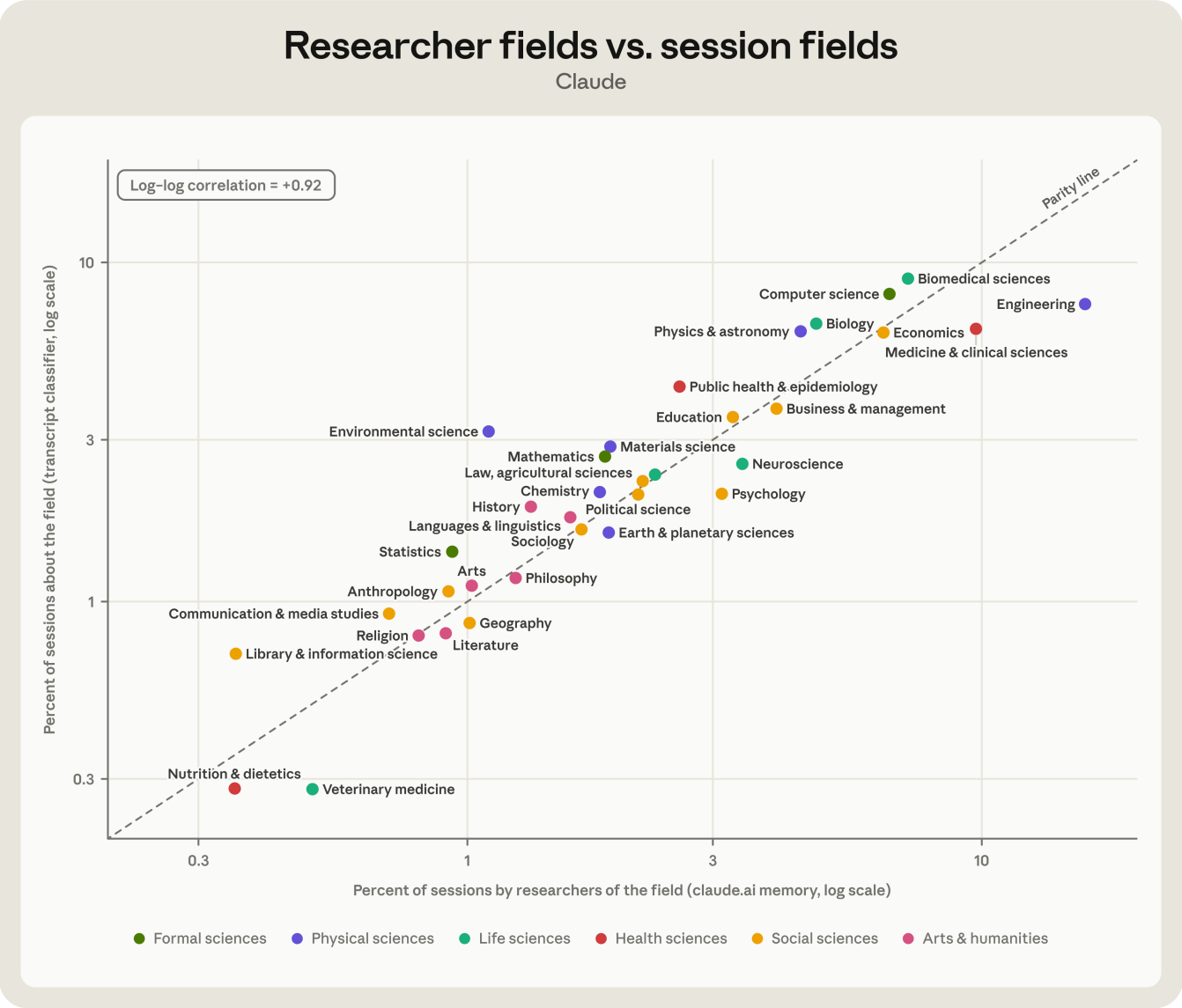


Figure A1. Researcher fields vs. session fields. Each dot is a field. The vertical axis represents the share of research sessions as classified by session content. The horizontal axis represents the share of research sessions based on the researcher's field, as recorded in their memory.

Robustness

We perform two additional robustness analyses with regard to our classifier results. We assess (1) the role of potential classification error in mapping sessions to SciNet tasks and fields, and (2) the role of sessions in which Claude contributes little to the researcher's goal.

First, we analyze the subsample of sessions for which our classifier maps a transcript to a SciNet task and field with high confidence. The classifier emits a confidence score from 1 to 5 for both the task and the field. We construct a high-confidence sample by restricting to sessions for which both scores are 4 or 5. This sample contains 62% of research usage of the main sample for Claude, and 60% for Claude Code. The main results hold in the high-confidence sample. The correlation of log usage per researcher between the high-confidence sample and the main sample, weighted by researchers, is 0.97 (Claude) and 0.98 (Claude Code) at the field level, and 0.91 (Claude) and 0.94 (Claude Code) at the subfield level. All field- or subfield-level correlations reported in the main text move by at most 0.04. Task category shares move by at most 1.5 percentage points. The use ratios rise or fall with each task characteristic just as in the main sample and, if anything, are slightly steeper.

Second, we analyze the subsample of sessions for which Claude's contribution is rated as at least co-producing (we label this the high-contribution sample). Our contribution classifier rates Claude's contribution to the identified task as none, providing form work (such as spell checks), advising, co-producing, or fully generating the task output. This sample contains 75% of research usage of the main sample for Claude, and 83% for Claude Code. The main patterns also hold in this high-contribution sample. The weighted correlation of log usage per researcher with the main sample is 0.99 at the field and subfield levels, on both products. All field- or subfield-level correlations reported in the main text move by at most 0.05. Task category shares move by at most 2.1 percentage points. The use ratios rise or fall with each task characteristic just as in the main sample.

Possible directions for future refinements of our task classifier

Our task classifier assigns each session to a single research task. Better understanding how AI integrates into extended workflows may require a more

fine-grained disaggregation of those sessions. In particular, future work could explore how sessions with Claude transition across research tasks, assisted by shared context and understanding. To support this, future versions of our classifier could assign multiple tasks to a session. We leave this extension for future research.

D. Researcher survey

We supplement our privacy-preserving analysis of Claude use with survey data on scientists. This complements our data on AI use with scientists' reports of how they use AI and their expectations of AI capability improvements. We survey researchers who applied for grants through one of several Anthropic token grant programs, from May 2025 through August 2026. These researchers are drawn primarily from biological and medical sciences and lean toward computational research. They are also expected to be stronger users of AI models in research, given their selection into the grant program. In September 2026, we contacted 1,813 researchers by email to complete the survey, offered as the baseline survey to initiate participation in a study of the effects of AI on scientific research. A total of 528 completed the survey.

