MIT FutureTech
Innovations that Shape the World

Massachusetts Institute of Technology
MIT FutureTech
32 Vassar Street, Office 386
Building 32
Cambridge, MA 02139-4307

Dear Selection Committee,

We are delighted to submit our recent working paper “Scientific Work” and the associated data project to be presented at the NBER Innovation Information Initiative (I3) Technical Working Group in December, 2026.

This working paper presents a detailed description of the Scientific Task Taxonomy. This taxonomy presents a new framework that maps scientific tasks across 232 subfields, including both commercial and non-commercial sectors. It is in essence an O*NET for science. We believe that this provides a valuable resource for the research community by allowing us to move the study of science to the task level. The first application of this data may naturally be to study the impact and adoption of AI in science, allowing the wealth of research on the task based production framework to be applied to science. However, given the flexibility of the framework developed, the data also reveal significant heterogeneity around how science is produced across academia versus industry, seniors/juniors, and in future iterations we plan to incorporate geographical differences between the U.S. and other regions.

We would benefit greatly from presenting the project at the working group. This project remains a work in progress as we continuously look to improve on the taxonomy built. The presentation would present a great opportunity to get feedback on both how we are building the taxonomy technically, and how it will be most useful to present and provide the data publicly.

We have already posted this working paper online, as well as a version 1 of the taxonomy here, as we are committed to developing this as a publicly available research resource. Our plan is to develop a data platform through which we can regularly update the taxonomy, which will allow us to capture how the tasks performed by scientists change over time.

Thank you for your consideration.

All the best,

Joseph Emmens
Assistant Professor | CUNEF Universidad
Affiliate Professor | MIT FutureTech

Scientific Work

Joseph Emmens[1,2,†], Jingnan Liu[3,2], Ana Trišović[2], and Neil Thompson[2,†]

[1]CUNEF Universidad
[2]MIT FutureTech
[3]University of Illinois Urbana-Champaign

July 2026

Abstract

Scientific discovery is a fundamental engine of human progress and economic growth, yet we lack detailed evidence on the nature of scientific work. This paper introduces the *Science Task Taxonomy*, a new framework that maps scientific work across 232 subfields, including both commercial and non-commercial sectors. In particular, we introduce three levels of task categories, from Level-1 tasks shared across the sciences to Level-3 tasks specific to individual subfields and sectors, with Level-2 providing an intermediate level of granularity. The taxonomy runs from 12 Level-1 task areas, such as "Analyze and model quantitative research data", through 114 Level-2 and 2,433 Level-3 areas, down to a pool of 208,202 representative tasks, such as "Develop cell-based assays to measure pharmacodynamic biomarkers". Using this taxonomy, we provide a systematic account of how the tasks performed by scientists differ from the tasks in the wider economy. We show that scientific work is more cognitive, requires fewer interpersonal interactions, is less manual across each of the physical, psychomotor and sensory dimensions. Science is also less codifiable, and the median scientific task takes over two hours longer to complete. We also find significant variation between scientific occupations along every dimension we measure. This paper moves the measurement of the idea production function to the task level, with important implications for our understanding of the effect of AI on science.

Keywords: Science, Research and Development, Task-based Modelling

[†]Corresponding authors: jemmens@mit.edu (J. Emmens); neil_t@mit.edu (N. Thompson). We are deeply indebted to both Peter Lin and Denis-Bogan Siminuic for their exceptional work and contributions as research assistants to the paper. Funding for this research was provided by Good Ventures / Coefficient Giving.

1 Introduction

The production of new scientific knowledge requires completing a wide range of tasks: developing a hypothesis and research strategy, managing personnel and data, running experiments, analysing results and writing them up, to name but a few. This bundle of tasks differs across scientific fields and from the work performed in the rest of the economy. Some scientists spend their time in wet labs or in the field, while others work with high-performance computing or provide patient care. This paper presents a new characterisation of scientific work, measures task characteristics across scientific fields, and compares those characteristics to the rest of the economy. In doing so, we extend the idea production function (Romer, 1990; Jones, 1995; Bloom et al., 2020) to the task level, with important implications for our understanding of the effect of AI on science.

We characterise each task along seven dimensions: its cognitive and interpersonal requirements; three components of manual work (physical, psychomotor, and sensory-perceptual demands); its codifiability, meaning the extent to which correct performance can be specified in advance as a sequence of replicable steps; and the human time it takes to complete. Scientific work differs from work in the rest of the economy on every one of them. Unsurprisingly, science makes greater cognitive demands, but it also requires fewer interpersonal interactions. Scientific work is less physically demanding and places lower demands on psychomotor and sensory-perceptual skills. It is also harder to codify.[1] Finally, scientific tasks take substantially longer to complete, with the median science task taking almost twice as long as the median task in the rest of the economy. Importantly, we find substantial variation across scientific fields on every dimension. Physical requirements, for instance, span more than an order of magnitude, from Mathematics at the bottom to Veterinary science at the top. The life and health sciences are markedly more physical than the relatively more interpersonal social sciences.

Documenting these findings requires fine-grained, cross-occupation data on scientific tasks. We therefore developed the *Science Task Taxonomy*, a hierarchy of the work performed by scientific occupations, which we define as a research subfield crossed with the commercial or non-commercial sector. The taxonomy is built on 3.7 million job postings for scientific research positions: positions whose primary responsibility is to generate new scientific knowledge through systematic investigation, experimentation, or theoretical work, or to directly support

[1] Though this gap is the smallest we measure and conceals the largest variation: the spread across scientific subfields is roughly nine times the science–economy gap itself.

others to do so. Job adverts have proved a valuable data source on changes in the supply of and demand for skills and technology at work (Deming and Kahn, 2018; Deming and Noray, 2020; Atalay et al., 2020; Acemoglu et al., 2022; Goldfarb, Taska, and Teodoridis, 2023). We process these postings through a novel pipeline that extracts 64.9 million atomic task instances and deduplicates them across subfields and sectors.[2] We then combine hierarchical clustering methods, embedding models and large language models (LLMs) to organise the scientific task instances into the task hierarchy.

The taxonomy is organised into three nested levels of increasing specificity: 12 Level-1 areas capturing the broad functions shared across the sciences, 114 Level-2 areas describing distinct bodies of work within them, and 2,433 Level-3 areas resolving that work into subfield-specific activities, grounded in a pool of 208,202 representative tasks, such as "Develop cell-based assays to measure pharmacodynamic biomarkers". For each subfield-sector pair we calculate a task importance weight for every task area in the hierarchy. The frequency with which a task is required by an occupation, captured by mentions in job adverts, provides a first approximation of its importance to scientific production in that occupation. We validate the taxonomy against two external sources of evidence. First, task overlap between subfield-sector pairs predicts observed job-to-job worker transitions in *Revelio Labs* career data. Second, task overlap between scientific subfields predicts knowledge flows through field-to-field citations.

Our taxonomy improves on existing characterisations of scientific work (O*NET, 2026; Althoff, 2026) in three respects. First, the O*NET occupation taxonomy describes science through a small number of broad occupations, which underspecifies tasks and omits many research subfields entirely. Under the occupation "Molecular and Cellular Biologists", O*NET lists *"Conduct research on cell organization and function, including mechanisms of gene expression, cellular bioinformatics, cell signalling, or cell differentiation"* as a single task. Even before splitting across sectors, our taxonomy separates that occupation into subfields which include Molecular Biology, Cell Biology, Genetics and Cellular and Molecular Neuroscience. The taxonomy then resolves that one task statement into many Level-3 task areas.

The distinction matters because the work being collapsed is not of one kind. Two of the Level-3 areas featured in our taxonomy correspond to phrases in the O*NET statement itself: "Analyse transcriptomic data using high-throughput sequencing" is the mechanisms of gene expression, and "Perform cell-based proliferation and differentiation assays" is cell differentiation. Both

[2]We apply the same pipeline to a random sample of non-science job adverts to build the comparison set. For this general economy comparison set, healthcare practitioners, sales and management occupations are the largest groups. The full occupational composition of the control sample is reported in Appendix Table A6.

areas are cognitively demanding, scoring 4.00 and 3.48 on a 0–4 scale but differ significantly along all other dimensions. The first requires much less physical effort (0.18 against 0.87) and far less psychomotor precision (1.03 against 2.72), is less codifiable (1.85 against 2.44), and takes roughly twice as long, a median of 9.0 hours against 4.7. This single O*NET statement assigns one characteristic profile to two areas that differ sharply. The taxonomy we present here allows us to to characterise the nature of scientific work more precisely.

Second, because our taxonomy is derived from job adverts, which carry posting-level microdata such as employer, sector, wages and required education, we can examine greater levels of heterogeneity, for instance by splitting across the commercial and non-commercial sectors. Third, job adverts describe the full set of tasks scientists perform, not only the work recorded in papers and lab notes but also management, budgeting and operations. These account for a fifth of the task areas we measure in non-commercial science and over a quarter in the commercial sector.[3]

Deriving a task taxonomy from job adverts also has its limitations. Defining the line separating science research from the broader STEM economy is difficult, and our science corpus therefore contains both false positives and false negatives. Our classifier tags about 0.85% of the universe of job adverts as science research positions, against an external benchmark of ∼2%.[4] We are therefore more likely to be under-recovering scientific positions than over-including them. Job adverts are also a selected sample of scientific positions, over-representing roles filled through public postings, which may capture junior positions better than senior ones.

Conceptualising the idea production function, and bringing it to data, has been a central concern of the economics of science and technology. This paper presents an empirical framework in which the production of science is described at the task level, in the spirit of task-based models of production (Autor, Levy, and Murnane, 2003; Acemoglu and Autor, 2011). Such models have been applied most prominently to automation and to the adoption

[3]These figures sum two Level-1 task areas: "Manage research projects and operational processes" (11.6% of Level-1 task coverage in non-commercial science and 13.6% in the commercial sector) and "Ensure quality, compliance, and regulatory documentation" (8.8% and 13.0% respectively). Shares are of total Level-1 coverage within each sector, and so measure the breadth of task areas a scientist is hired into rather than time allocation.

[4]Science-research jobs are roughly 2% of U.S. employment: 2,126,000 R&D employees in industry (NCSES BERD 2023, Table 50) plus 1,064,295 R&D personnel in higher education (NCSES HERD FY 2023, Table 78, headcount) against 161,037,000 employed persons aged 16+ (BLS CPS 2023 Annual Averages, Table 1). This is a lower bound as it omits federal, FFRDC, state government and nonprofit R&D performers, and both surveys apply size thresholds. It is also the share of employment, which may differ from the share of vacancies posted. However, it serves to give a rough approximation and better analyse the classifier performance. The NSF's ∼25% STEM share is far broader, covering all occupations in five scientist/engineer categories rather than only jobs that produce or directly support new research.

and impact of AI (Acemoglu and Restrepo, 2018; Eloundou et al., 2024; Lashkari et al., 2026).

Prior literature has shown how progress in AI capabilities and wider AI adoption could produce a permanent jump in the rate of new ideas (Cockburn, Henderson, and Stern, 2019; Besiroglu, Emery-Xu, and Thompson, 2024). That work has so far been largely theoretical (Agrawal, McHale, and Oettl, 2026) or field-specific (Gans and Mohnen, 2026; Hill and Stein, 2026). Jones (2025) argues that whether AI delivers transformational growth in R&D depends on the complementarity between scientific tasks and the strength of the bottleneck tasks that remain. Claims of this form cannot be taken to data without knowing what scientists actually do, and how those tasks differ across fields.

Beyond science, AI is being adopted across the broader economy, with a growing body of evidence on its effects on entry-level jobs, tasks, wages and employment (Brynjolfsson, Chandar, and Chen, 2026; Brynjolfsson, Li, and Raymond, 2025). Further more, Autor and Thompson (2025) show that whether adoption is favourable for wages and employment depends on which tasks within the bundle are automated. As such, interest in measuring AI use at the task level has grown accordingly (Handa et al., 2025; Chatterji et al., 2025; Appel et al., 2026).

Our measurements imply that science should not be expected to follow the same trajectory as the rest of the economy. Scientific tasks are more cognitive, a characteristic associated with greater language-model exposure (Eloundou et al., 2024), and non-routine cognitive work is the most represented task group in observed LLM use (Iscenko et al., 2026). Scientific tasks also take significantly longer to complete, and AI performance degrades as task horizons lengthen (Kwa et al., 2025; Mertens et al., 2026). Finally, numerous subfields remain substantially physical. We therefore hypothesise that the adoption and impact of AI across science will be determined as much by the requirements of the work as by model reasoning capabilities, thus potentially diverging from adoption patterns across the economy as a whole. Considering the cross-subfield differences, this may lead to the emergence of scientific fast lanes, where particular processes that can be fully automated quickly become commoditised. This paper builds the framework that would allow future work to test such a hypothesis.[5]

The rest of the paper is structured as follows. Section 2 outlines the method used to build the Science Task Taxonomy and a number of validation tests. Section 3 describes how we measure seven dimensions of task characteristics and the strategy to compare science to the

[5]This is the scope of the work presented in Codreanu et al. (2026) which provides early insights into the use of AI in Science at the task level.

general economy. Section 4 presents the results of this comparison and Section 5 develops an argument for what these differences mean for the production of science and technology.

2 The Science Task Taxonomy

2.1 Data

The science task taxonomy is built on the universe of U.S. job postings provided by Lightcast (2026), who aggregate and deduplicate job postings from a wide range of online sources. The full sample covers 487,557,017 unique job adverts covering the United States for positions posted between 2010 to 2025. Job posting and career data such as *Lightcast* over-represent high-skill and formal positions (Hershbein and Kahn, 2018), however as we are targetting scientific research positions, the standard representativeness concern is less significant.

2.2 Building the science task taxonomy

The Science Task Taxonomy is built through an iterative pipeline combining embedding-based machine learning with frontier LLMs. All LLM prompts are included for reference in Appendix Section B. The pipeline is summarised in Figure 1.

[Figure 1 Here]

2.2.1 Filtering science research positions

We first filter the science research positions from the full sample of U.S. job postings. Sending the full population of *Lightcast* jobs to an LLM for classification would be prohibitively expensive, and regex-hueristic based classifiers are inherently noisy. We therefore build the following pipeline through which we identify scientific research roles which optimises for both precision and efficiency.

Given that scientific research jobs are a small minority of all postings, in order to train a discerning classifier, we use a heuristic score to extract a stratified sample. We apply a fast regex-heuristic scorer that assigns each posting a science score from -6 to +12 given 8 signals. These signals are summarised in Appendix Table A1. The heuristic assigns each job advert to one of five bins ≤ -3, -2 to 0, 1 to 2, 4 to 6, and ≥ 8. An equal number of postings is drawn from each bin, alongside an additional unstratified random draw that preserves a representative

slice of the corpus. Stratifying the sample across the heuristic score distribution ensures that the LLM labels span unambiguous positives, unambiguous negatives, and the ambiguous border in roughly equal measure.

To train the model, we transform the job adverts into 1024-dim embedding vector representations using the Qwen3-Embedding-0.6B model.[6] In addition to stratifying the sample, we employ a method of active learning whereby we train an initial classifier on a train-val-test sample drawn as described above using the heuristic score. We then draw a second set of train-val-test job adverts for which the predicted science probabilities from the initial model are between 0.3 and 0.7. This supplementary sample enables us to target the uncertain boundary in this binary classification problem. We also draw an additional bucket of confidently-rejected, low-signal postings (heuristic score ≤ 0 and Pr(Science) ≤ 0.5). This supplementary sample targets the long tail of science positions that both the heuristic and the initial classifier reject in an effort to reduce the number of false negatives by finding long tail examples of science research positions. The classifier is then retrained on the full sample of LLM labels. In total, 72,768 job adverts are LLM labelled across the initial and supplementary draws, giving a total of 56,000 training, 5,590 validation, and 11,178 test labelled adverts.

We use a frontier LLM model (`gpt-5.4`) to produce labels on which we can train the science-position classifier. The full prompt is given in Appendix section B.1. We define a position as a science research role given the following definition:

> *"A scientific research job is one where the primary responsibility is to generate new scientific knowledge through systematic investigation, experimentation, or theoretical work. This includes academic, industry, non-profit, and government research roles."* ... *"Include research operations and support roles that directly enable scientific research (e.g. research project managers, lab technicians, equipment maintenance specialists, research data managers, research procurement/purchasing staff, grant managers)."* ... *"Exclude roles where 'research' refers primarily to market, user, policy, financial, or investment analysis, or where research activities are incidental to commercial, sales, marketing, or client-facing responsibilities."*

Table A2 compares three candidate classifiers at a threshold of 0.9, reporting both raw test-set

[6] Job adverts with a body shorter than 200 characters are dropped, since they carry too little text to meaningfully describe or to extract tasks from. This drops 6,662,713 adverts (1.4%), leaving 480,894,304 that are scored, embedded, and passed to the classifier.

metrics and metrics adjusted to our real-world science prior of 2%. We chose a conservative threshold of 0.9 to favour a clean rather than an exhaustive science corpus. The models are near-indistinguishable on the ROC-AUC which spans 0.977–0.978 and the average precision 0.935–0.940, selection is therefore determined by the adjusted-precision measure.[7]

The calibrated LinearSVC is chosen as the final classifier model. It produces a low false-positive rate (0.0025) which yields a prior-adjusted precision of 0.83 and the best prior-adjusted F_1 (0.70). Its prior-adjusted precision of 0.83 implies that, applied to the full corpus, roughly five of every six postings it labels as science are genuinely science, while its recall of 0.60 recovers about three in five true science postings. Finally, applying the calibrated classifier at the 0.9 threshold, to all 480.9M embedded postings, yields 4,105,313 science research positions, ~0.85% of the corpus. This falls below our prior of 2%, which assuming the prior is realistic, means that we are not capturing the entire population of science positions.