E. Technical details on methods and analysis

Use ratios

We compute task-specific use ratios by comparing the volume of Claude use on a specific research task to the typical time researchers spend completing that task (based on SciNet). We define overuse versus underuse as the share of Claude sessions specific to a given task being larger versus smaller than the same task's share of typical research time.

Formally, let each researcher $i \in I_s$ in subfield $s \in S = \{1,...,318\}$ work on tasks $t \in T_s$ for $h_{t,s}$ hours per year, where T_s is the set of core research tasks in subfield

s and T is the union of these sets across subfields. Let the number of researchers in a subfield be denoted by N_s.

Typical time spent. The share of time researchers spend on a given task across scientific fields can be computed as

$$x_t = (\Sigma_s N_s \cdot h_{t,s}) / (\Sigma_t \Sigma_s N_s \cdot h_{t,s}).$$

Claude use. We compute the observed Claude use for task t for product $g \in$ {Claude, Claude Code} as the number of research sessions classified as corresponding to the task. The number of sessions is weighted to correct for differences in sampling rates across products and hours. Use shares are computed as the number of sessions with task t on product g as a share of the overall number of sessions on product g.

Claude use ratios. We define the use ratio for task t on product g as the use share of sessions on product g divided by its share of typical research time, x_t. A ratio above 1 means that the task receives more use than its share of research time implies, and a ratio below 1 means less.

Binning by task characteristics. A task's characteristic y can differ across the subfields in which a task is performed. We assign each task the mean of its characteristic over those subfields, weighted by the total time researchers spend on it in each subfield:

$$y_t = (\Sigma_s y_{t,s} \cdot h_{t,s} \cdot N_s) / (\Sigma_s N_s \cdot h_{t,s}).$$

We then sort tasks into bins by y_t. A bin's use ratio is the sum of its tasks' use shares divided by the sum of their time shares x_t.

Aggregation of task-level characteristics at the subfield and field level

For several analyses, we aggregate task-level characteristics, such as AI capability, to the subfield and field level. In both cases, each task is weighted by the time researchers spend on it. For subfield-level aggregation, we take the average of the characteristic across all tasks performed in the subfield, weighted by the hours a researcher in that subfield spends on each task per year. For field-level aggregation, we take the weighted average across all

task-subfield combinations in the field. The weight of each combination is the number of researchers in the subfield times the hours each of them spends on the task, so larger subfields in terms of researcher hours count for more.

AI share of research work by field

We estimate how much researcher time AI can take on in producing a paper, given how researchers report using AI today. We estimate the *AI share* as the research time freed step by step, which is then aggregated to tasks, subfields, and fields.

Research time by step. SciNet estimates the hours per year that researchers in each subfield spend on each task. As in our other analysis, we keep core research tasks aimed at generating new knowledge, such as data gathering, analysis, and writing, and drop non-research tasks such as teaching and refereeing. We split each task's hours across its steps using per-step hours estimated by a language model (Claude Opus 5.5). Each subfield has many tasks, and each task has many steps. A step's weight ω_j is its share of the subfield's core research hours: its task's share of those hours times the step's share of the task's hours. Weights therefore sum to one within a subfield.

Research time freed for each step. Estimates for researcher time freed draw on our survey of researchers. For each of four kinds of research work (ideation, data gathering, data analysis, and writing), respondents reported a distribution of how they used AI on their past few papers: none, assistive, shared, delegated, or autonomous. We treat these descriptors as saving 0%, 25%, 50%, 75%, and 100% of researcher time, respectively. Each step joins one of the four groups according to its task's SciNet category. We further assign design and development tasks to ideation, theoretical analysis to data analysis, and reading and knowledge acquisition to writing. The remaining research tasks, which include tasks such as coordinating with collaborators, are assigned the average of the four survey distributions.

Within each task group we assume that researchers use AI most on the steps AI performs best. We pool the steps of all fields, counting a step that appears in several fields once, and rank them by the AI capability score described in the previous sections. Steps the judge consistently rates impossible for AI, and the few steps without a rating, are tied at the bottom. We then assign levels of AI use from the bottom up, so that the share of steps at each level matches the

share of respondents reporting it. In ideation, for example, the lowest-rated 17% of steps save nothing, and the next 53% by the AI capability score save 25%. Physical and interpersonal steps are ranked like all others.

Researcher time freed by field. A subfield's time freed, its AI share, is $S=\sum_j \omega_j \sigma_j$, where ω_j is the time-weight for a given step when performed without AI and σ_j is the share of researcher time freed when using AI. This measure leaves out the cost of AI and the time AI takes, so it should be interpreted as human time freed for a given project. (In the research acceleration analysis, AI cost and time are included.) A field's value is the average over its subfields, weighted by their numbers of researchers. Across the 34 fields, time freed ranges from 27% (arts) to 52% (statistics), with a median of 38%. A field will see relatively high time freed if it has mostly non-physical steps and those steps have high AI capability scores.

Research acceleration estimates by field

To estimate potential accelerations for each research subfield and field, we incorporate the empirical findings into an ideas production function. The framework follows Jones (2026) with extensions to allow (i) different research steps to take different amounts of time and (ii) researchers and AI to work together on a given step. The next subsection develops the formal framework. The following subsection describes the empirical implementation.

Framework

The ideas production function. Let $\dot{Z}$ be a subfield's rate of research output, such as papers per year or a quality-weighted rate of papers, and let Z be the total prior output. Research is a set of steps j whose outputs r_j combine as

$$\dot{Z} = \zeta Z^{\phi} \left[\sum_j \alpha_j r_j^{\theta} \right]^{1/\theta}, \qquad \theta < 0 \tag{1}$$

where r_j is the output of research step j, $\zeta > 0$ is a scale parameter, ϕ determines how output depends on the cumulative stock of prior work Z, and the step weights $\alpha_j > 0$ sum to 1. Because $\theta < 0$, steps are complements, and a step that remains hard becomes a bottleneck. Steps can also be grouped into tasks. If steps combine within each task and tasks combine within a research project with the

same θ, then the result is again (1). We assume a common θ in the analysis, so none of the results below depend on how steps are grouped into tasks.