2.2.2 Classifying positions into subfields

We define a scientific occupation as a (subfield, sector) pair. Subfields are drawn from the OpenAlex taxonomy (Priem, Piwowar, and Orr, 2022), which allows us to link each subfield to OpenAlex's bibliometric record of research. The sector classification is assigned separately during the task extraction (Section 2.2.3). OpenAlex organises science hierarchically into broad fields, each subdivided into subfields, and we classify positions at both levels. We make three adjustments to the raw taxonomy before classifying. First, we merge five of the broad fields into two: Dentistry and Health Professions fold into Medicine, while Economics, Econometrics and Finance, Decision Sciences, and Business, Management and Accounting fold into a single synthetic field, Business, Economics and Decision Sciences. Second, we collapse the Veterinary field, whose OpenAlex subfields are too narrow for its job-ad population, so that its broad field and subfield coincide.[8] Third, we drop the eleven General [Field] subfields (e.g. General Engineering), which are catch-all buckets that add no discriminating detail. Together these leave 22 broad fields and 232 subfields which are described in Table 1.

[Table 1 Here]

[7] Since the test set is deliberately stratified to contain confidently science positions (24.0% science, against a ~2% corpus base rate), raw test precision overstates the classifier's precision. The relevant quantity is the prior-adjusted precision, which recomputes precision for a 2%-science population by holding each model's observed true- and false-positive rates fixed.

[8] These choices are reversible down the line and statistics on the task taxonomy can be reported on the original OpenAlex structure, however this was done to reduce the noise from small and highly similar fields.

The same frontier LLM (`gpt-5.4`) is then prompted to classify a training, validation and test sample at each of the broad and subfield levels. The prompts are included in the Appendix section B.2 and B.3. As in the science filter, the field classifier learns from GPT labels rather than the full corpus, and the sample sent for labelling is chosen to oversample from small fields and subfields. The classifier reuses the same 1024-dimensional Qwen3-Embedding-0.6B body embeddings computed in Section 2.2.1.

For broad-field labelling, the embedding representations of the 4.1M science positions are first clustered with FAISS k-means ($k = 500$). We again employ a two step procedure to improve coverage. An initial stratified 5,000 pilot sample is drawn across the k-cluster × Pr(science) quartile. After which a 50,000-posting additional sample up-weights fields that received fewer than 50 pilot labels, yielding roughly 55,000 broad labels. Subfield labelling repeats the design one level down: within each broad field the positions are re-clustered and sampled to produce a 15% / 85% pilot/boost split per subfield, giving roughly 170,000 subfield labels. In both passes the two-stage design ensures that rare fields and subfields accumulate enough labelled examples to train on, rather than being swamped by the few large fields.

Each classifier is a multi-layer perceptron (MLP) model that maps a job's 1024-dimensional body embedding through a single 512-unit hidden layer to one output per class. Because a posting may legitimately belong to more than one field, this is a multi-label problem. We compared three candidate classifiers, including a calibrated LinearSVC, a logistic regression and the MLP model (Table A3), trained on a 70/10/20 split of 38,500 / 5,500 / 11,000 labelled postings. The subfield layer comprises 21 per-broad MLP classifiers which use the same architecture (Veterinary is skipped, as it has no usable subfield taxonomy). Finally, rather than applying one global probability cut-off, we tune a per-class threshold on the validation set which rescues rare classes that a single threshold would suppress.

Given that this is a multi-label prediction problem the standard performance metric is the micro-F1, as opposed to using the accuracy statistic. At the tuned per-class thresholds, the broad-field MLP attains a micro-F_1 of 0.841 and a macro-F_1 of 0.812 on the held-out test set (Table A3; per-class metrics are reported in Appendix Table A4). The 21 subfield heads perform comparably one level down, with a mean micro-F_1 of 0.813 and mean macro-F_1 of 0.739 across fields, and per-field micro-F_1 ranging from 0.746 to 0.875. Performance is uneven across the taxonomy: the weakest broad classes are Chemical Engineering ($F_1 = 0.675$, on a recall of 0.614), Pharmacology, Toxicology and Pharmaceutics (0.734, on a precision of 0.690) and Environmental Science (0.764), against Nursing (0.921) and Medicine (0.888) at the

top, while the weakest subfield heads are Neuroscience (0.746), Biochemistry, Genetics and Molecular Biology (0.746) and Environmental Science (0.752), against Arts and Humanities (0.875) and Business, Economics and Decision Sciences (0.871) at the top.

Field assignment follows a hierarchical procedure. Instead of training a single classifier over all 232 subfields at once, we decompose it into a broad-field stage and a conditional subfield stage. We calibrate the classifier in order to produce true independent probabilities that can be combined to calculate the conditional expectation that a given job advert belongs a to a specific subfield. Each class is calibrated independently on the validation split. Calibration barely changes the classifications themselves (tuned micro-F_1 is 0.840 calibrated versus 0.841 raw). The broad classifier first assigns each position up to two OpenAlex fields. Each position is then routed to the subfield classifier belonging to its predicted broad field, and that classifier chooses only among that field's own child subfields; by construction the assigned subfield is therefore always a child of the assigned field. Each of the 21 subfield classifiers is trained and applied only within its broad field.

The two stages compose into a joint field–subfield probability that factorises as $P(\text{broad}) \times P(\text{subfield} \mid \text{broad})$. A position assigned a single broad field takes its top two subfields from that field's classifier, whereas a position spanning two broad fields takes the top subfield from each.[9] Positions whose top broad field is Unclassified—research-like roles that map onto no recognised OpenAlex field—are dropped, leaving 3,820,744 classified job adverts.[10]

Rather than treat each position as a full member of its assigned field–subfield cells, we weight membership by the classifier's calibrated confidence. For a position j, with vector embedding x_a and a field–subfield pair (f, s), the weight is

$$w\big(a, (f, s)\big) = Pr(s \mid f, x_a) \times Pr(f \mid x_a).$$

These weights are not renormalised to sum to one: the broad probabilities are independent draws and renormalising would impose a single-field assumption and therefore remove genuine dual-field membership, or hide cases in which the model is not confident. Membership is capped at two pairs per position. The primary (argmax) pair is always retained, and a second pair is added only if its class probability clears that class's tuned threshold.

[9] Job adverts classified into the broad field Veterinary skip the second stage, and the broad label is carried down as the subfield.

[10] Unclassified is fit as its own class in the broad classifier, so the model learns to predict it rather than being forced into a spurious field. This is effectively a second, field-based science gate on top of the binary filter from Section 2.2.1: a posting must be not only a research role but one locatable within the scientific field taxonomy.

2.2.3 Extracting task instances

Each classified science advert is sent once to `gpt-5-mini` with reasoning effort `low`, with its body truncated to 12,000 characters.[11] A single call performs four jobs; a further "Is this job a science research position" filter to remove additional false positives, job title standardisation, the task extraction, and sector classification (the full prompt is reproduced in Appendix B.4).

We extract each task instance as an atomic statement with a fixed structure following the template used by O*NET (2026), ACTION + OBJECT + [ENABLER] + [CONTEXT]: a base-form verb (e.g., *analyse*, *design*) and its direct object. We optionally include information on how the task is performed (tools or methods, using. . .) and, when specified, the setting (where, when, or for whom). Enablers and contexts are included only when supported by the advert. A task is extracted only if the position will personally perform it and it is explicitly stated (responsible for. . . , duties include. . .). The LLM is instructed to not included additional information inferred from its training data. We exclude requirements, qualifications, skills (knowledge of. . . , ability to. . . , preferred. . .), mission statements, and organisational goals (advance patient care, further the mission of. . .). Composite duties are decomposed into atomic tasks. Tasks are split on multiple actions ("design and analyse experiments" becomes two tasks), coordinated object lists, and do-X-to-do-Y purpose constructions. Shared enablers and contexts are propagated to each resulting task, and exact duplicates are removed. Table 2 reports descriptive statistics for the extracted task instances.

[Table 2 Here]

Sector is assigned in the same call. To do so we supply the hiring companies name, removing recruitment firms, which is the strongest signal. The model first verifies the listed company and then classifies that employer into one of five sectors: academia, government, non-profit, industry, or unclassified. Ex-post we group the academia, government and non-profit sectors into on non-commercial sector to compare against the commercial (industry) sector.

2.2.4 Organising the hierarchy

The 64.9 million extracted task instances are not 64.9 million distinct tasks. Two adverts writing "analyse experimental data" and "analyse experiment data" describe the same work. We therefore build the taxonomy in three steps: reduce the task instances to a set of distinct

[11] This truncation only bites for 1.81% of job adverts and is included to avoid one off large LLM costs. In addition, this is likely not to cut many tasks since many job adverts dedicate a long section at the end to generic text, for example application instructions and diversity commitments.

"representative tasks", discover the hierarchy over those representative tasks by clustering them, and use a frontier LLM to name and improve the structure that this clustering produces. Table A5 reports the counts at each step. This is a data-driven process that combines the compute power of clustering high-dimensional embedding representations of task instances with the reasoning capabilities of frontier LLMs.

The 64.9 million instances reduce to 20.7 million unique task statements. We embed each task statement using the Qwen3-Embedding-8B model, a larger model than the one used for advert bodies since task statements are short. We then find its nearest neighbours, using the cosine similarity, to then merge statements into clusters above a similarity threshold of 0.72.[12] The LLM writes one atomic label per cluster, and that label is the representative task. These labels are written in the same style as in the underlying task extraction prompt. Two filters then apply. First a frequency bar: a representative task must appear in at least 30 adverts to enter the taxonomy. This is the largest cut in the pipeline, and it leaves 255,618 representative tasks that cover 35.4 million raw task instances, or 83% of those mapped to a cluster. Second, the LLM performs a second check for statements that are not well-defined work tasks e.g. hollow obligations or role tautologies ("be responsible for assigned duties") and recruitment instructions. This drops a further 33,652 (13.2%), leaving 221,966 representative tasks.

We then cluster the representative tasks in two passes, reducing the embedding space with a UMAP projection and clustering with HDBSCAN at each pass. A broad pass over all representative tasks gives the Level-2 task areas. Sub-clustering within each Level-2 area gives the Level-3 areas. HDBSCAN does not assign a cluster to outlier representative tasks, and the broad pass leaves 79,753 of these tasks unassigned. These pass through three stages in an attempt to find a place for them in the taxonomy. First, by soft-assignment to the nearest existing centroid above a threshold cosine similarity floor, then LLM assignment to existing areas only. Finally, we recluster those tasks that remain within each area. Tasks that remain unassigned to a cluster, after the three steps outlined, are dropped from the taxonomy (8,291, or 3.9%), leaving 213,874 placed in a Level-2 and Level-3 area.

Clustering provides data driven partitions of the task space. Every cluster at both levels is named by the LLM from a sample of its members. The clustering process over embeddings captures a range of connections that a domain expert, or frontier LLM, would not. Therefore

[12] We included a number of guardrails to representative tasks that pass this threshold. An edge survives only if the two statements rank each other in their top-k neighbours and the edge is within a margin of each node's best neighbour. Without those guards clusters chain transitively: A merges with B and B with C even though A and C are unrelated. This leads to many tasks collapsing into a few giant components.

we allow the LLM to adjudicate the hierarchy proposed. The LLM is permitted to merge duplicate areas, move Level-3 areas to alternative Level-2 areas, and split mixed areas into finer grained descriptions. Third, the Level-2 areas are grouped into Level-1 areas and the resulting top layer is named and audited. Finally, a row-level pass judges every representative task against the path it sits on, reassigning those that fit another area better.

A representative task can sit on the correct branch and still carry no information beyond the label of the area it occupies. The row-level pass therefore returns one of four verdicts for each task. Tasks that add detail beyond their Level-3 label are retained as written. Tasks that merely restate their Level-3 label are flagged as *generic* and remain under that area. Tasks that are only as specific as the Level-2 label are also flagged generic and re-homed directly to the Level-2 area, so that they no longer count towards the Level-3 area they had been placed under. Tasks that belong on another branch are reassigned, and judged again under their new labels. In total, 97,523 of the 210,669 task–area placements (46.3%) are flagged as generic restatements. Being flagged however does not mean that they are deleted, generic tasks remain in the taxonomy and count towards the task weights defined in Section 2.2.5, but they are excluded wherever individual task statements are listed or matched. The same redundancy is repaired at the node level: a Level-3 area whose label merely restates its Level-2 parent is dissolved, and its representative tasks re-map directly to the Level 2. A Level-2 area left with a single labelled child has that child dissolved into it on the same grounds, so no labelled edge in the final tree merely restates the edge above it.

After consulting the taxonomy we intentionally define two additional Level-1 task areas, "Manage research data, integrity, and quality" and "Conceptualise, design, and propose research". This is the most heavy handed change we make to the taxonomy as we propose the name, a definition, example areas, and explicit exclusions, but the assignment is data-driven. The LLM decides from the task content which Level-2 areas belong in each target, and which Level-3 sub-areas to move onto the new path.

The final taxonomy has 12 Level-1 areas, 114 Level-2 areas and 2,433 Level-3 areas over a pool of 208,202 representative tasks. The task hierarchy is a tree, and task areas within it differ in the level of specificity with which they describe the work performed. The taxonomy is flexible to this and allows branches of different lengths, as illustrated in Figure 2.

[Figure 2 Here]

2.2.5 Task weights

Each subfield–sector pair carries a weight in $[0, 1]$ on every task area and representative task of the hierarchy. This weight measures how frequently that task is required by that scientific occupation. Weights are defined by union coverage. Denote o as a subfield–sector pair, $\mathcal{A}_o$ as the set of job adverts classified into o, and $\mathcal{T}(\tau)$ as the set of representative tasks descended from node τ. The weight is the share of that pair's adverts mentioning at least one task under the node,

$$\omega(o, \tau) \;=\; \frac{\left|\{\, a \in \mathcal{A}_o \;:\; \exists\, t \in \mathcal{T}(\tau) \text{ mentioned in } a \,\}\right|}{\left|\mathcal{A}_o\right|}.$$

A weight of 1 means every advert in the pair mentions at least one task in that area. An advert that mentions several representative tasks under a node is counted only once, therefore the weights do not sum to one across the areas of a given level. In that sense, they measure the breadth of task areas a scientist performs, and do not represent shares of time allocated.

Weights at higher levels are computed from the same advert–task mapping. A Level-1 weight is the union over every representative task descended from that Level-1 area. This method assigns weights monotonically as you move up the tree, so a parent's weight is always at least that of its largest child. At the representative task level the two coincide: the union coverage of a single representative task is exactly the share of the pair's adverts that mention it.

Generic tasks, those flagged as restating the label of the area they sit on, still count towards the node weights. When a task restates its Level-3 label it stays where it is and counts towards all Level-1, -2 and -3 weight calculations: an advert stating "analyse collected data" is genuine evidence that the pair performs work in that Level-3 area. Discarding these tasks would understate coverage for precisely those areas whose work is most often described in broad terms. Where a task restates only the broader Level-2 label, carrying none of what distinguishes the Level-3 it was placed under, it is re-homed to the Level-2 task area directly. The task then only contributes to the Level-2 and Level-1 union weight calculations.

Finally, membership in a subfield-sector pair is weighted by the confidence $w\big(a, (f, s)\big)$ with which the classifier assigned that job advert to the subfield, as defined in Section 2.2.2. The numerator becomes the sum of membership weights over distinct adverts mentioning that task area, and the denominator $W_o = \sum_{a \in \mathcal{A}_o} w_a$ the total membership weight of the pair. This reduces to the original expression above when every membership weight equals one.

2.2.6 Descriptive statistics

Table 3 summarises the structure of the taxonomy. Panel A counts the nodes at each level: 12 Level-1 areas capturing the broad functions shared across the sciences, 114 Level-2 areas describing distinct bodies of work within them, and 2,433 Level-3 areas resolving that work into subfield-specific activities, grounded in a pool of 208,202 representative tasks. Panel B shows how the tree branches outwards as you move from Level-1 down to the representative tasks. Each Level-1 area has between 3 and 19 Level-2 children (median 8), each Level-2 area a median of 17 Level-3 children (up to 82), and each Level-3 area a median of 27 representative tasks, with a mean of 55 and a maximum of 5,482. Three Level-2 areas have no Level-3 child and map directly to representative tasks, which is why the unconditional minimum in the Level-2 to Level-3 row is zero.

[Table 3 Here]

Panel C reports the coverage of each task area across the global hierarchy, which is the share of all science adverts mentioning at least one task under it, pooled across subfield–sector pairs. Coverage falls steeply as you move down the hierarchy, from a mean of 31.3% at Level 1 to 5.3% at Level 2 and 0.20% at Level 3. The largest Level-1 area, "Manage research projects and operational processes", appears in 53% of adverts, while the median Level-3 area appears in 0.07%, roughly 2,300 adverts. The hierarchy begins by outlining broad task areas, common to almost every scientific occupation, and moves down to activities specific to a small number of subfield-sector pairs.

Table 4 lists all 12 Level-1 areas in order of coverage, each with its three highest-coverage Level-2 children and the number of Level-2, Level-3 task areas and representative tasks beneath it. This provides an aggregate view of the taxonomy.[13]

[Table 4 Here]

2.3 Validation

We validate the science task taxonomy using two external sources of evidence.

[13]The full Level-1, Level-2 and Level-3 taxonomy, with task weights for every subfield–sector pair, is included in the data repository here.

2.3.1 Job-to-job transitions

The first test asks whether the task bundle we assign to a scientific occupation (a subfield-sector pair) carries information about the skills and knowledge that occupation actually requires. If it does, two occupations whose weighted task sets overlap should be close substitutes for a worker, and workers should move between them more often.

Those histories come from *Revelio Labs*, which assembles individual career records as a chronological sequence of positions, each with an employer, a job title and (in some cases) a textual description. We apply a similar pipeline as the one build for the Lightcast data to job observations. Position descriptions pass the a binary science filter and are then embedded with the same Qwen3-Embedding-0.6B model. The job positions are then classified by the existing broad and subfield classifiers, so a Revelio position is assigned a subfield exactly as a Lightcast advert is. We classify job positions into the same commercial / non-commercial split using the employer, NAICS codes and keyword rules. We use an LLM label for the companies that remain unassigned. This leaves 2,368,581 position spells classified into subfield-sector pairs, held by 1,434,202 workers.

Transitions are consecutive positions within a worker's history. Positions are sorted by start date, and a move is counted when the gap between the end of one and the start of the next falls between -30 and $+92$ days, which admits a short overlap or a brief spell out of work but not a career break. That yields 360,431 moves made by 246,984 workers, of which 95,021 stay within the same occupation, leaving 265,410 moves between occupations. For an unordered pair of occupations (a, b) we form the symmetric transition rate

$$r_{ab} = \frac{n_{a\to b} + n_{b\to a}}{\text{moves out of } a + \text{moves out of } b},$$

which nets out the fact that large occupations send workers everywhere. Requiring at least ten observed transitions between them leaves 4,298 pairs, spanning 334 of the 449 occupations in the comparison universe.

Task overlap is measured on the same task-weight vectors used throughout the paper: for occupation a and task area g, the weight v_{ag} is the share of a's adverts mentioning at least one task in g. The overlap between two occupations is the Ružička (weighted-Jaccard) index $\sum_g \min(v_{ag}, v_{bg}) \,/\, \sum_g \max(v_{ag}, v_{bg})$, which is 0 for disjoint task profiles and 1 for identical ones, and which can be computed at any level of the tree.

Figure 3 plots binned mean overlap at the Level-3 task vector level against the rank of the transition rate, and the relationship is monotone and increasing, with a Spearman correlation of 0.46 over the 4,298 pairs. Partialling out the pair's combined advert count leaves $\rho = 0.39$ Because the same field classifier labels both the job postings and the résumé positions, subfield classification confusion could in principle inflate overlap and fabricate transitions. To reduce these concerns, we have replicated this analysis using the NAICS code classifications provided by Lightcast, instead of the subfield-sector pair occupation definition, and the relationship holds.

[Figure 3 Here]

Table 5 reads the same relationship as a regression at each level of the tree, with the transition rate entered as a percentile rank so that the coefficient is the fitted change in overlap between the least and the most mobile pair. It is positive and significant everywhere, and largest in the middle of the hierarchy: 0.214 at Level 2 against a mean overlap of 0.395, and 0.170 at Level 3 against a mean of 0.236. The two ends are weaker for opposite reasons. Level 1 has only twelve areas and almost every scientific occupation touches most of them, so overlap is both high (0.565 on average) and compressed before mobility can discriminate between pairs. Level 4 is the individual representative task, where two occupations share little by construction (mean overlap 0.082) and the measure is correspondingly noisy.

[Table 5 Here]

2.3.2 Tracing knowledge flows through citations

The second validation test doubles down on tracking the required knowledge implied by each weighted set of tasks assigned to a subfield pair. We collapse the (subfield, sector) pair task vector to the subfield level by counting the number of job adverts categorised under that subfield, across both sectors, that require task t. We then normalise by the same count across all tasks. Again, Figure A1 demonstrates that the overlap between two subfield task vectors is informative of the knowledge required to perform those occupations given that it is positively associated with the frequency with which papers in those scientific subfields cite each other.

3 Describing Task Characteristics

We describe tasks along a range of seven dimensions, and compare science against the rest of the economy along each of them. Each dimension is summarised and described across fields

and subfields in the Figures 9 to 15.

3.1 Cognitive, manual and interpersonal

In the spirit of Autor, Levy, and Murnane (2003), we score each task on how cognitive, manual and interpersonal it is, on a scale from 0 to 4. We score the tasks by prompting a frontier LLM, following the method developed in Autor et al. (Forthcoming).[14] As a quick visual check, refer to the set of Figures 9 through 15, which include the distribution across task scores and a set of example representative tasks.[15] The manual dimension we split further into 3 groups. Physical tasks require both static and explosive strength, as well as stamina and flexibility. Psychomotor tasks require control, precision, reaction times and dexterity. While sensory tasks require the ability to use both near, far and peripheral vision, detect sounds, colours and depth. The definitions for cognitive and each of the three manual task characteristics are taken from the Handbook of Human Abilities Fleishman and Reilly, 1992. We refer to these five dimensions as the task requirements.

Each task statement is scored by `gpt-5.4-mini`. The prompt scores each dimension on a five-point scale from not required (0) to essential or defining (4) in the context of typically being performed by a competent worker. The five dimensions are scored in a single call, so that the model holds the contrastive frame between them. Each task is scored three times and the mean of the three votes taken as its score. The prompt is given in Appendix B.5.

3.2 Codifiable

An important question around automation at the task level is whether it is targetted at routine tasks, those which make up regular and repetitive tasks, or non-routine tasks. This can be captured along a dimension of whether a task is codifiable.

We score codifiability in a separate prompt, so that the routine judgement is not anchored on the requirement scores. The prompt asks whether a competent worker who has never performed this particular task could carry it out correctly, guided only by written rules and instructions. The scale runs from fundamentally non-codifiable (0), where correct performance rests on

[14] Autor et al. (Forthcoming) validate LLM scoring against human judgement on similar dimensions. Across 100 example tasks, the LLM's classification of whether a task is AI-addressable—that is, not inherently physical—agreed with the human classification 98% of the time.

[15] This stage requires further rigorous validation, especially for the cardinal ordering, however these figures show that the ordinal ranking of the tasks is sensible.

tacit knowledge that cannot be fully written down, to fully codifiable (4), where explicit ordered rules cover essentially every situation typical performance presents. The prompt is deliberately anchored neither on frequency (a task performed daily can demand fresh judgement each time) nor on the level of difficulty. Scoring follows the same three-vote mean. The prompt is given in Appendix B.6. For full transparency, this is the least well-established of our seven measures. The five requirement dimensions inherit definitions and referents from the Handbook of Human Abilities (Fleishman and Reilly, 1992), and our duration estimates are calibrated against survey measurements. Further work is required to establish that LLMs can accurately score tasks along this dimension.

Codifiability is not a restatement of the other characteristics, but neither is it independent of them. Appendix Figure A5 presents the correlation between the codifiability score and each of the other task requirement dimensions, across the scored tasks. Codifiability is positively correlated with manual tasks. A task that demands psychomotor precision ($\rho = +0.42$), sensory discrimination (+0.35) or physical effort (+0.25) is more likely to be one whose method can be written down in advance. While cognitive requirement (-0.42) and interpersonal requirement (-0.35) are negatively correlated. Task length is the strongest single correlate ($\rho = -0.46$). The longer a task takes a human, the less of it can be specified in advance.

3.3 Human time length

Kwa et al. (2025) showed that AI model performance degrades over task horizons for computer science tasks, and (Mertens et al., 2026) established that this relationship holds in general across the economy. We extend this concept here to examine the length of time required to complete scientific tasks given it's role as an important predictor of AI model performance. Task duration is not scored on a scale but estimated in hours, allowing tasks to range from requiring a few minutes to multiple weeks. We estimate task length in the Science Task Taxonomy with an anchored few-shot prompt and then correct the estimates onto the human scale. The method developed here builds again on that presented in Autor et al. (Forthcoming).

`gpt-5.4-mini` is first prompted to produce an estimate for the total active working time needed to complete one instance of the task, for a worker with five to nine years of experience in the role, excluding calendar time, waiting and interruptions, within an allowed range of one minute to four weeks. Each prompt carries 100 anchor tasks whose durations were measured in a survey of the workers who perform them (Mertens et al., 2026). These tasks are O*NET

tasks and so we first implement a matching between the science task taxonomy and O*NET. We supply the 40 most similar O*NET tasks and their human estimates and 60 additional examples sampled from across the surveyed O*NET log-duration distribution. Two draws per task are combined into their geometric mean. The prompt is given in Appendix B.7. The raw estimates are then corrected by a log-log regression, $\log_{10} T^{\text{human}} = a + b \log_{10} T^{\text{LLM}}$, fitted on 1,454 calibration tasks that carry both a survey measurement and an estimate from the identical pipeline ($a = 0.21$, $b = 0.76$).

[Table 6 Here]

3.4 Comparison design

Every comparison in this section compares science to the rest of the economy, and the two comparison sets are built so that they differ in what the work is rather than in how it was written down or selected. To avoid building a full taxonomy for the general economy and applying the same pipeline developed here for scientific work, significant care is taken to ensure a fair comparison across domains.

The scorer only sees verbatim advert text. A representative task is an LLM-written label, and the label does not read like text from an advert as they are often more generic. We calculate each dimension score for the lowest level of the taxonomy tree and then aggregate up to higher levels. This is either the Level-3 task area, where one exists, or the Level-2 where it does not. In total, this leaves 2,513 areas to score along the 7 dimensions. Within each of these areas, we draw up to 20 of its associated representative tasks, and from each of those one raw statement at random from that representative task's own cluster. This leads to 43,472 raw task statements, a mean of 17 per Level-2 or -3 area. An area's score on a characteristic is the mean over the associated raw task instances.

The control set is drawn from the corpus of job adverts not classified as science research postings. We sample adverts at random across the corpus from among the classifier's clear negatives, defined as those with a predicted science probability below 0.3. These adverts are then put through the same task extraction prompt, discarding any advert the extractor itself then reports as science. Each contributing advert supplies one task, drawn at random from the tasks extracted from it, which leaves 9,826 raw non-science task instances. Sampling one task per advert neutralises the correlation between tasks in a single advert, which can be substantial. The non-science sample tasks are then cleaned identically to those within the

science task taxonomy: task instances with more than 150 characters are dropped, and we apply the same the non-task (generic statement) classifier.

Area scores aggregate to higher levels in the taxonomy through union coverage—the share of science adverts mentioning at least one task under a node. Level-3 task areas are averaged within their Level-2 parent, Level-2 within Level-1 and so on. We measure uncertainty throughout using 95% cluster bootstraps over 2,000 draws. The science set resamples the 2,513 areas and recomputes the union-coverage weights on every draw, and the control set resamples adverts.[16]

3.5 Interpreting the scores

The five requirement dimensions and codifiability are ordinal scales running from 0 to 4, so they carry no natural external unit: a task scoring 3 does not demand three times what a task scoring 1 demands. To fix ideas when reading the results, Table 7 gives the tasks that sit at each end of each scale. Examples are endpoints and are not representative of the typical task. Given that comparing across dimensions without natural units is challenging, we report standardised differences throughout the results, which express each gap in units of the standard deviation of the underlying task-level distribution. This can be compared across dimensions, alongside the raw levels for reference.

[Table 7 Here]

4 Results

Figure 7 plots the distribution of each of the seven dimensions across the Level-1 task areas. This serves as a second visual validation check. "Teach and train students and research staff" scores highest on interpersonal requirement (3.51), and "Conduct experimental and sample processing operations" scores highest on all three manual dimensions, physical (1.13), psychomotor (2.12) and sensory-perceptual (2.33), while scoring lowest on cognitive requirement (3.15). The ordering is equally sensible on the remaining dimensions: "Conceptualise, design, and

[16]The subfield and sector assignments, the task extractions, and the characteristic scores are generated by an LLM or a classifier trained on LLM labels rather than directly observed. Battaglia et al. (2025) show that treating such generated variables as data can produce biased estimates and invalid inference, and develop two remedies. Our cluster bootstrap resamples task areas and adverts and so reflects sampling variation, but not the uncertainty introduced by the generation step. Incorporating these corrections, which would require a validated subsample against which to characterise the generation error, is left to future work.

propose research" is one of the most cognitive (3.92) and among the least codifiable (1.36), whereas "Ensure quality, compliance, and regulatory documentation" is one of the most codifiable (2.40).

4.1 Science compared to general work

Science tasks differ from general-economy tasks on every dimension we measure. Scientific work is more cognitive (3.63 against 3.12, $d = +0.79$) and it takes longer: the median science task takes 4.6 hours against 2.4 in the control, a ratio of 1.93 and a standardised gap of +0.80 on log duration. Every manual dimension runs the other way: scientific tasks are less physical ($d = -0.64$) and less interpersonal ($d = -0.22$). They are also less codifiable, however the difference is smaller than for the other dimensions ($d = -0.14$).

[Figure 4 Here]

Figure 4 reports the levels and Figure 5 the standardised differences. The standardised differences allow us to compare across each dimension, including the task characteristics and human time estimates. Figure A5 shows how each of the seven characteristics are correlated within both the science task and control sets and Table 6 provides the underlying distributions.[17]

[Figure 5 Here]

4.2 Comparing cross scientific occupations

We now examine heterogeneity across scientific subfields and sectors, in comparison to the general economy. Figure A2 shows the distribution for the 22 broad scientific fields across the Level-1 task area descriptions. The Level-1 areas are ordered by their coefficient of variation across fields. There are some areas which are highly field specific, such as "Coordinate clinical study execution and participant operation" being concentrated in the Health fields. While others are broadly performed across science, such as communication and management tasks.

These cross-field differences are reflected in Figure 8 which provides the distribution across the seven dimensions by Level-1 task area and by field. Across fields the picture is one of a common cognitive core with variation concentrated in the manual dimensions. Panel B of

[17]The statistics in Table 6 are unweighted over the 43,472 scored task instances, whereas the means reported here and in Figures 4 and 5 weight areas by union coverage, so the two differ slightly (for example, 3.58 against 3.63 on the cognitive dimension).

Figures 9 to 15 ranks the 22 fields on each characteristic in turn, along with the spread across each of the associated field's subfields. Unsurprisingly, all of science is relatively cognitive with a range of scores from 2.95 to 3.83 against an advert-weighted all-science mean of 3.58, with Mathematics and Computer Science at the top and Veterinary at the bottom. There is however significant variation across manual task requirements. Psychomotor ranges 0.50 to 1.80 and the physical requirement from 0.04 to 1.14, a spread of more than an order of magnitude. Interpersonal requirement separates the fields almost as sharply (1.52 to 3.03), codifiability runs 1.54 to 2.45, and median task length runs from 1.8 hours in Veterinary to 6.8 in Computer Science. Within-field spread is also significant. For example, the medicine subfields spread across nearly the whole range on every characteristic, so the field mean conceals significant variation.

Figure 6 compares the distribution across Level-1 task areas for each of the commercial and non-commercial sectors. Commercial scientists spend time on different tasks than scientists within the same subfield but who are working in the non-commercial sector. This translates into differences in the task characteristic requirements of each sector. Part of any gap is composition. The share of a subfield's job adverts coming from the commercial sector ranges from 0 to 92%. Therefore, a sector effect can be a field effect in disguise. Figure A4 therefore reports the difference estimated within subfield, identified off the 217 of 232 subfields with scored tasks in both sectors. The manual dimensions survive that control. Commercial science is less psychomotor ($d = -0.19$), less sensory-perceptual (-0.16) and less physical (-0.15). Cognitive requirement is higher in commercial science but marginally so ($+0.11$), and interpersonal requirement, codifiability and task length do not differ once the subfield is held fixed.

5 Discussion

Scientific work is different to that performed across the general economy. First, science tasks are defined over longer periods of time. There is a growing body of research showing that AI performance degrades over task length (Kwa et al., 2025; Mertens et al., 2026), which could restrict the adoption of AI models due to worse performance.

However, scientific work is also, unsurprisingly, more cognitive and places less importance on interacting with both the physical and social world in which it sits. Combine this with the fact that science work is slightly less codifiable and therefore less routine. These features of

knowledge work may leave parts of scientific work more exposed to advances in language models, consistent with task-based assessments of potential LLM exposure (Eloundou et al., 2024). Iscenko et al., 2026 find that non-routine cognitive tasks are over-represented in LLM (Gemini) use. These results align with Tomlinson et al. (2025), who when analysing 200,000 Copilot conversations, find that AI-assisted work is concentrated in information tasks, which are also characteristic of science.

Our results show substantial variation in manual requirements across fields and subfields, which aggregate descriptions of science can obscure. We find that the life sciences require a significant share of manual work, especially when compared to the social and physical sciences. This speaks to the increased interest in self-driving labs in which machine learning, automated experimentation and robotics are being combined across the chemical and materials sciences (Abolhasani and Kumacheva, 2023).