Producing each step. Each step's output can be produced in different ways: by researchers alone, by researchers and AI working together in some combination, or by AI alone if the step becomes fully automated. A step's output is measured in completed instances of that step, such as a dataset cleaned or a section drafted. Working alone, a researcher completes an instance of step j in T_{0j}^{L} hours. If L_j researcher-hours are applied, then output is

$$r_j = H_j L_j, \qquad H_j \equiv \frac{1}{T_{0j}^{L}} \tag{2}$$

where H_j is the researcher productivity at the step. With AI, an instance requires $T_{1j}^{L} < T_{0j}^{L}$ researcher-hours, which may include instructing, checking, and fixing the AI's work, plus T_{1j}^{AI} hours of AI running time. An AI-assisted instance may also have greater value, producing q_j instances of researcher output. Setting $q_j = 1$ means working with AI produces work of the same quality as researchers working alone. Combining L_j researcher-hours and K_j AI-hours, the step output is

$$r_j = q_j \min\left\{\frac{L_j}{T_{1j}^{L}},\ \frac{K_j}{T_{1j}^{AI}}\right\} \tag{3}$$

If AI fully automates the step, $T_{1j}^{L} = 0$ and output depends on AI time alone, $r_j = q_j K_j / T_{1j}^{AI}$.

The cost advantage. A researcher-hour costs w (the wage, common across steps) and an AI-hour costs μ. By (2), a unit of a step's output produced by researchers alone costs $w/H_j = w \cdot T_{0j}^{L}$. By (3), a unit produced with AI costs $p_j^{AI} = (w \cdot T_{1j}^{L} + \mu \cdot T_{1j}^{AI})/q_j$, which includes any researcher time the step still needs. The cost advantage of AI at step j is then

$$c_j \equiv \frac{w/H_j}{p_j^{AI}} \tag{4}$$

Researchers pick the cheaper method, so a human alone performs the step when $c_j < 1$. The cost for producing one unit of the step's output is $p_j = \min\{w/H_j, p_j^{AI}\}$.

The research problem. Given an R&D budget D, researchers choose step outputs to maximize their research output. Researchers solve

$$\max_{\{r_j\}} \left[\sum_j \alpha_j r_j^{\theta} \right]^{1/\theta} \quad \text{subject to} \quad \sum_j p_j r_j = D \tag{5}$$

Let $b \equiv \theta/(\theta - 1)$. Let ω_j be step j's share of the budget at the optimum.

Solving (5) gives

$$\omega_j = \frac{\alpha_j^{1-b} p_j^b}{\sum_k \alpha_k^{1-b} p_k^b} \tag{6}$$

Before AI, researchers perform every step, so $p_j = w/H_j$, and the common wage cancels:

$$\omega_j = \frac{\alpha_j^{1-b} H_j^{-b}}{\sum_k \alpha_k^{1-b} H_k^{-b}} \tag{7}$$

Therefore, a step takes a larger share of the initial budget, and of researcher time, when it matters more for output (higher α_j) or when researchers are less productive at it (lower H_j). The effect of H_j reflects complementarity. Because each step is needed, researchers cannot avoid steps where their productivity is low. Rather, a lower H_j (a harder step) raises the researcher time devoted to that step and consequently that step's share of the budget.

Time freed and AI running time. It is convenient to express the times in (2) and (3) relative to the researcher's original time. Define

$$\sigma_j \equiv 1 - \frac{T_{1j}^L}{T_{0j}^L}, \qquad \tau_j \equiv \frac{T_{1j}^{AI}}{T_{0j}^L} \tag{8}$$

Here σ_j is the share of researcher time freed, and τ_j is the AI's running time relative to the researcher's original time.

The AI share. The first of the two measures we report is the AI share: the fraction of baseline researcher time that AI could free. This is the time freed at each step, weighted by that step's share of researcher time,

$$S = \sum_j \omega_j \sigma_j \tag{9}$$

The AI share is a direct accounting of researcher time freed.

The research acceleration. The second measure, the research acceleration, is the resulting gain in research output for a fixed budget, after researchers rebalance their effort and pay for AI. We derive this acceleration result next.

Proposition 1 (acceleration). (i) At the optimum, $\dot{Z} = \zeta \cdot Z^{\phi} \cdot D/P$, where P is the cost of a unit of combined research output. (ii) If AI multiplies each step's unit cost by $\hat{p}_j$, then, holding Z fixed, the rate of research output ($\dot{Z}$) per R&D dollar increases by a factor

$$A = \left[\sum_j \omega_j \hat{p}_j^b\right]^{-1/b} \tag{10}$$

where ω_j are the budget shares before the change. (iii) Because researchers are paid the same wage w across research areas, ω_j is also step j's initial share of research time.

We call A the research acceleration. This is how many times as much research output can be produced per unit of research expenditure. This is the immediate acceleration, given the current state of knowledge, Z. Over the longer run, the knowledge stock Z will increase, which can further change the rate of progress. For example, if there is a "fishing out" effect, where the good or easy ideas are found earlier, then the research acceleration would decline with time.

Part (iii) of the proposition is what makes the theory measurable. Given the degree of complementarity θ, the proposition reduces the empirical estimation to two central objects. The first is how researchers spend their time across steps. These are the baseline time shares ω_j. These time shares absorb the step weights α_j and productivities H_j, so neither needs to be measured separately. The second is how much AI cuts the cost of each step, $\hat{p}_j$. Where AI is used, $\hat{p}_j = 1/c_j$, which depends on the researcher time AI frees, the time AI itself runs, the price of AI relative to the wage, and any gain in quality.

Cost changes in measurable terms. Substituting (2), (3), and (8) into (4), the cost advantage of AI at step j is

$$c_j = \frac{q_j}{(1-\sigma_j) + (\mu/w)\,\tau_j} \tag{11}$$

Researchers use AI where $c_j \geq 1$, so AI changes step j's unit cost by

$$\hat{p}_j = \min\left\{1,\ \frac{(1-\sigma_j)+(\mu/w)\ \tau_j}{q_j}\right\} \tag{12}$$

Substituting (12) into (10) gives the research acceleration in terms of measurable quantities: the baseline time shares ω_j, the share of researcher time freed σ_j, AI running time τ_j, the relative price of AI μ/w, and quality q_j. The empirical exercise therefore needs these quantities step by step, applying (12) to every step. Where researchers still do a step alone, (12) gives $\hat{p}_j = 1$, so that step's cost is unchanged.

The acceleration as cost per paper. By Proposition 1(i), the research acceleration has a direct interpretation. With AI, a paper of a given value costs $1/A$ as much to produce as before. Equivalently, a fixed research budget produces A times as much research output. This gain reflects a change in how researchers work. They shift effort toward the steps AI does not make cheaper, which become the bottlenecks. The ratio $1/A$ therefore compares the cost of the best mix of steps with AI to the best mix without it, where both mixes yield the same research output under (1).

What compute can do

Equation (11) shows three ways AI can raise the cost advantage. It can cut the researcher's remaining time $1 - \sigma_j$. It can cut the cost of its own running time $(\mu/w)\tau_j$. Or it can raise the quality q_j of the output. Each type of improvement has a different limit for how AI can accelerate research.

Faster or cheaper compute. Cheaper or faster compute lowers $(\mu/w)\tau_j$. Its gain is limited, however, while researchers still have work to do on the step. Specifically, recall that $c_j = q_j / [(1 - \sigma_j) + (\mu/w)\tau_j]$. As compute cost goes to zero, $c_j \to q_j/(1 - \sigma_j)$, a ceiling set by the share of researcher time $1 - \sigma_j$ that remains. That is, even with infinitely cheap compute, the cost advantage of AI may be limited, because the parts of a step that still rely on human researchers (e.g., physical elements or verification elements) remain expensive. Relatedly, compute matters more when researchers have little left to do. As $\sigma_j \to 1$ and with finite AI costs, $c_j \to q_j \cdot w / (\mu \cdot \tau_j)$. Then halving the price or running time of compute doubles c_j. Cheaper compute has the biggest effect for work AI does alone.

Quality. Quality raises c_j in proportion, with no ceiling. If $q_j > 1$, then AI-assisted output is simply better than what researchers produced at the same step, possibly through better ideas, wider search, or fewer errors. If researchers could not produce that output at any feasible cost, q_j is effectively infinite and so is c_j. This is the channel through which AI can change what research produces, in addition to what it costs.

Empirical implementation

Researcher time freed. We compute time freed for each step. Each step's share of researcher time freed is σ_j. The methodology for calculating time freed at the step level, and its aggregation to subfields and fields, is described earlier in this appendix.

AI running time and prices. No data source measures AI running time by step, so we make an assumption that $\tau_j = 0.2 \cdot \sigma_j$, meaning that it takes 0.2 AI-hours for each researcher-hour saved. A researcher-hour is assumed to cost $w = \$95$, which is taken as the median salary in the 2023 Survey of Doctorate Recipients ($155,000 in 2026 dollars) plus 28% benefits, over 2,080 hours. For AI, we assume one agent actively working for an hour costs $\mu = \$30$, an illustrative value based on public API list prices for Claude Opus 5.5. It is not a measured average cost of Claude use or of Anthropic's costs. We therefore have $\mu/w = 0.32$, and AI's running time costs are $(\mu/w) \cdot \tau_j = 0.063 \cdot \sigma_j$. In other words, AI costs about one-third as much per hour as a researcher, and it runs for only a fifth of the researcher time it saves.

Other calibration parameters. We set $\theta = -1$, so $b = ½$, which is the approximate median ideas production function parameter across fields from Ahmadpoor and Jones (2019). We also set $q_j = 1$, so work done with AI is of the same quality as researchers' work alone.

Subfield acceleration. We combine these calibration assumptions with Proposition 1 to compute the acceleration for each subfield. With these values, $c_j = 1 / (1 - 0.937 \cdot \sigma_j) \geq 1$. Notably, AI lowers the cost of every step where it frees researcher time. Equation (12) then gives $\hat{p}_j = (1 - \sigma_j) + 0.063 \cdot \sigma_j$. That is, a step keeps a share $1 - \sigma_j$ of its researcher time and adds AI time that costs $0.063 \cdot \sigma_j$ of its original researcher cost. The acceleration (10) becomes

$$A_s = \left[\sum_j \omega_j \left(1 - 0.937\,\sigma_j\right)^{1/2} \right]^{-2} \tag{13}$$

Field acceleration. A field's acceleration is the average of its subfields' accelerations, weighted by SciNet's researcher counts for each subfield.

Scenarios. We compute the AI share and the research acceleration under three scenarios. Scenario 1 reflects current AI capabilities, using the time freed σ_j estimated from our survey and AI capability scores. In Scenario 1, every step has σ_j equal to 0, 0.25, 0.5, 0.75, or 1, depending on its AI capability score and the levels of AI use reported in the survey. Scenarios 2 and 3 are hypothetical futures in which AI fully automates 75% or 100% of non-physical (cognitive and interpersonal) research steps, counted equally and starting with the steps with the highest AI capability scores. Automated steps have $\sigma_j = 1$ and $\tau_j = 0.2$. The added automation never applies to physical steps. Every other step, physical steps included, keeps its Scenario 1 values of σ_j and $\tau_j = 0.2 \cdot \sigma_j$.

We hold AI costs fixed across scenarios. In other words, AI extends what it can do, but it does not become cheaper at its research tasks on average. To the extent that AI inference costs fall per unit of time, or AI completes steps in less time, the constant cost assumption is conservative. On the other hand, harder tasks that AI does not yet perform, and more autonomous work on current tasks, may require more tokens, more expensive models, or longer running times, raising the cost of using AI. The overall AI cost trajectory in research will be a balance of these forces.

F. Additional results

Comparison of economic primitives across scientific and non-scientific usage

We compare sessions that the scientific work screener flags as scientific use with sessions on the same product that it does not. This allows us to assess how scientists' use of Claude differs from that of non-scientists in terms of AI autonomy, the time a task takes to complete for a human alone, the time AI can save in a given task, the years of education needed to understand the user's

prompt, and the level of education somebody would need to understand Claude's response. For this purpose, we use classifiers introduced in Appel et al. (2026).

We find that scientists' use is higher across virtually all of those dimensions, both on Claude and Claude Code. Scientists' usage provides AI with more autonomy, especially on Claude Code. The task time-to-completion is often more than twice as long for scientists' applications, irrespective of Claude product. Scientists use Claude for tasks where more time can be saved—with the exception of administrative tasks—though that difference is stronger on Claude than on Claude Code. Scientific use often requires four to seven more years of education to conduct the underlying task or understand Claude's response associated with it.