While we are restricted by a lack of scientific task level AI performance data, we can conclude that the adoption and potential impact of AI across science will differ to the broader economy. More broadly, assistance with individual tasks need not translate directly into autonomous research: gains from faster analysis or drafting may depend on complementary progress in experiment execution, validation and coordination.

An important feature of scientific work is that the share of employment in these fields is small compared to the broader economy. It is particularly low in comparison to the net social gains from its output, given that scientific knowledge is a public good. As such it produces significant spillover effects. Therefore, in a world of AI automation the typical concerns around reduced employment in return for productivity gains are somewhat reduced, if those gains lead to a jump in the rate of new ideas.

The idea production function remains a debated object within the economics of science and technology. This paper provides the empirical foundation which allows us to understand how science is produced at the task level. To that end, the framework presented in this paper naturally enables the development of a task-based model of scientific knowledge production. Building on the task-based framework (Acemoglu and Restrepo, 2019), future research should connect the taxonomy to direct measures of human and AI performance, task costs and scientific outputs. Resources such as the Intelligence Task Ontology (Blagec et al., 2022) provide complementary measures of AI capabilities, which could be linked to our taxonomy to measure scientific exposure to AI. Therefore, allowing us to better model the impact of AI and design policy to support the continued pursuit of scientific discovery.

References

Abolhasani, **Milad** and **Eugenia Kumacheva** (2023). "The rise of self-driving labs in chemical and materials sciences". *Nature Synthesis* 2, pp. 483–492. DOI: 10.1038/s44160-022-00231-0.

Acemoglu, **Daron**, **David Autor**, **Jonathon Hazell**, and **Pascual Restrepo** (2022). "Artificial Intelligence and Jobs: Evidence from Online Vacancies". *Journal of Labor Economics* 40.S1, S293–S340. DOI: 10.1086/718327.

Acemoglu, **Daron** and **David H. Autor** (2011). "Skills, Tasks and Technologies: Implications for Employment and Earnings". *Handbook of Labor Economics*. Vol. 4B, pp. 1043–1171. DOI: 10.1016/S0169-7218(11)02410-5.

Acemoglu, **Daron** and **Pascual Restrepo** (2018). "The Race Between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment". *American Economic Review* 108.6, pp. 1488–1542. DOI: 10.1257/aer.20160696.

– (2019). "Automation and New Tasks: How Technology Displaces and Reinstates Labor". *Journal of Economic Perspectives* 33.2, pp. 3–30. DOI: 10.1257/jep.33.2.3.

Agrawal, **Ajay K.**, **John McHale**, and **Alexander Oettl** (2026). *AI in Science*. NBER Working Paper 34953. National Bureau of Economic Research. DOI: 10.3386/w34953.

Althoff, **Lukas** (2026). *SciNet: The Anatomy of Science*. https://github.com/lukasalthoff/scinet.

Appel, **Ruth**, **Maxim Massenkoff**, **Peter McCrory**, **Miles McCain**, **Ryan Heller**, **Tyler Neylon**, and **Alex Tamkin** (2026). *Anthropic Economic Index Report: Economic Primitives*. URL: https://www.anthropic.com/research/anthropic-economic-index-january-2026-report.

Atalay, **Enghin**, **Phai Phongthiengtham**, **Sebastian Sotelo**, and **Daniel Tannenbaum** (Apr. 2020). "The Evolution of Work in the United States". *American Economic Journal: Applied Economics* 12.2, pp. 1–34. DOI: 10.1257/app.20190070. URL: https://www.aeaweb.org/articles?id=10.1257/app.20190070.

Autor, **David**, **Lucy Hampton**, **Neil Thompson**, and **Can Yeşildere** (Forthcoming). "It's about time: Forecasting AI automation andexpertise demands from task durations". Working Paper.

Autor, **David H.**, **Frank Levy**, and **Richard J. Murnane** (2003). "The Skill Content of Recent Technological Change: An Empirical Exploration". *The Quarterly Journal of Economics* 118.4, pp. 1279–1333. DOI: 10.1162/003355303322552801.

Autor, **David H.** and **Neil Thompson** (2025). "Expertise". *Journal of the European Economic Association* 23.4, pp. 1203–1271. DOI: 10.1093/jeea/jvaf023.

Battaglia, **Laura**, **Timothy Christensen**, **Stephen Hansen**, and **Szymon Sacher** (2025). *Inference for Regression with Variables Generated by AI or Machine Learning*. Working Paper 2402.15585. arXiv. URL: https://arxiv.org/abs/2402.15585.

Besiroglu, **Tamay**, **Nicholas Emery-Xu**, and **Neil Thompson** (2024). "Economic Impacts of AI-Augmented R&D". *Research Policy* 53.7, p. 105037. DOI: 10.1016/j.respol.2024.105037.

Blagec, **Kathrin**, **Adriano Barbosa-Silva**, **Simon Ott**, and **Matthias Samwald** (2022). "A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks". *Scientific Data* 9.1, p. 322. DOI: https://doi.org/10.1038/s41597-022-01435-x.

Bloom, **Nicholas**, **Charles I. Jones**, **John Van Reenen**, and **Michael Webb** (2020). "Are Ideas Getting Harder to Find?" *American Economic Review* 110.4, pp. 1104–1144. DOI: 10.1257/aer.20180338.

Brynjolfsson, **Erik**, **Bharat Chandar**, and **Ruyu Chen** (2026). *Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence*. Tech. rep. Stanford Digital Economy Lab.

Brynjolfsson, **Erik**, **Danielle Li**, and **Lindsey Raymond** (2025). "Generative AI at Work". *The Quarterly Journal of Economics* 140.2, pp. 889–942. DOI: 10.1093/qje/qjae044.

Chatterji, **Aaron**, **Thomas Cunningham**, **David J. Deming**, **Zoe Hitzig**, **Christopher Ong**, **Carl Yan Shan**, and **Kevin Wadman** (2025). *How People Use ChatGPT*. NBER Working Paper 34255. National Bureau of Economic Research. DOI: 10.3386/w34255.

Cockburn, **Iain M.**, **Rebecca Henderson**, and **Scott Stern** (2019). "The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis". *The Economics of Artificial Intelligence: An Agenda*, pp. 115–148. DOI: 10.7208/chicago/9780226613475.003.0004.

Codreanu, **Mihai**, **Alex Imas**, **Juan Mateos-Garcia**, **Joseph Emmens**, **Evalyne Muiruri**, **Arthur Turrell**, **Julian Jacobs**, **Atoosa Kasirzadeh**, **Ana Trišović**, **Yiyuan Chen**, **Tanya Rodchenko**, **Catherine Pollard**, **Scott Strand**, **Daniel Rock**, **Zanna Iscenko**, **Fabien Curto Millet**, **Neil Thompson**, and **James Manyika** (2026). "AI in Science: Early Insights". Google Working Paper.

Deming, **David** and **Lisa B. Kahn** (2018). "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals". *Journal of Labor Economics* 36.S1, S337–S369. DOI: 10.1086/694106.

Deming, **David J** and **Kadeem Noray** (Nov. 2020). "Earnings Dynamics, Changing Job Skills, and STEM Careers". *The Quarterly Journal of Economics* 135.4, pp. 1965–2005. ISSN: 0033-5533. DOI: 10.1093/qje/qjaa021. URL: https://doi.org/10.1093/qje/qjaa021.

Eloundou, **Tyna**, **Sam Manning**, **Pamela Mishkin**, and **Daniel Rock** (2024). "GPTs are GPTs: Labor market impact potential of LLMs". *Science* 384.6702, pp. 1306–1308. DOI: 10.1126/science.adj0998.

Fleishman, **Edwin A.** and **Maureen E. Reilly** (1992). *Handbook of Human Abilities: Definitions, Measurements, and Job Task Requirements*.

Gans, **Joshua S.** and **Myra Mohnen** (2026). "How AI Impacts the Quest for Knowledge: Evidence from AlphaFold2". Working Paper.

Goldfarb, **Avi**, **Bledi Taska**, and **Florenta Teodoridis** (2023). "Could Machine Learning Be a General Purpose Technology? A Comparison of Emerging Technologies Using Data from Online Job Postings". *Research Policy* 52.1, p. 104653. DOI: 10.1016/j.respol.2022.104653.

Handa, **Kunal**, **Alex Tamkin**, **Miles McCain**, **Saffron Huang**, **Esin Durmus**, **Sarah Heck**, **Jared Mueller**, **Jerry Hong**, **Stuart Ritchie**, **Tim Belonax**, **Kevin K. Troy**, **Dario Amodei**, **Jared Kaplan**, **Jack Clark**, and **Deep Ganguli** (2025). *Which Economic Tasks Are Performed with AI? Evidence from Millions of Claude Conversations*. DOI: 10.48550/arXiv.2503.04761. arXiv: 2503.04761.

Hershbein, **Brad** and **Lisa B Kahn** (2018). "Do recessions accelerate routine-biased technological change? Evidence from vacancy postings". *American Economic Review* 108.7, pp. 1737–1772. DOI: 10.1257/aer.20161570.

Hill, **Ryan R.** and **Carolyn Stein** (2026). *How Artificial Intelligence Shapes Science: Evidence from AlphaFold*. NBER Working Paper 35143. National Bureau of Economic Research. DOI: 10.3386/w35143.

Iscenko, **Zanna**, **Scott Strand**, **Yiyuan Chen**, **Guillaume Aimard**, **Mihai Codreanu**, **Vivek Sampathkumar**, **Alex Imas**, **Julian Jacobs**, **Evalyne Muiruri**, **Juan Mateos-Garcia**, **Jia Jen Ng**, **Samirah Javed**, **Josh Martin**, **Omar Ajmeri**, **Denis Calin**, **Andrew Kim**, **Fabien Curto Millet**, and **James Manyika** (2026). *Google's AI & Economy ATLAS v1.0: Mapping Gemini Usage in the Economy*. DOI: 10.48550/arXiv.2608.00038. arXiv: 2608.00038.

Jones, **Benjamin F.** (2025). *Artificial Intelligence in Research and Development*. NBER Working Paper 34312. National Bureau of Economic Research. DOI: 10.3386/w34312.

Jones, **Charles I.** (1995). "R&D-Based Models of Economic Growth". *Journal of Political Economy* 103.4, pp. 759–784. DOI: 10.1086/262002.

Kwa, **Thomas**, **Ben West**, **Joel Becker**, **Amy Deng**, **Katharyn Garcia**, **Max Hasin**, **Sami Jawhar**, **Megan Kinniment**, **Nate Rush**, **Sydney Von Arx**, **Ryan Bloom**, **Thomas Broadley**, **Haoxing Du**, **Brian Goodrich**, **Nikola Jurkovic**, **Luke Harold Miles**, **Seraphina Nix**, **Tao Lin**, **Chris Painter**, **Neev Parikh**, **David Rein**, **Lucas Jun Koba Sato**, **Hjalmar Wijk**, **Daniel M. Ziegler**, **Elizabeth Barnes**, and **Lawrence Chan** (2025). *Measuring AI Ability to Complete Long Software Tasks*. DOI: 10.48550/arXiv.2503.14499. arXiv: 2503.14499.

Lashkari, **Danial**, **Li Wensu**, **Christina Qui**, and **Neil Thompson** (2026). "AI and Scale: A Quantitative Task-Based Theory of Automation". Working Paper.

Lightcast (2026). *Lightcast Data*. URL: https://lightcast.io.

Mertens, **Matthias**, **Adam Kuzee**, **Brittany S. Harris**, **Harry Lyu**, **Wensu Li**, **Jonathan Rosenfeld**, **Meiri Anto**, **Martin Fleming**, and **Neil Thompson** (2026). *Crashing Waves vs. Rising Tides: Findings on AI Automation from Thousands of Worker Evaluations of Labor Market Tasks*. DOI: 10.48550/arXiv.2604.01363. arXiv: 2604.01363.

National Center for O*NET Development (2026). *O*NET 30.3 Database*. URL: https://www.onetcenter.org/database.html.

Priem, **Jason**, **Heather Piwowar**, and **Richard Orr** (2022). *OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts*. DOI: 10.48550/arXiv.2205.01833. arXiv: 2205.01833.

Romer, **Paul M.** (1990). "Endogenous Technological Change". *Journal of Political Economy* 98.5, Part 2, S71–S102. DOI: 10.1086/261725.

Tomlinson, **Kiran**, **Sonia Jaffe**, **Will Wang**, **Scott Counts**, and **Siddharth Suri** (2025). *Working with AI: measuring the applicability of generative AI to occupations*. DOI: https://doi.org/10.48550/arXiv.2507.07935. arXiv: 2507.07935.

Tables and Figures

Table 1

Scientific fields covered

Broad field	# Subfields	Example subfields
Medicine	55	*Pharmacology; Genetics*
Social Sciences	21	*Public Administration; Education*
Engineering	15	*Biomedical Engineering; Industrial and Manufacturing Engineering*
Biochemistry, Genetics and Molecular Biology	14	*Molecular Biology; Biotechnology*
Business, Economics and Decision Sciences	14	*Economics and Econometrics; Finance*
Arts and Humanities	12	*History; Philosophy*
Computer Science	11	*Computer Science Applications; Artificial Intelligence*
Environmental Science	11	*Management, Monitoring, Policy and Law; Health, Toxicology and Mutagenesis*
Mathematics	10	*Statistics and Probability; Applied Mathematics*
Agricultural and Biological Sciences	10	*Plant Science; Animal Science and Zoology*
Physics and Astronomy	8	*Instrumentation; Statistical and Nonlinear Physics*
Earth and Planetary Sciences	8	*Atmospheric Science; Oceanography*
Neuroscience	8	*Cellular and Molecular Neuroscience; Neurology*
Materials Science	7	*Biomaterials; Polymers and Plastics*
Chemistry	6	*Analytical Chemistry; Organic Chemistry*
Chemical Engineering	6	*Bioengineering; Filtration and Separation*
Psychology	6	*Developmental and Educational Psychology; Applied Psychology*
Immunology and Microbiology	5	*Immunology; Microbiology*
Pharmacology, Toxicology and Pharmaceutics	4	*Pharmacology; Pharmaceutical Science*
Nursing	4	*Research and Theory; Leadership and Management*
Energy	4	*Renewable Energy, Sustainability and the Environment; Fuel Technology*
Veterinary	1	*Veterinary*

Notes: The 22 broad fields into which the 3.8 million classified science positions are classified, with the number of subfields per field and illustrative examples. The taxonomy spans 22 broad fields and 232 subfields in total; the counts above sum to 240 because eight subfields sit under more than one broad field. These fields are taken from the OpenAlex taxonomy. Five of the broad fields are combined into two: Dentistry and Health Professions fold into Medicine, while Economics, Econometrics and Finance, Decision Sciences, and Business, Management and Accounting fold into a single synthetic field, Business, Economics and Decision Sciences.

Table 2

Task extraction: descriptive statistics

Statistic	Value
Corpus	
Job adverts processed	3,820,740
confirmed science (retained)	3,736,980 (97.8%)
dropped by science filter	83,760 (2.2%)
Task instances extracted	64,899,085
Tasks per science advert	
Mean	17.4
Median [IQR]	15 [8, 24]
90th percentile	34
Adverts with no extractable task	1.9%
Task length	
Words: mean (median)	8.2 (7)
Words: 10th–90th percentile	3–15
Sector (share of science adverts)	
Industry	43.3%
Academia	37.9%
Non-profit	9.2%
Government	6.3%
Unclassified	3.2%

Notes: Counts are over the full corpus. "confirmed science" are the adverts the Stage-1 filter retained (the remainder are dropped and contribute no tasks). Tasks-per-advert and sector statistics are over confirmed-science adverts.

Table 3

THE TASK TAXONOMY: STRUCTURE AND WEIGHTS

Panel A: nodes at each level

Level 1	12
Level 2	114
Level 3	2,433
Level 4	210,542
distinct statements	208,202
non-suppressed	113,080

Panel B: number of children

	Parents	Children	Min	Median	Mean	Max
Level 1 → Level 2	12	114	3	8	9.5	19
Level 2 → Level 3 (all)	114	2,433	0	17	21.3	82
Level 2 → Level 3 | match exists	111	2,433	2	17	21.9	82
Level 3 → Level 4	2,433	134,239	1	27	55.2	5,482

Panel C: distribution of weights

	Nodes	Min	p25	Median	Mean	p75	Max
Level 1	12	12.97%	24.44%	28.25%	31.33%	40.14%	53.23%
Level 2	114	0.114%	2.14%	4.08%	5.31%	6.40%	41.36%
Level 3	2,433	0.0003%	0.035%	0.074%	0.202%	0.165%	23.05%
Level 4	210,542	0.0003%	0.0011%	0.0017%	0.0043%	0.0032%	4.59%
non-suppressed	113,080	0.0003%	0.0011%	0.0015%	0.0035%	0.0028%	1.07%

Notes: The taxonomy has three levels built on top of a set of representative task statements. Level-4 counts are distinct representative tasks (208,202 distinct statements). Suppressed entries are generic restatements of their parent labels, excluded from leaf task lists but retained in node weights. Panel B's Level-2 → Level-3 row is unconditional, so the three Level-2 areas with no Level-3 child enter at zero however, the second Level-2 → Level-3 row conditions on a match existing. Panel B's Level-3 → Level-4 row covers the 134,239 Level-4 entries that sit under a Level-3 node. Panel C reports the weight of a node: the share of the 3,174,773 job adverts with at least one assigned task that mention at least one task under that node. Weights are computed directly at each level as a union over distinct adverts, never summed or renormalised, so a single advert contributes to several nodes and column sums exceed one (Level 1 3.76, Level 2 6.05, Level 3 4.91).

TABLE 4

LEVEL-1 AREAS AND A SAMPLE OF THEIR LEVEL-2 CHILDREN

Area	Weight	L2	L3	L4
Manage research projects and operational processes	53.23%	19	451	38,190
Attend and coordinate meetings and conferences	15.03%		52	4,742
Manage projects and project plans	13.51%		78	5,274
Communicate with clients and stakeholders	10.41%		59	4,365
Ensure quality, compliance, and regulatory documentation	44.96%	12	351	25,861
Perform testing and validation	11.52%		82	5,086
Prepare regulatory submissions and documentation	11.32%		66	4,638
Maintain records and documentation	10.74%		44	3,048
Communicate research findings and stakeholder information	44.15%	3	76	15,710
Prepare research reports, publications, and briefings	41.36%		41	12,903
Conduct literature and information searches	6.45%		26	1,544
Prepare tables and figures	3.87%		9	1,263
Conduct experimental and sample processing operations	38.80%	16	345	29,956
Conduct experimental investigations	19.27%		53	6,288
Collect and process samples	11.33%		30	5,127
Perform laboratory facility cleaning and upkeep	8.01%		25	1,626
Analyze and model quantitative research data	33.60%	9	201	15,314
Perform statistical analysis and planning	24.62%		51	6,058
Analyze high-throughput experimental data	5.14%		59	2,191
Analyze and process imaging data	2.43%		22	985
Conceptualise, design, and propose research	29.62%	7	93	10,592
Research Strategy and Program Design	11.86%		51	4,056
Research Funding and Proposal Development	11.50%		8	2,239
Clinical Trial Protocol Design and Development	9.84%		17	2,297
Teach and train students and research staff	26.88%	6	128	10,075
Teach courses and curricula	8.11%		57	2,993
Supervise and mentor students and trainees	7.81%		15	1,285
Train and onboard personnel	7.76%		31	2,292
Coordinate clinical study execution and participant operations	26.84%	13	228	24,760
Clinical Trial Operations and Protocol Management	14.24%		33	6,428
Recruit and enroll research participants	8.93%		16	2,723
Conduct study site selection, initiation, and monitoring visits	6.37%		27	2,691
Manage research data, integrity, and quality	25.41%	14	238	13,605
Multi-Source Data Integration and Preparation	6.54%		34	1,889
Validate dataset integrity and accuracy	5.96%		28	1,626
Study Database Development, Quality, and Management	5.53%		29	1,947
Develop and troubleshoot research assays and software	21.55%	6	133	11,286
Troubleshoot processes and technical issues	11.22%		23	3,828
Develop and validate assays	6.05%		33	2,661
Research Software Engineering and Infrastructure	3.74%		33	2,427
Collect research data and participant information	18.02%	5	63	5,266
Collect research data	10.07%		25	1,992
Case Report Form Completion, Validation, and Annotation	5.56%		4	1,096
Clinical Research Data Collection and Management	4.18%		21	1,142
Develop products, prototypes, and process technologies	12.97%	4	126	9,927
Develop research methods and application technologies	8.53%		63	4,916
Design instruments and prototype systems	2.97%		24	2,160
Transfer processes and technologies	2.60%		22	1,605

Notes: All 12 Level-1 areas, ordered by total weight wthin the taxonomy across scienntific subfield-sector pairs. Each Level-1 area is described by its 3 highest-weight Level-2 children. Weight is the share of job adverts with at least one assigned task that mention at least one task under the node and is calculated as the union over distinct adverts computed at each level. Columns L2, L3 and L4 count nodes at each level descending from the row's area; L4 includes representative tasks with no Level-3 parent.

Table 5

Task overlap regressed on the job-to-job transition rate, by taxonomy level

	Level-1	Level-2	Level-3	Level-4
J2J transition rate (percentile rank)	0.152***	0.214***	0.170***	0.084***
	(0.015)	(0.014)	(0.012)	(0.009)
	[0.125, 0.181]	[0.189, 0.243]	[0.147, 0.193]	[0.067, 0.103]
Spearman ρ	0.331***	0.463***	0.462***	0.366***
R^2	0.109	0.219	0.218	0.145
Mean task overlap	0.565	0.395	0.236	0.082
Pairs	4,298	4,298	4,298	4,298

Notes: OLS of the weighted Jaccard (Ružička) task overlap between occupation (subfield × sector) pairs on the percentile rank of their symmetric job-to-job transition rate, so the coefficient is the fitted change in overlap from the least to the most mobile pair. Standard errors in parentheses and 95% intervals in brackets come from a dyadic cluster bootstrap over occupations (2,000 draws), the same procedure that draws the confidence band in Figure 3. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 6

Descriptive statistics for the seven task-requirement dimensions

	N	Mean	SD	Min	P25	Median	P75	Max
Panel A. Science tasks								
Cognitive	43,472	3.58	0.58	0.33	3.00	4.00	4.00	4.00
Psychomotor	43,472	1.00	0.89	0.00	0.33	1.00	1.33	4.00
Physical	43,472	0.33	0.61	0.00	0.00	0.00	0.33	4.00
Sensory-perceptual	43,472	1.72	0.65	0.00	1.00	1.67	2.00	4.00
Interpersonal	43,472	2.01	1.14	0.00	1.00	2.00	3.00	4.00
Codifiability	43,472	1.97	0.90	0.00	1.00	2.00	3.00	4.00
Task length (hours)	43,472	6.11	6.27	0.08	2.47	4.56	7.75	197.71
Panel B. General-economy control								
Cognitive	9,826	3.12	0.85	0.00	2.67	3.00	4.00	4.00
Psychomotor	9,826	1.23	0.98	0.00	0.67	1.00	2.00	4.00
Physical	9,826	0.70	1.01	0.00	0.00	0.00	1.33	4.00
Sensory-perceptual	9,826	1.83	0.73	0.00	1.33	1.67	2.00	4.00
Interpersonal	9,826	2.31	1.27	0.00	1.00	2.00	3.67	4.00
Codifiability	9,826	2.06	0.93	0.00	1.00	2.00	3.00	4.00
Task length (hours)	9,826	3.79	6.50	0.02	1.22	2.37	4.67	252.17

Notes: One observation per scored task over the sample of tasks used in Figures 9 – 15. The first six dimensions are means of three votes on the prompt's 0–4 rubric; task length is in hours and is not comparable to the rows above it.

Table 7

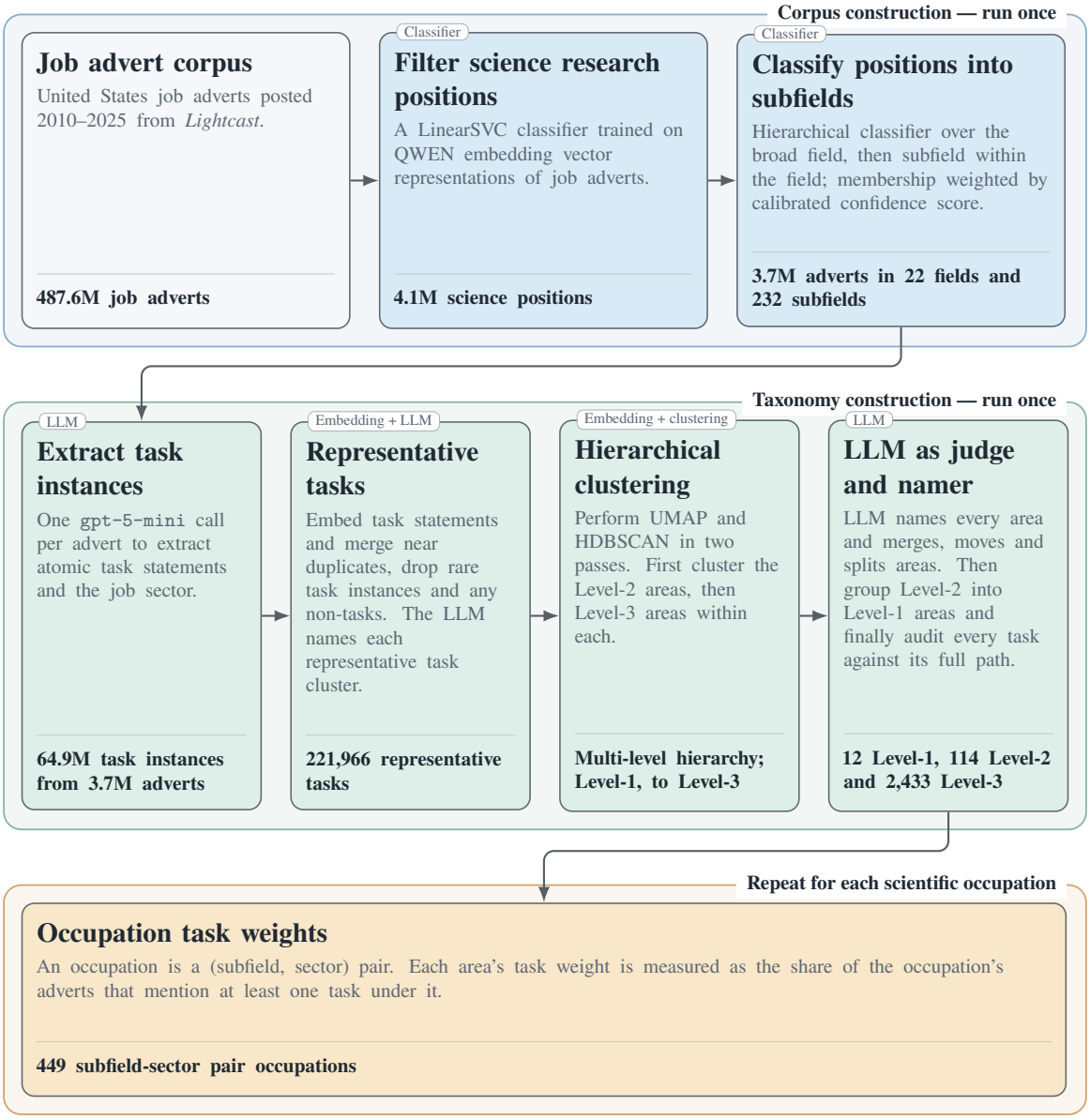
Anchoring the task characteristic scales

	Lowest scoring		*Highest scoring*	
Dimension	Example task	Score	Example task	Score
Cognitive	make copies at the Xerox machine	1.0	analyze literature and patent search results	4.0
Psychomotor	approve CRA travel expenses and timesheets	0.0	collect blood via submandibular vein	4.0
Physical	take detailed minutes and notes	0.0	move large and heavy equipment	4.0
Sensory-perceptual	provide guidance on randomization	0.0	perform blastocyst microinjection of embryos	4.0
Interpersonal	transcribe and code data	0.0	teach courses in research and statistics	4.0
Codifiability	generate patentable invention disclosures	0.0	record blood pressure measurements	4.0
Task length	take heart rate	0.11 h	execute multi-center clinical trials	197.7 h

Notes: Example task statements at each end of each scale. The first six dimensions are scored 0–4; task length is estimated in hours of active working time and is reported here in hours. Examples are endpoints and are not representative of the typical task: the unweighted all-science mean over scored tasks is 3.58 on cognitive, 1.00 on psychomotor, 0.33 on physical, 1.72 on sensory-perceptual, 2.01 on interpersonal and 1.97 on codifiability, with a median task length of 4.6 hours. No task scores 0 on the cognitive dimension, so the low anchor there is 1.0.

FIGURE 1

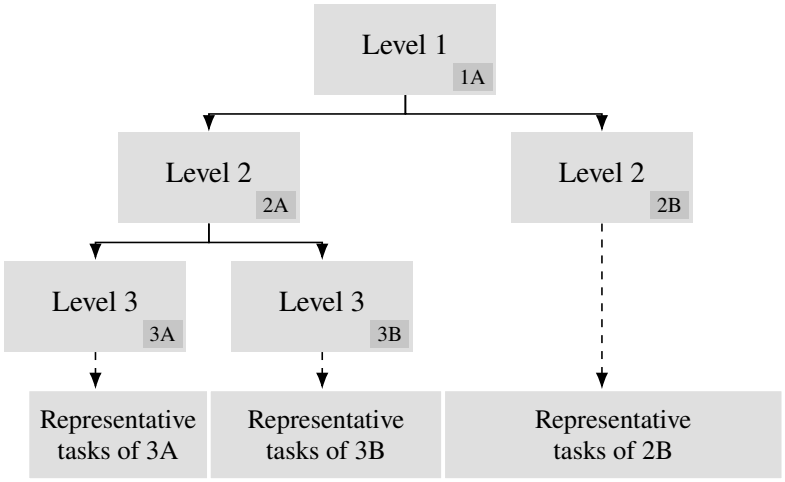
BUILDING THE SCIENCE TASK TAXONOMY: PIPELINE OVERVIEW



Notes: Each box is one step in building the taxonomy. The three shaded plates group the steps. First the corpus construction (Section 2.2.1), second the taxonomy construction (Section 2.2.3), and third the weighting that is repeated for every scientific occupation, a (subfield, sector) pair. The mechanism is listed where relevant; a trained classifier, embedding-based clustering, or a frontier LLM. Table A5 reports the full sequence of counts.

FIGURE 2

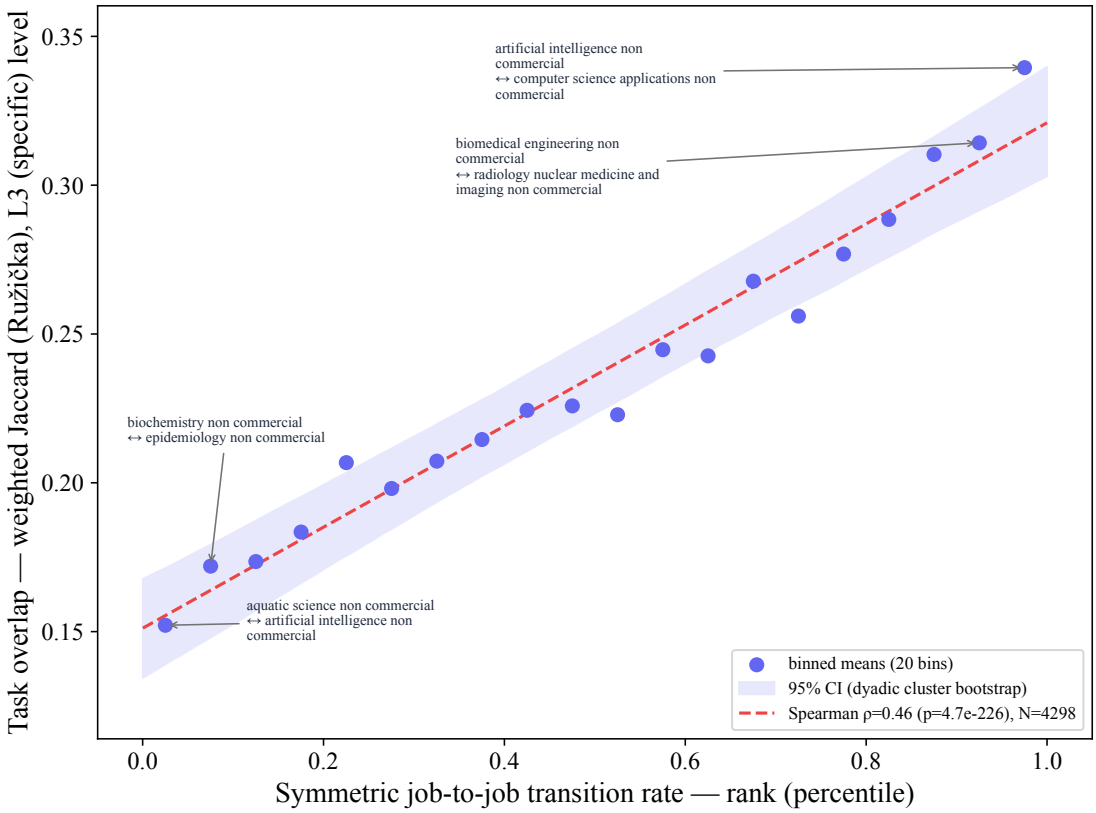
TAXONOMY STRUCTURE, LEVEL 1 TO REPRESENTATIVE TASKS



Notes: The taxonomy is a tree of variable depth. Task areas differ in how specifically they describe the work performed, and the taxonomy allows branches of different lengths. Level-2 area 2A is specified further into two Level-3 areas, 3A and 3B; Level-2 area 2B is not, so its representative tasks attach directly to it. Dashed arrows run from the deepest labelled area on each branch to its slice of the representative-task pool. The pool is partitioned by these areas: every representative task belongs to exactly one Level-2 or Level-3 area.