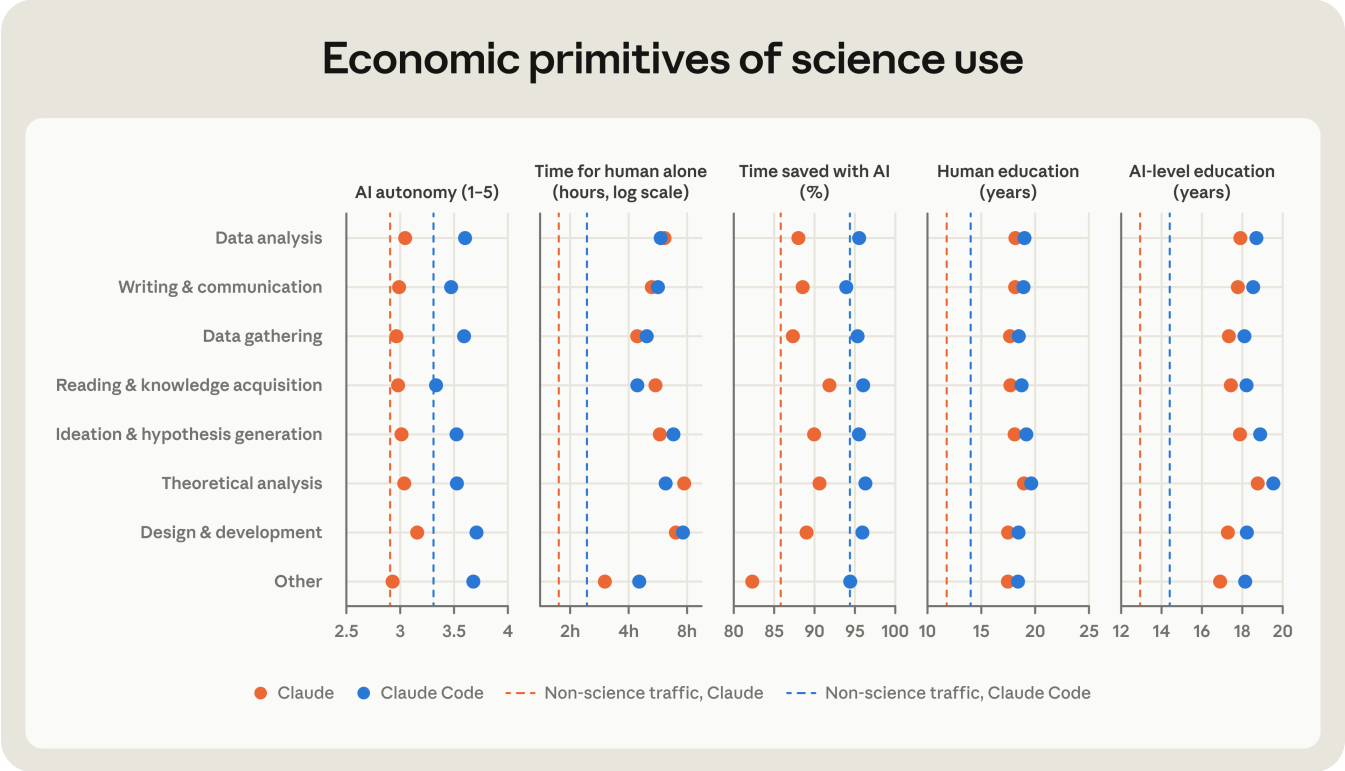


Figure A2. Economic primitives of science use. Each panel shows a different measure of task autonomy, size, time savings, and difficulty from our economic primitives framework introduced in Appel et al. (2026). Each dot is the average over research sessions in one task category. Orange dots correspond to Claude research sessions; blue dots correspond to Claude Code research sessions. The dashed lines correspond to all other use of the same product.

Claude use by career stage

For users with Claude's memory feature turned on, we use a classifier on Claude sessions to identify their career stage and role from what their memory states. As with our other classifiers, the memory-based classifiers in this study run in a privacy-preserving analysis pipeline. Users can modify their memory or turn it off entirely. Claude use among scientists differs by career stage. Doctoral students use Claude most for data analysis (34.4%), whereas professors use it most for writing and communication (44.0%). This is consistent with the division of research tasks across career stages, as documented by Robinson-Garcia et al. (2020), suggesting that AI usage tracks researcher needs. The largest relative difference in use is in reading and knowledge acquisition, where the share is about twice as high for doctoral students (8.8%) as it is for professors (4.9%). This could imply that doctoral students spend more time on this category of work, or that they have adopted AI more intensely for this category of work than their faculty counterparts. Differences in the remaining categories are small, at most 1.3 percentage points.

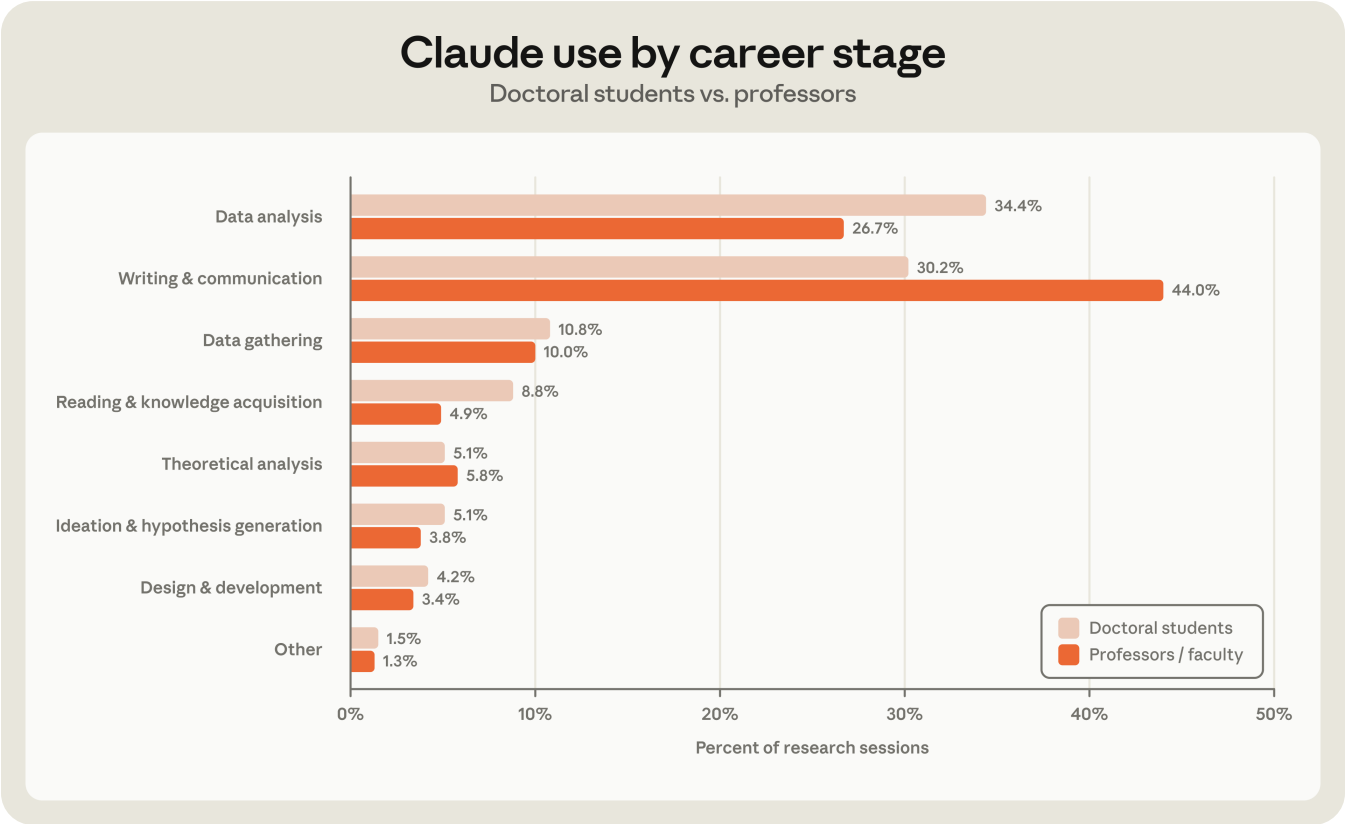


Figure A3. Claude use by career stage. Each row is one of the eight task categories. Light orange bars represent how doctoral students' Claude research is split across categories; dark orange bars represent the same for professors.