FIGURE 3

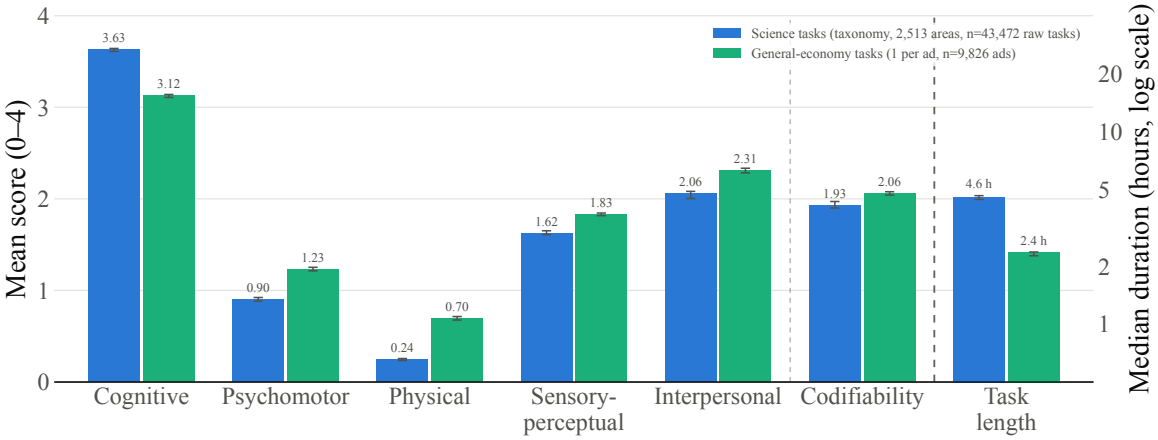
JOB-TO-JOB TRANSITIONS AND TASK OVERLAP



Notes: Binned means of weighted Jaccard (Ružička) task overlap against the rank of the symmetric job-to-job transition rate between occupation pairs, at the Level-3 task level; 20 bins, 95% CI from a dyadic cluster bootstrap; Spearman $\rho = 0.46$, $N = 4{,}298$ pairs.

FIGURE 4

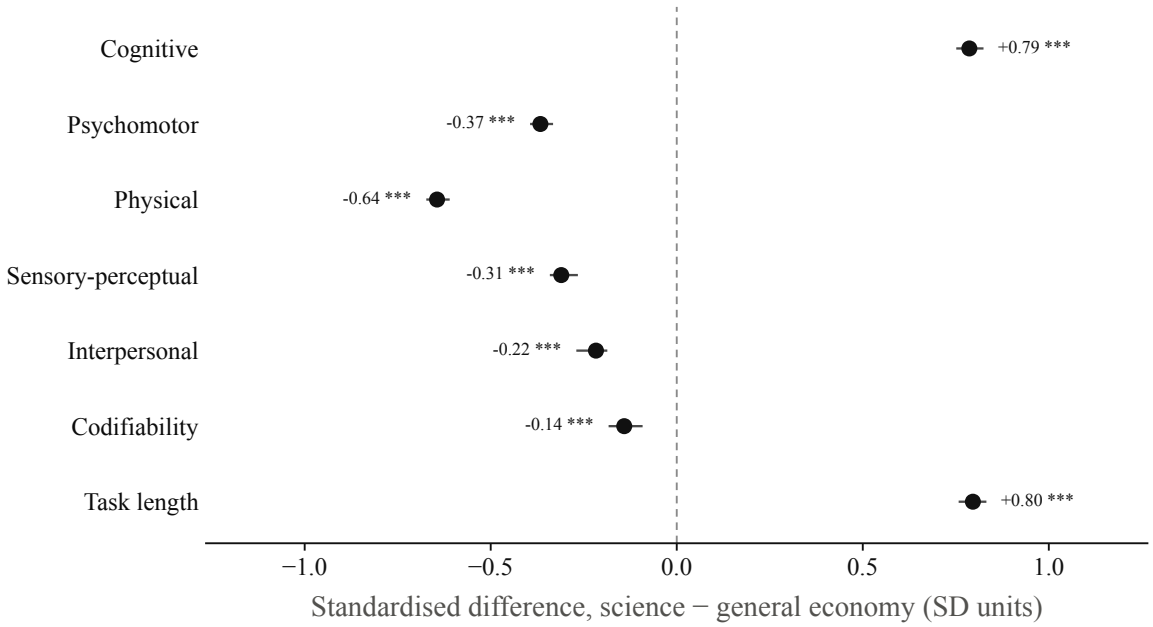
TASK REQUIREMENT PROFILES: SCIENCE VS THE GENERAL ECONOMY



Notes: Mean task requirement and codifiability scores (0–4) for science tasks against the general-economy control group. The science sample is 43,472 raw task instances drawn from within the 2,513 taxonomy areas, weighted by nested union coverage. The control comes from 9,826 non-science adverts contributing one task each. Confidence intervals are 95% cluster-bootstrap intervals per arm over 2,000 draws; the science sample resamples the 2,513 areas and recomputes weights on every draw, the control resamples adverts. Significance is from the bootstrap on the difference and every dimension differs at $p < 0.001$. *Task length is the median corrected duration on the right-hand log axis.*

Figure 5

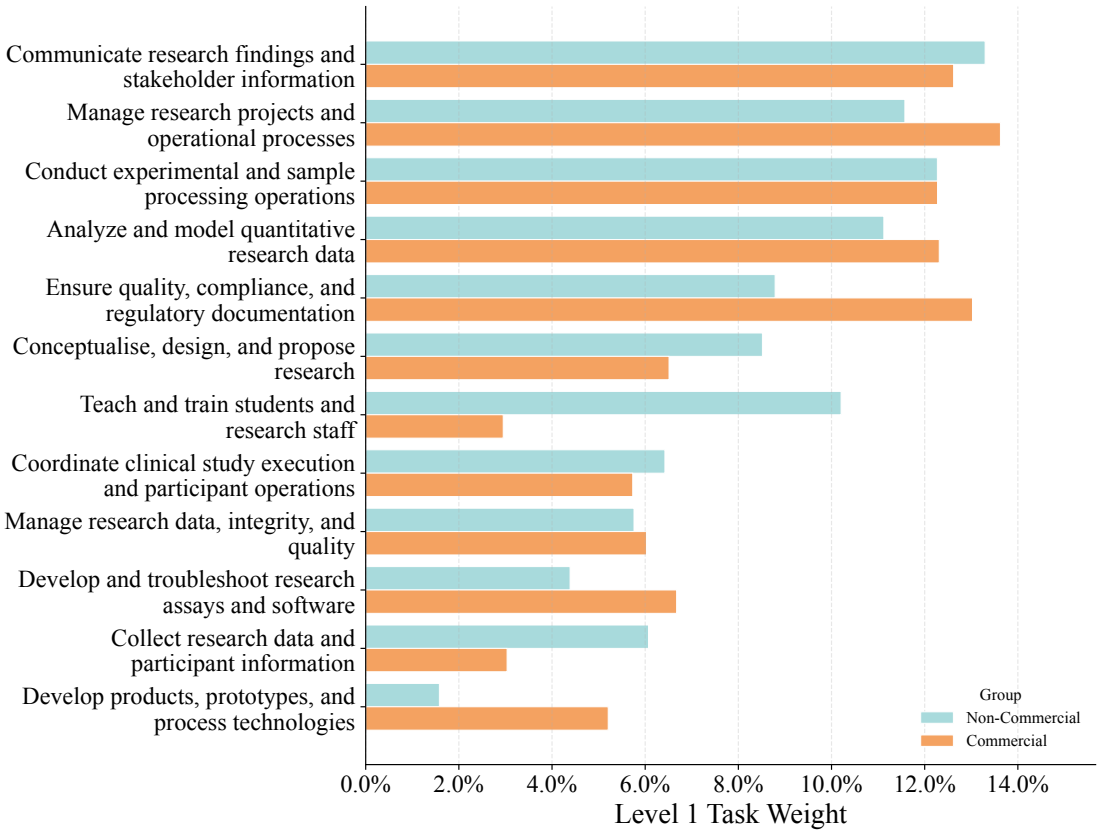
Standardised differences: science vs the general economy



Notes: Each point is the coefficient on a science dummy from a regression whose outcome has been z-standardised by the pooled task-level standard deviation. Confidence intervals are the 95% cluster-bootstrap interval on the difference over that same denominator. Because the units are standard deviations, task length sits on the axis with the scored dimensions; its median ratio is 1.93× long (4.6 against 2.4 hours). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Figure 6

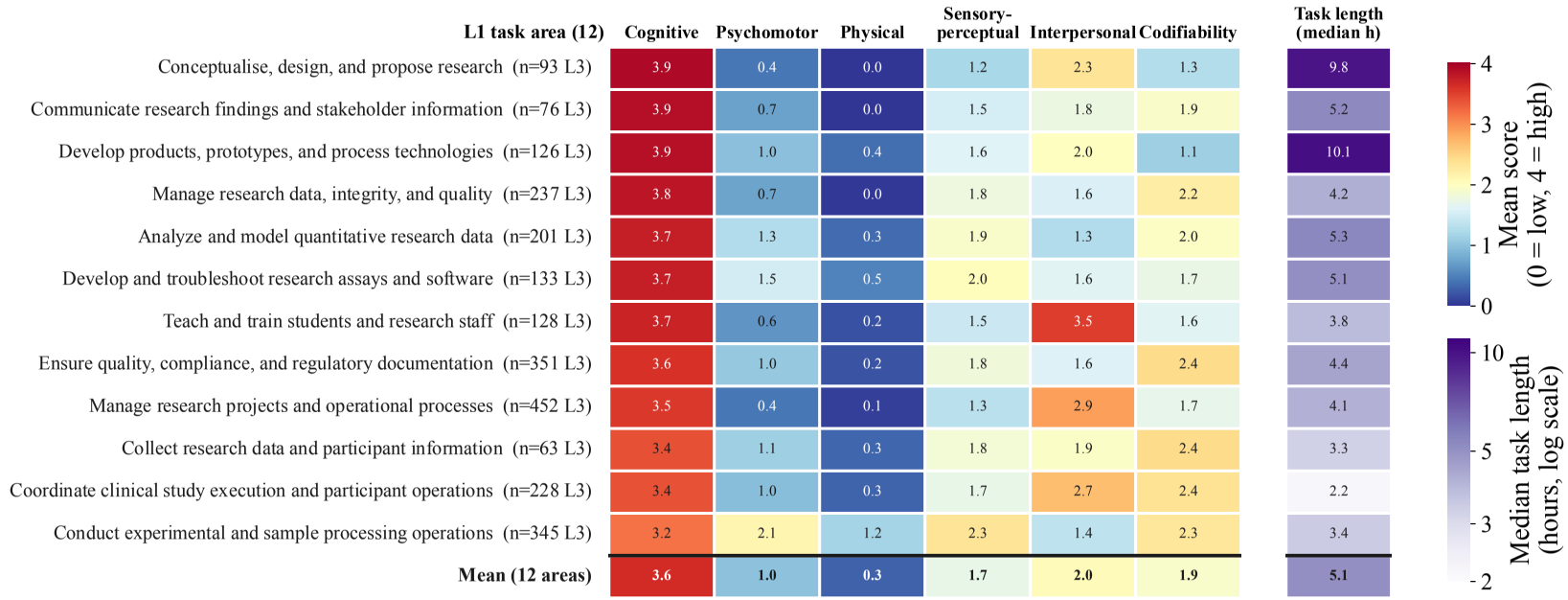
Level-1 task distribution, commercial vs non-commercial



Notes: Weighted mean of Level-1 union-coverage task shares across all occupations in each group, renormalised within group to sum to 100%. A Level-1 category's weight is the share of the occupation's adverts mentioning at least one child task under it. Commercial = industry; non-commercial = academia, government and non-profit.

Figure 7

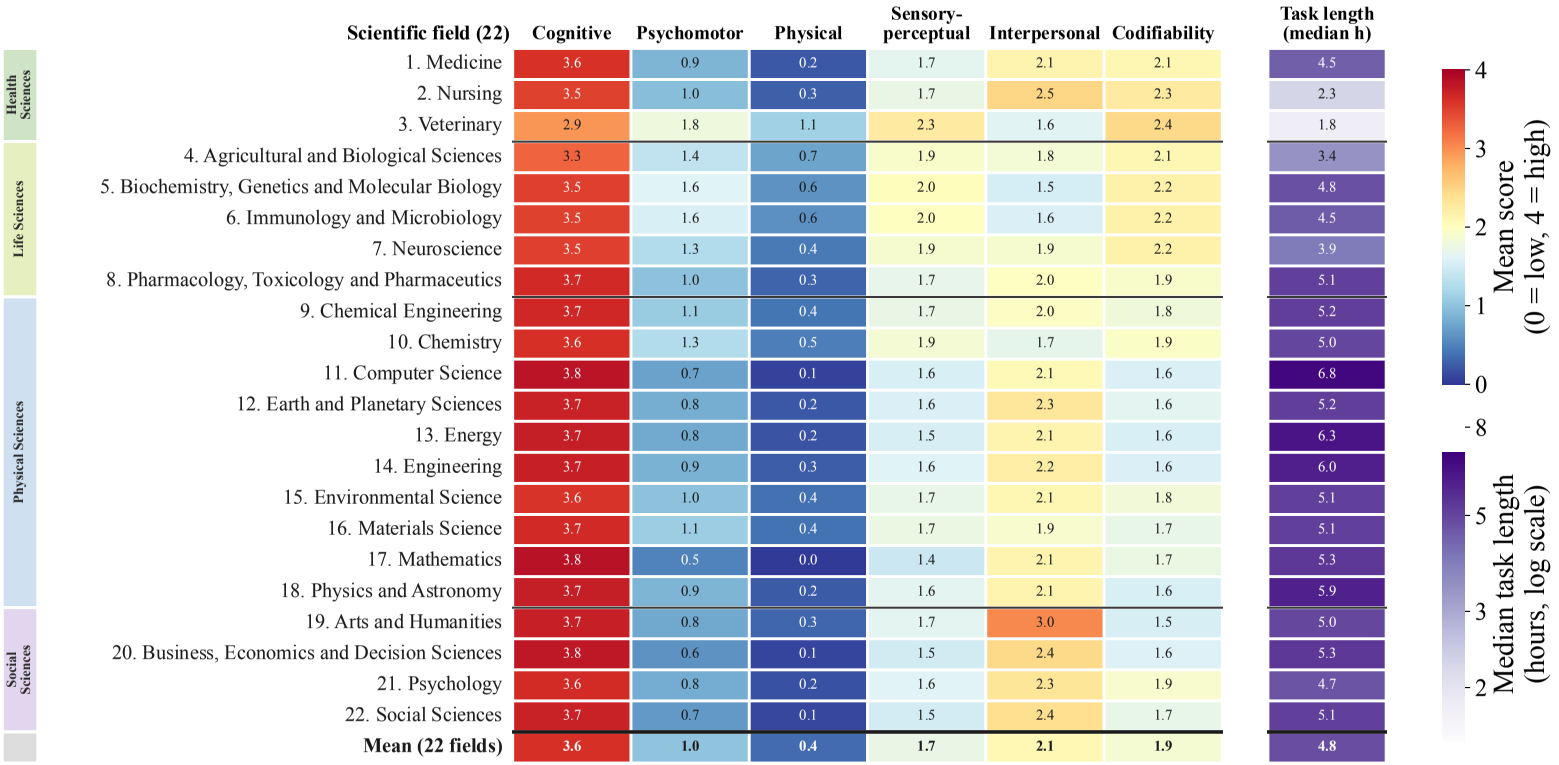
Task requirements by Level-1 task area



Notes: Mean requirement scores and median task length across the 12 Level-1 task areas. Per-area scores are the means over raw raw advert text described in Section 3, aggregated up the taxonomy weighted by union coverage. The rows are ordered by cognitive score. The six score columns share a single 0–4 scale and so are comparable within a row. Task length is the advert-weighted median duration in hours on its own log colour scale. The bottom row is the mean across the 12 areas.

Figure 8

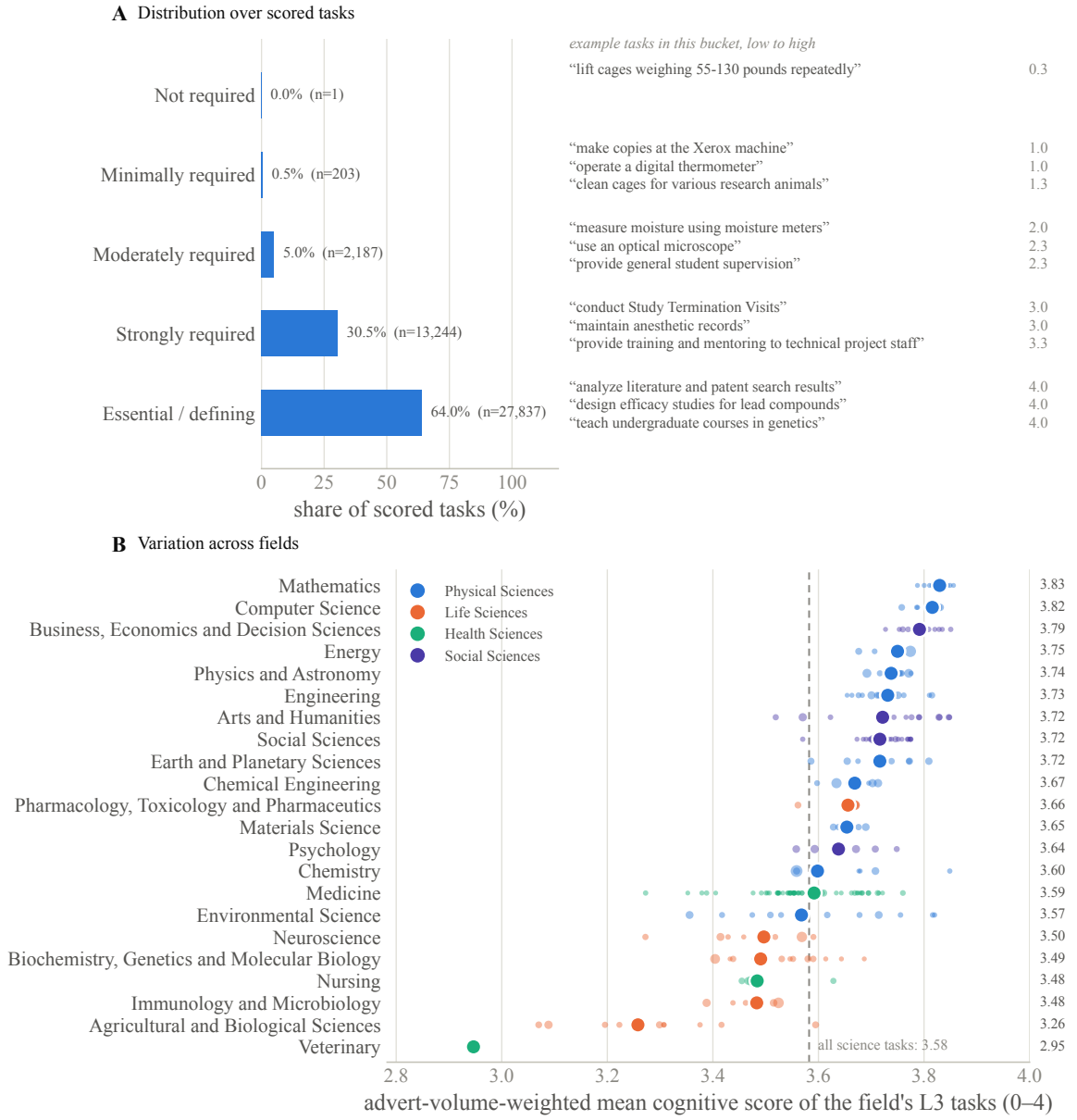
Task requirements by scientific field



Notes: This plot shows the mean requirement scores and median task length, aggregated to the 22 OpenAlex broad fields. The rows are grouped by OpenAlex domain. The six score columns share a single 0–4 scale and so are comparable within a row. Task length is the advert-weighted median duration in hours on its own log colour scale. The bottom row is the mean across the 12 areas.

FIGURE 9

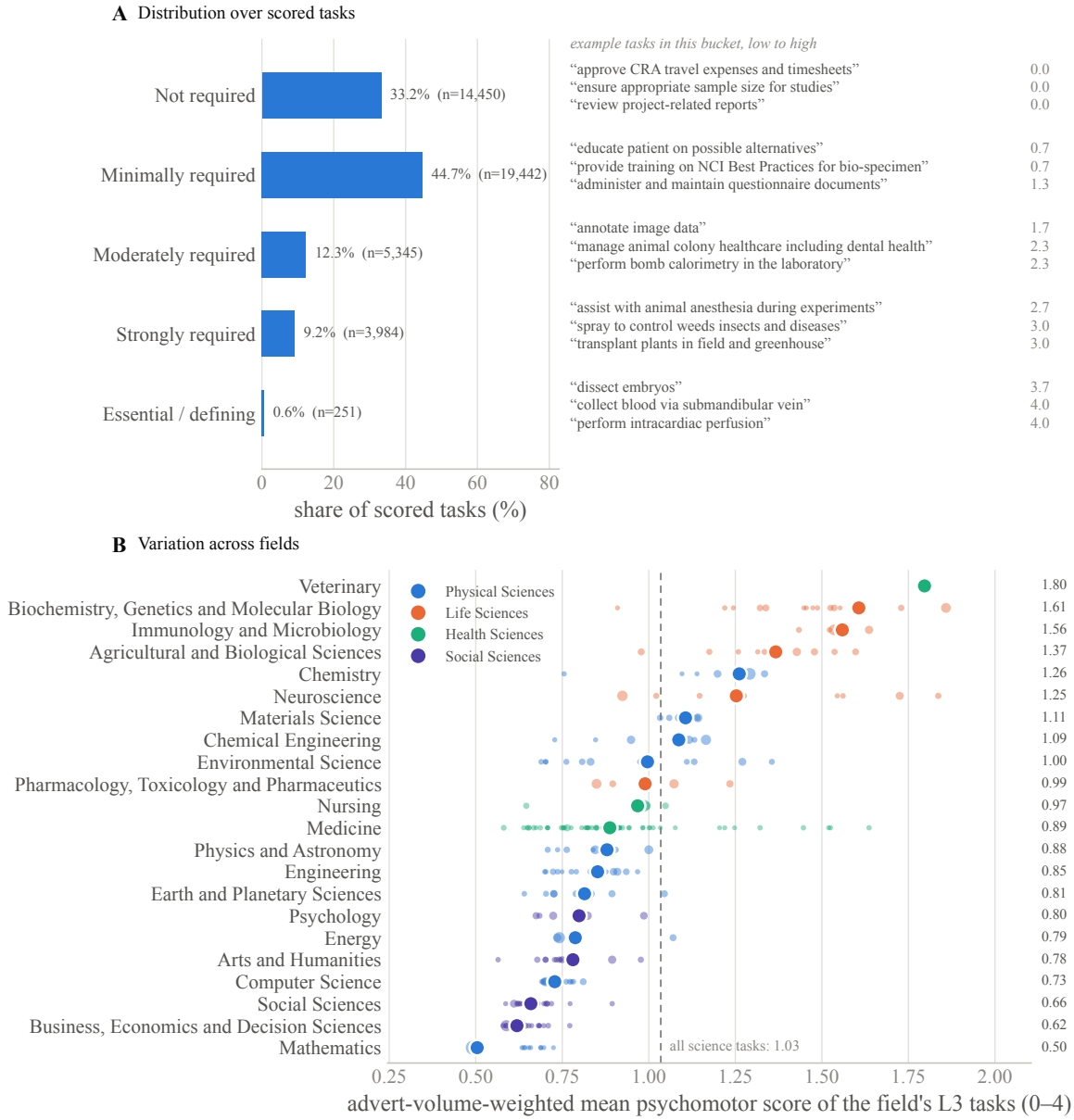
COGNITIVE REQUIREMENT ACROSS SCIENCE TASKS



Notes: Panel A) is the distribution of estimated cognitive requirement score over the 43,472 raw text science tasks, bucketed onto the five rubric anchors the model was given, by rounding the three-vote mean. Example tasks are raw text science tasks drawn evenly across each bucket's value range and printed low to high with their own value, Panel B) ranks the 22 broad fields on the advert-weighted cognitive requirement score; the dashed line is the advert-weighted mean over all science tasks, which differs slightly from the coverage-weighted mean in Figure 4.

Figure 10

Psychomotor requirement across science tasks



Notes: Panel A) is the distribution of estimated psychomotor requirement over the 43,472 raw text science tasks, bucketed onto the five rubric anchors the model was given, by rounding the three-vote mean. Example tasks are raw text science tasks drawn evenly across each bucket's value range and printed low to high with their own value, Panel B) ranks the 22 broad fields on the weighted psychomotor requirement score; the dashed line is the advert-weighted mean over all science tasks, which differs slightly from the coverage-weighted mean in Figure 4. The field is marker is largest, with each of the field's subfields size weighted by advert volume.

FIGURE 11

PHYSICAL REQUIREMENT ACROSS SCIENCE TASKS



Notes: Panel A) is the distribution of estimated physical requirement over the 43,472 raw text science tasks, bucketed onto the five rubric anchors the model was given, by rounding the three-vote mean. Example tasks are raw text science tasks drawn evenly across each bucket's value range and printed low to high with their own value, Panel B) ranks the 22 broad fields on the weighted physical requirement score; the dashed line is the advert-weighted mean over all science tasks, which differs slightly from the coverage-weighted mean in Figure 4. The field is marker is largest, with each of the field's subfields size weighted by advert volume.

FIGURE 12

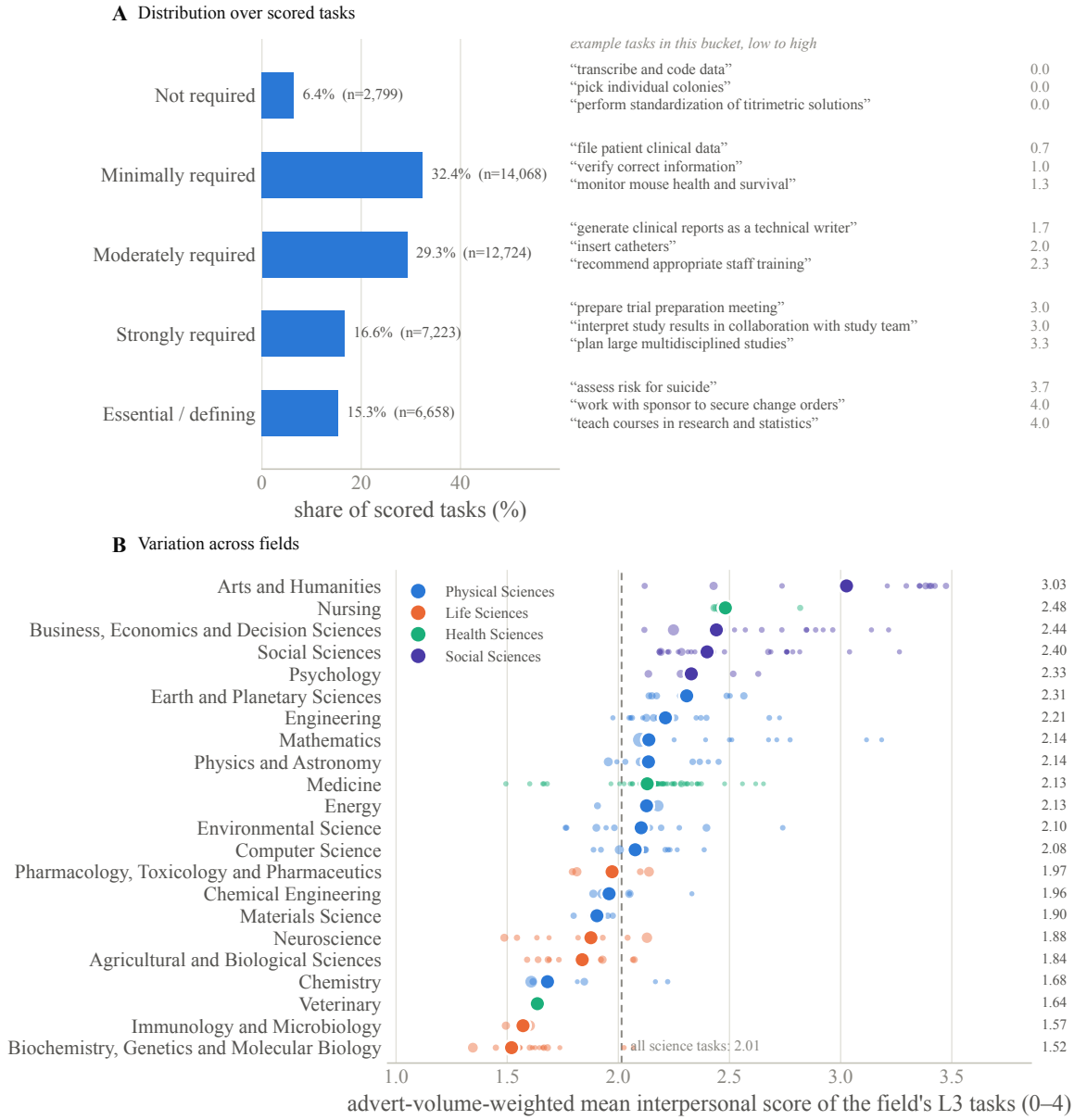
SENSORY-PERCEPTUAL REQUIREMENT ACROSS SCIENCE TASKS



Notes: Panel A) is the distribution of estimated sensory-perceptual requirement over the 43,472 raw text science tasks, bucketed onto the five rubric anchors the model was given, by rounding the three-vote mean. Example tasks are raw text science tasks drawn evenly across each bucket's value range and printed low to high with their own value, Panel B) ranks the 22 broad fields on the weighted sensory-perceptual requirement score; the dashed line is the advert-weighted mean over all science tasks, which differs slightly from the coverage-weighted mean in Figure 4. The field is marker is largest, with each of the field's subfields size weighted by advert volume.

Figure 13

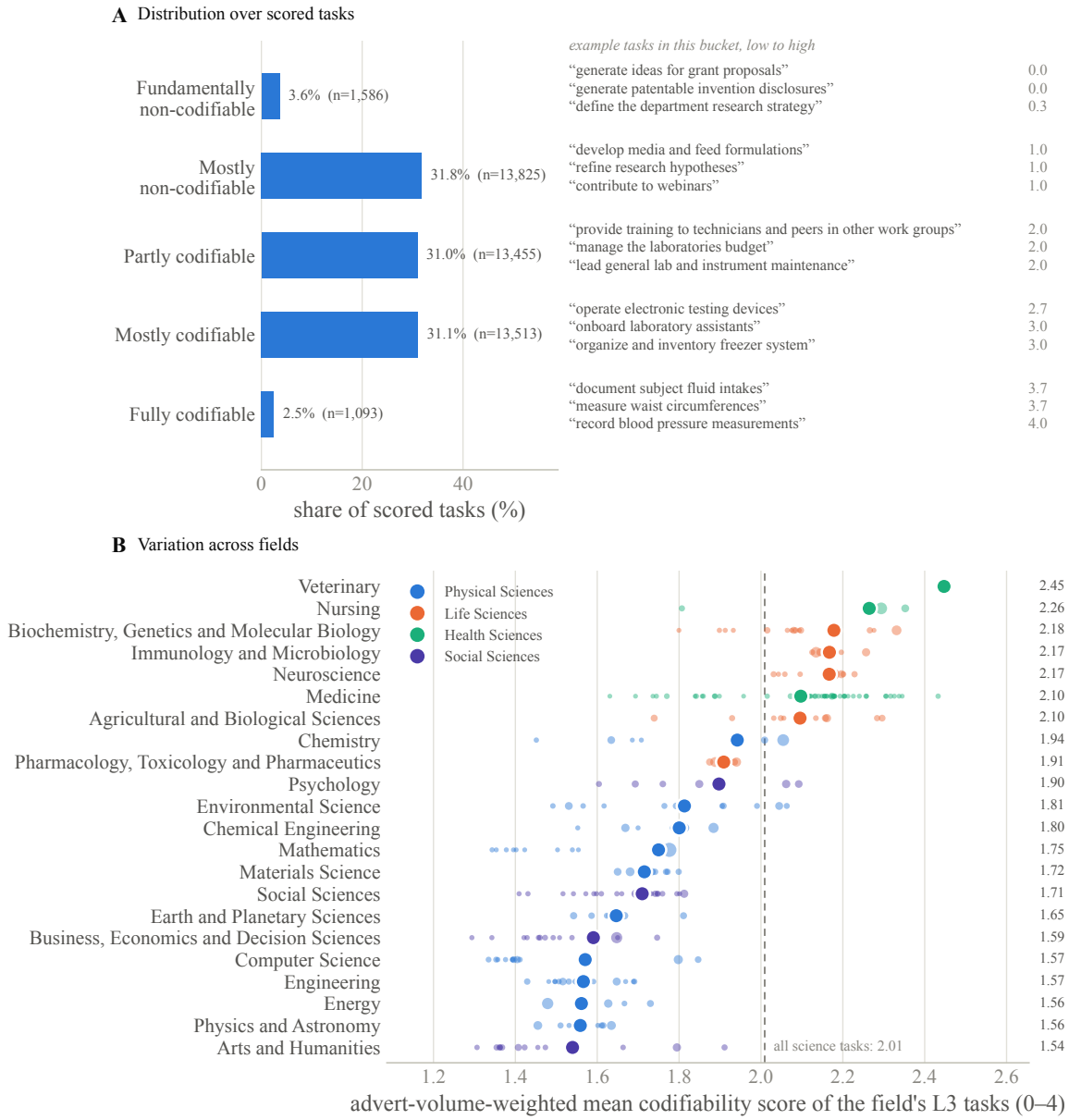
Interpersonal requirement across science tasks



Notes: Panel A) is the distribution of estimated interpersonal requirement over the 43,472 raw text science tasks, bucketed onto the five rubric anchors the model was given, by rounding the three-vote mean. Example tasks are raw text science tasks drawn evenly across each bucket's value range and printed low to high with their own value, Panel B) ranks the 22 broad fields on the weighted interpersonal requirement score; the dashed line is the advert-weighted mean over all science tasks, which differs slightly from the coverage-weighted mean in Figure 4. The field is marker is largest, with each of the field's subfields size weighted by advert volume.