Determinants of Claude usage across fields

At the field level, Claude usage per researcher rises with theoretical AI capability, as it does across subfields (see Figure 11 in the main text). The correlations are 0.52 for Claude and 0.55 for Claude Code. Beyond this strong correlation, scientists in certain domains tend to use Claude more or less than AI capabilities would predict, with patterns similar to those in the subfield-level analysis in the main text.

Claude use and theoretical AI capability across fields

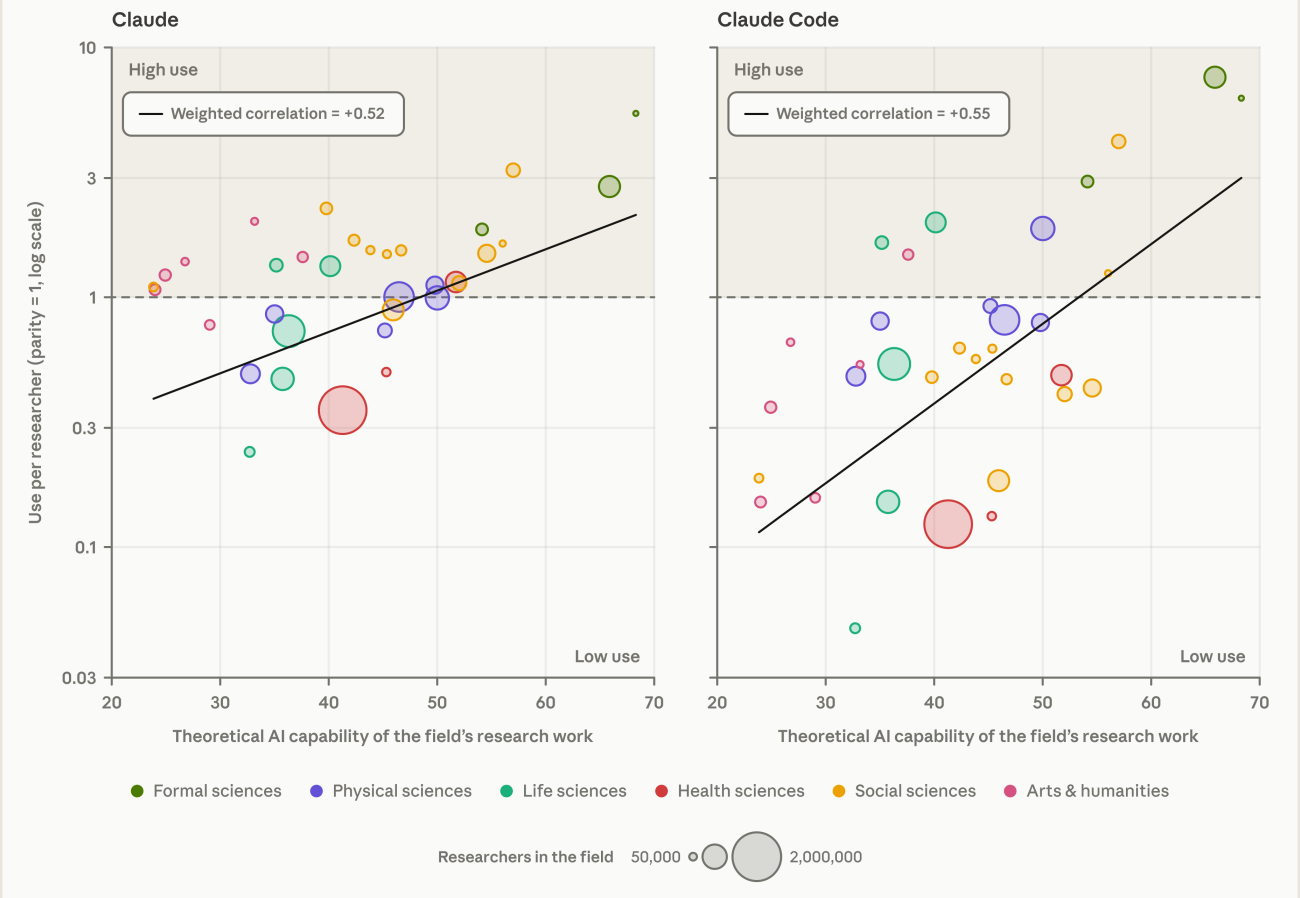


Figure A4. Claude use and theoretical AI capability across fields. This figure is the field-level counterpart of Figure 11. Both panels show each field as a dot, sized by the number of researchers in the field. The vertical axis shows the usage per researcher on a log scale. Usage per researcher is computed as the field's share of research sessions divided by the field's share of total researchers. The horizontal axis shows the field's theoretical AI capability score, computed by aggregating across all of the field's research tasks weighted by researcher hours. The black line shows the best fit, and each panel reports the correlation weighted by field size. The left panel shows results for Claude research sessions; the right panel for Claude Code research sessions.

Outputs of Claude Code usage

The artifacts produced by Claude Code tend to be more technical than those produced by Claude (Figure 7, main text). However, many of the broad patterns from Figure 7 hold, with written artifacts more common in the arts and humanities, social sciences, and health sciences, and technical artifacts more common in the life sciences, physical sciences, and formal sciences.

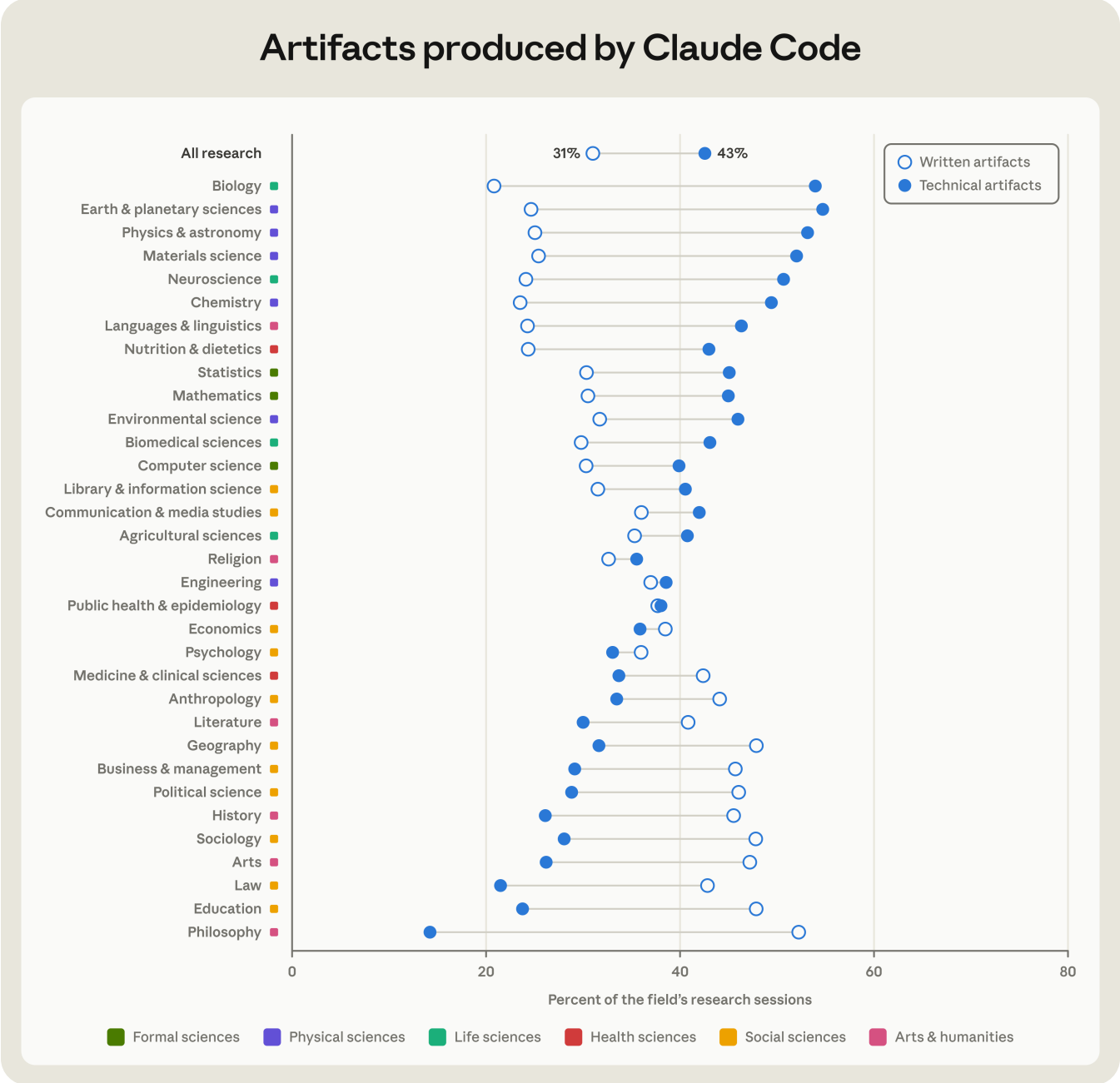


Figure A5. Artifacts produced by Claude Code. This figure computes the share of sessions on Claude Code that produce written artifacts and technical artifacts by field. Fields are ordered by the difference between the share of written artifacts and technical artifacts.

Claude Code usage vs. Claude usage

Fields whose tasks AI handles especially well make relatively more use of Claude Code than Claude. The left panel of Figure A6 shows this positive relationship across subfields. The right panel ranks fields by their Claude Code-to-Claude ratio: a field's share of Claude Code sessions divided by its share of Claude sessions. Computer science's share is 2.7x higher on Claude Code than on Claude, followed by physics and astronomy (1.9x) and mathematics (1.6x). By contrast, the shares of arts, anthropology, and veterinary medicine are more than five times as high on Claude as on Claude Code. Data analysis tasks may drive much of this difference. They are the type of research task where AI is rated most capable, and they account for nearly half of Claude Code sessions.

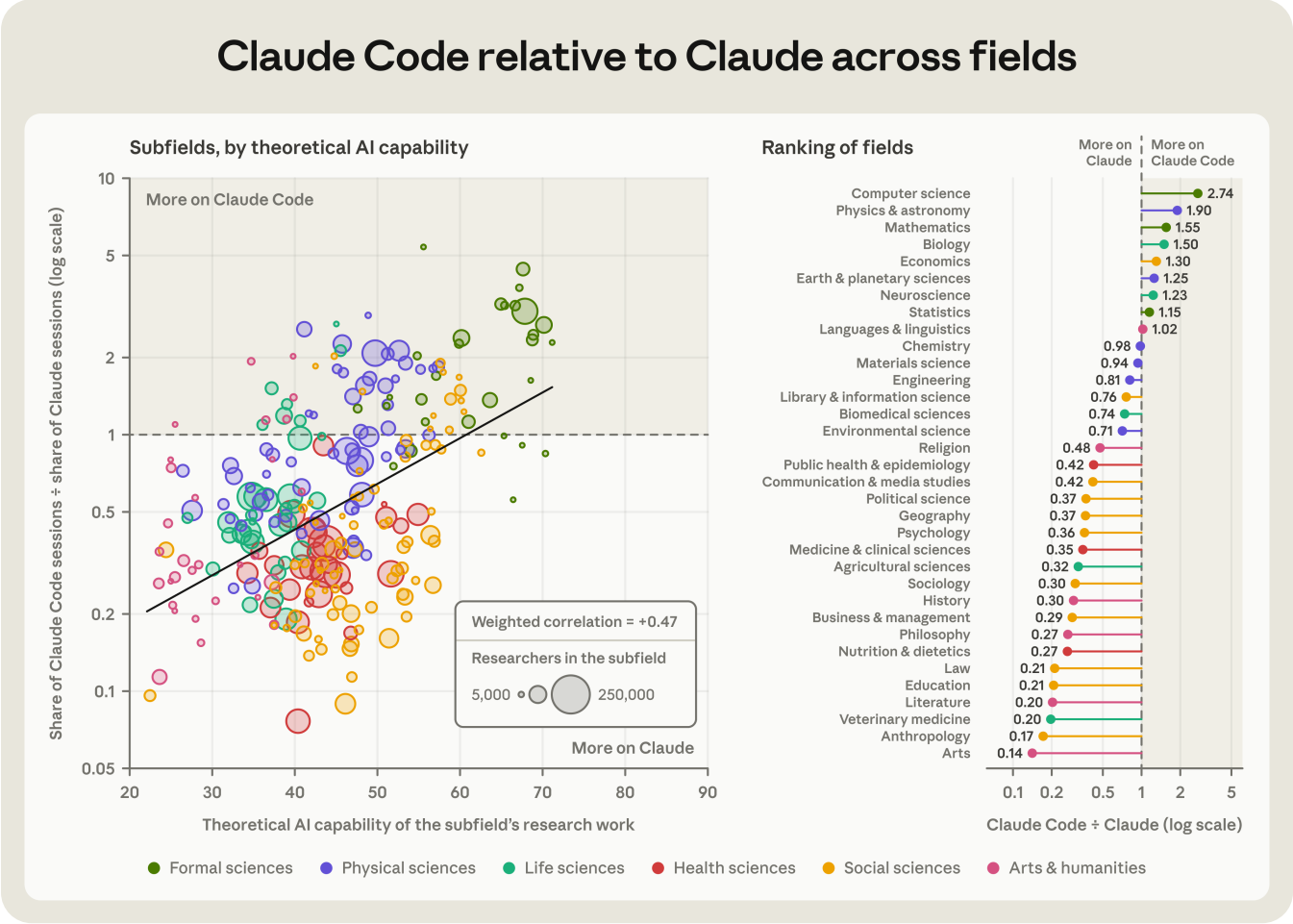


Figure A6. Claude Code relative to Claude across fields. The left panel shows each subfield as a dot, sized by the number of researchers in the subfield. The vertical axis shows the Claude Code-to-Claude ratio, defined as (subfield's share of research sessions on Claude Code) / (subfield's share of research sessions on Claude). The horizontal axis shows the subfield's AI capability score. The right panel ranks individual fields by their Claude Code use relative to Claude use.

Autonomy and theoretical AI capability across fields

As with subfields, fields with higher theoretical AI capability display more AI autonomy in their research sessions. This is true for sessions on both Claude and Claude Code.

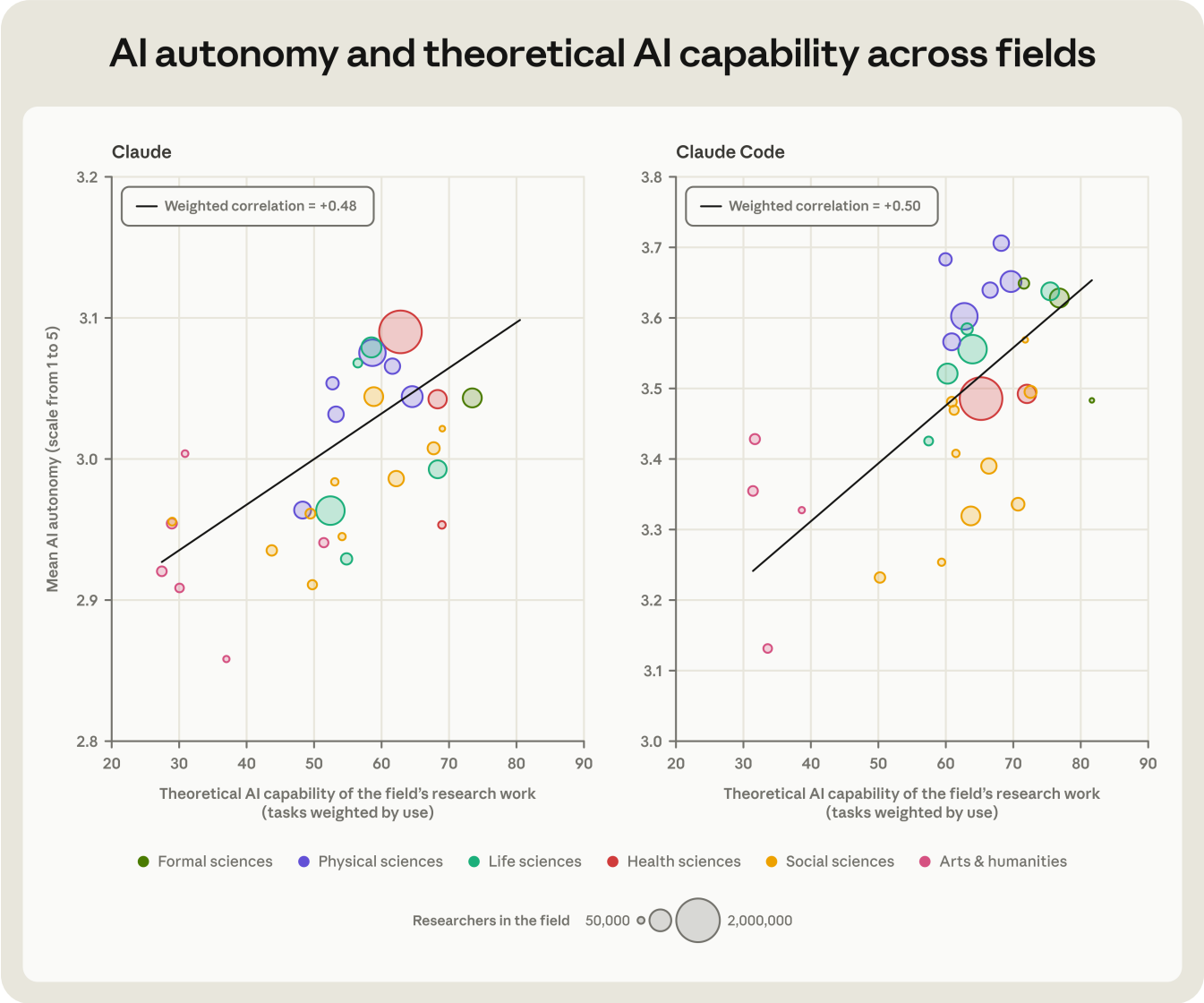


Figure A7. AI autonomy and theoretical AI capability across fields. This figure is the field-level counterpart of Figure 12. Both panels show each field as a dot, sized by the number of researchers in the field. The left panel focuses on Claude; the right panel focuses on Claude Code. The vertical axis shows the mean autonomy score for sessions in the field, weighted by session volume. The horizontal axis shows the mean theoretical capability of the field's tasks, also weighted by use. Tasks that are not covered by any session are excluded from the analysis, since they cannot be scored on autonomy. For legibility, outlier fields have been dropped, but they are included in the line of best fit and correlation calculation. Correlations are weighted by field size.