FIGURE 14

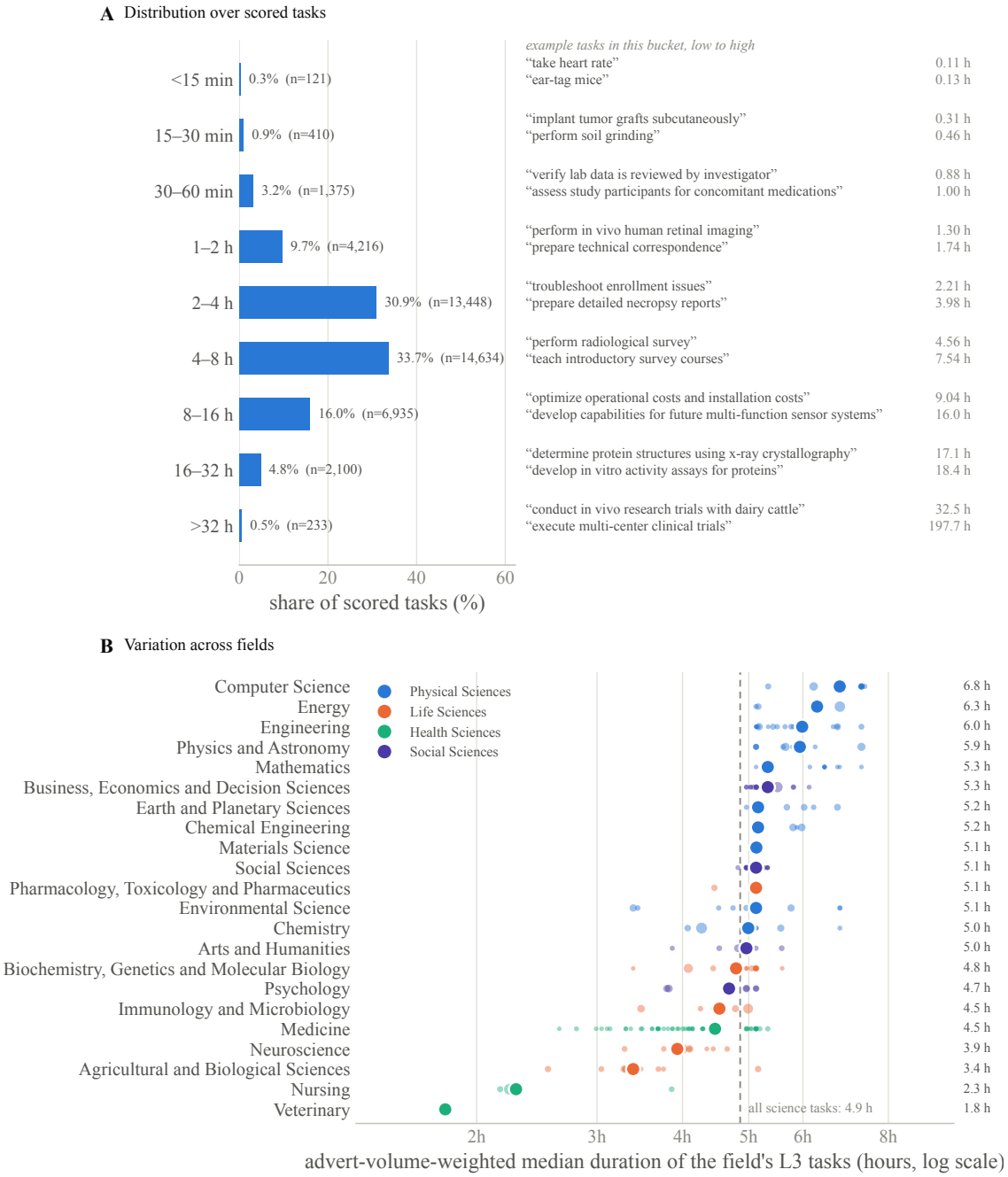
CODIFIABILITY ACROSS SCIENCE TASKS



Notes: Panel A) is the distribution of estimated codifiability requirement over the 43,472 raw text science tasks, bucketed onto the five rubric anchors the model was given, by rounding the three-vote mean. Example tasks are raw text science tasks drawn evenly across each bucket's value range and printed low to high with their own value, Panel B) ranks the 22 broad fields on the weighted codifiability requirement score; the dashed line is the advert-weighted mean over all science tasks, which differs slightly from the coverage-weighted mean in Figure 4. The field is marker is largest, with each of the field's subfields size weighted by advert volume.

Figure 15

Human task length across science tasks



Notes: Panel A) is the distribution of estimated task duration over the 43,472 raw text science tasks, in nine doubling buckets from under 15 minutes to over 32 hours. Example tasks are raw text science tasks drawn evenly across each bucket's value range and printed low to high with their own value, Panel B) ranks the 22 broad fields on the weighted median duration. The field is marker is largest, with each of the field's subfields size weighted by advert volume.

A Additional Tables and Figures

Each posting receives an integer score equal to the sum of the weights of the signals it triggers (Table A1). The eight signals are evaluated independently on the posting's title, body, and employer-name fields; no signal gates any other. Scores range from −6 (both negative signals present) to +12 (all six positive signals present).

Table A1

Heuristic science-score signals

#	Field	Signal	Weight
Positive signals			
1	Title	Research-role tokens: *research(er), scientist, postdoc(toral), investigator, R&D, research director/manager/lead/head, [modifier] fellow*	+2
2	Body	Mentions "research" at least three times	+2
3	Title	Research support / operations roles: *lab(oratory) technician, research technician/coordinator/administrator, research data manager, lab manager, grants manager, clinical research coordinator/associate, research nurse/practitioner/engineer*	+2
4	Employer	Research-oriented organisation: *university, college, institute, hospital, foundation, laborator(y/ies), research cent(re/er), clinic*	+2
5	Body	Research activity: *conduct/lead/design/carry out/undertake research, research programme/project/agenda/study, experimental design/methods/protocols, study design, scientific/laboratory methods, clinical protocols*	+2
6	Body	Research outputs: *publications, publish(ed), peer-reviewed, manuscripts, journal articles, conference/scientific papers, preprints, dissertations, thes(is/es)*	+2
Negative signals			
7	Body	Non-scientific "research": *market/consumer/user/UX/equity/investment/policy/opinion/financial/ competitive/media/audience research, competitive intelligence, due diligence*	−3
8	Title or Body	Commercial / client-facing role: *sales, marketing, business development, account manager/management, customer success, commercial, client-facing, revenue targets*	−3

Notes: Each signal is a case-insensitive regular-expression match returning a binary indicator; the score is the weighted sum of triggered signals. Signals are evaluated independently on each posting, giving a range of −6 *to* +12*. The score is used only to stratify the sample sent for LLM labelling and is not applied as a filter on the corpus.*

Table A2

Science classifier: model comparison

Metric	Logistic Regression	XGBoost	LinearSVC
Test set metrics			
Accuracy	0.9167	0.9014	0.9016
Precision	0.9619	0.9690	0.9870
Recall	0.6792	0.6079	0.5971
F1	0.7962	0.7471	0.7441
ROC-AUC	0.9772	0.9765	0.9781
Avg. Precision	0.9383	0.9349	0.9401
Adjusted for real-world prior (2.00%)			
TPR / Recall	0.6792	0.6079	0.5971
FPR	0.0085	0.0061	0.0025
Adj. Precision	0.6207	0.6697	0.8314
Adj. F1	0.6487	0.6373	0.6950

Notes: Three candidate classifiers for the binary science model, each fitted and evaluated on the same score-stratified splits of the LLM-labelled sample. The upper panel reports test-set metrics at a 0.9 probability threshold. The lower panel re-states precision and F_1 at a real-world baseline prior of 2.00% science adverts. Test-set precision is measured on a sample enriched for science, so it overstates performance on the corpus, whereas recall and the false-positive rate are prevalence-invariant and carry over unchanged. Adjusted precision is $\mathrm{TPR}\pi/(\mathrm{TPR}\pi + \mathrm{FPR}(1-\pi))$ *at* $\pi = 0.02$.

Table A3

Broad-field classifier: three-head comparison

	LinearSVC			Logistic reg.			MLP		
	Estimate	95% CI		Estimate	95% CI		Estimate	95% CI	
		LL	*UL*		*LL*	*UL*		*LL*	*UL*
Validation micro-F_1	0.794	–	–	0.746	–	–	0.774	–	–
Test micro-F_1 (global)	0.792	.788	.797	0.748	.743	.752	0.774	.769	.779
Test micro-F_1 (tuned)	0.825	.820	.831	0.815	.809	.820	0.841	.836	.846
Test macro-F_1 (global)	0.768	.760	.775	0.702	.695	.709	0.716	.708	.723
Test macro-F_1 (tuned)	0.794	.786	.801	0.787	.779	.794	0.812	.805	.819

Notes: Held-out test set (n = 11,000*; 23 classes = 22 fields + Unclassified). Point estimates with 95% non-parametric bootstrap CIs (1000 resamples). LL and UL denote the lower and upper confidence limits, respectively. "Global" applies a fixed 0.1 probability threshold; "tuned" applies per-class thresholds selected on the validation set. The MLP classifier, highlighted in bold, is selected as the headline model.*

Table A4

Broad-field classifier: per-class metrics

Field	Test n	Prec.	Rec.	F_1	95% CI
Agricultural and Biological Sciences	486	0.770	0.893	0.827	[0.804, 0.850]
Arts and Humanities	98	0.833	0.867	0.850	[0.795, 0.896]
Biochemistry, Genetics and Molecular Biology	1,654	0.824	0.864	0.843	[0.829, 0.857]
Business, Economics and Decision Sciences	182	0.853	0.731	0.787	[0.741, 0.836]
Chemical Engineering	233	0.749	0.614	0.675	[0.618, 0.725]
Chemistry	543	0.834	0.775	0.803	[0.777, 0.831]
Computer Science	1,078	0.879	0.875	0.877	[0.861, 0.890]
Earth and Planetary Sciences	129	0.847	0.775	0.810	[0.750, 0.863]
Energy	199	0.750	0.874	0.807	[0.766, 0.844]
Engineering	1,002	0.845	0.817	0.831	[0.814, 0.850]
Environmental Science	286	0.748	0.780	0.764	[0.724, 0.800]
Immunology and Microbiology	718	0.851	0.772	0.809	[0.785, 0.832]
Materials Science	293	0.815	0.799	0.807	[0.770, 0.839]
Mathematics	308	0.812	0.802	0.807	[0.770, 0.839]
Medicine	3,706	0.867	0.910	0.888	[0.881, 0.896]
Neuroscience	364	0.815	0.775	0.794	[0.760, 0.828]
Nursing	152	0.885	0.961	0.921	[0.888, 0.951]
Pharmacology, Toxicology and Pharmaceutics	625	0.690	0.786	0.734	[0.707, 0.761]
Physics and Astronomy	354	0.775	0.856	0.813	[0.782, 0.844]
Psychology	315	0.814	0.749	0.780	[0.742, 0.813]
Social Sciences	466	0.766	0.841	0.802	[0.773, 0.828]
Unclassified	1,781	0.912	0.848	0.879	[0.866, 0.890]
Veterinary	192	0.750	0.797	0.773	[0.726, 0.815]

Notes: MLP classifier, per-class metrics on the held-out test set at tuned thresholds, with 95% bootstrap CIs on the F_1 score. Test n is the number of positive instances of each class in the test set; because a posting may be assigned to more than one field, the column sums to 15,164 across 11,000 test adverts. "Unclassified" is the LLM opt-out class (positions with this top field are dropped from the corpus); "Veterinary" receives no subfield classifier, since we collapse its OpenAlex subfields into the broad field (Section 2.2.2).

Table A5

Building the taxonomy: counts at each step

	Distinct tasks	Task instances
Step 1: representative tasks		
Task instances extracted	—	64,899,085
Unique task statements	20,725,623	62,957,146
Near-duplicate clusters ($\tau = 0.72$)	6,820,270	62,957,146
representatives seen in ≥ 10 adverts	1,073,177	42,636,698
after the second pass ($\tau = 0.90$)	700,562	42,636,698
representatives seen in ≥ 30 adverts	255,618	35,380,567
Specific work tasks	221,966	—
Step 2: hierarchical clustering		
Placed in a Level-2 and Level-3 area	213,874	—
Step 3: LLM as judge and namer		
Final taxonomy (12 / 114 / 2,433 task areas)	208,202	—

Notes: Statements longer than 150 characters (3% of instances) are dropped before deduplication. Uniqueness is mesasured by exact string matching. Near-duplicate is measured by the cosine similarity on embedding representations using the Qwen3-Embedding-8B model. Specific work tasks is the count of tasks that pass the final LLM check for whether that task statement reflects a specific task, and not a generic or non-descript work related activity e.g. company-wide objectives or background knowledge required.

Table A6

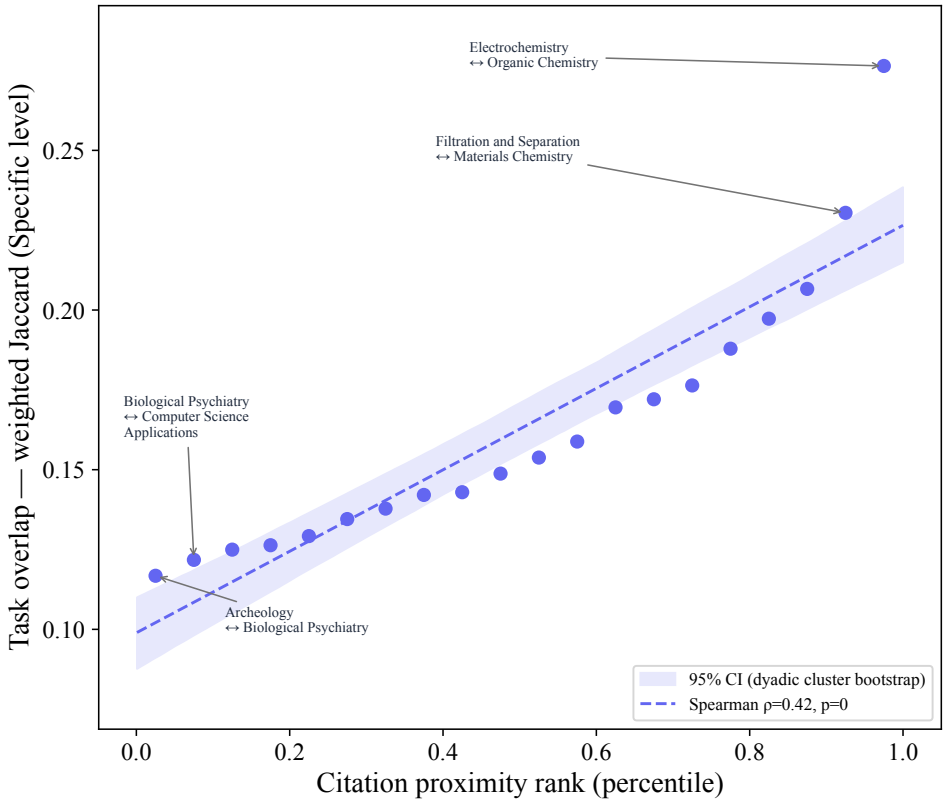
Occupational composition of the non-science control sample

SOC	Occupation group	Adverts	%
29-0000	Healthcare Practitioners and Technical Occupations	1,143	11.6
41-0000	Sales and Related Occupations	1,064	10.8
11-0000	Management Occupations	1,049	10.7
15-0000	Computer and Mathematical Occupations	956	9.7
43-0000	Office and Administrative Support Occupations	922	9.4
13-0000	Business and Financial Operations Occupations	837	8.5
53-0000	Transportation and Material Moving Occupations	623	6.3
35-0000	Food Preparation and Serving Related Occupations	408	4.2
51-0000	Production Occupations	369	3.8
49-0000	Installation, Maintenance, and Repair Occupations	352	3.6
31-0000	Healthcare Support Occupations	300	3.1
27-0000	Arts, Design, Entertainment, Sports, and Media Occupations	280	2.8
17-0000	Architecture and Engineering Occupations	276	2.8
25-0000	Educational Instruction and Library Occupations	242	2.5
37-0000	Building and Grounds Cleaning and Maintenance Occupations	173	1.8
39-0000	Personal Care and Service Occupations	139	1.4
47-0000	Construction and Extraction Occupations	133	1.4
21-0000	Community and Social Service Occupations	130	1.3
33-0000	Protective Service Occupations	120	1.2
99-0000	Unclassified Occupation	115	1.2
19-0000	Life, Physical, and Social Science Occupations	97	1.0
23-0000	Legal Occupations	70	0.7
55-0000	Military-only occupations	15	0.2
45-0000	Farming, Fishing, and Forestry Occupations	13	0.1
	Total	*9,826*	*100.0*

Notes: Lightcast SOC 2021 occupation codes for the 9,826 adverts in the non-science control sample. The sample is defined by two screening filters that never see the occupation code. First the science classifier scores the advert below 0.3, and the Stage-1 extractor is then separately asked whether the advert describes science research activity.

Figure A1

Citation proximity and task overlap



Notes: Binned means of weighted Jaccard (Ružička) task overlap against the rank of citation proximity between subfield pairs, at the Level-3 task level; 95% CI from a dyadic cluster bootstrap; Spearman $\rho = 0.42$.

Table A7

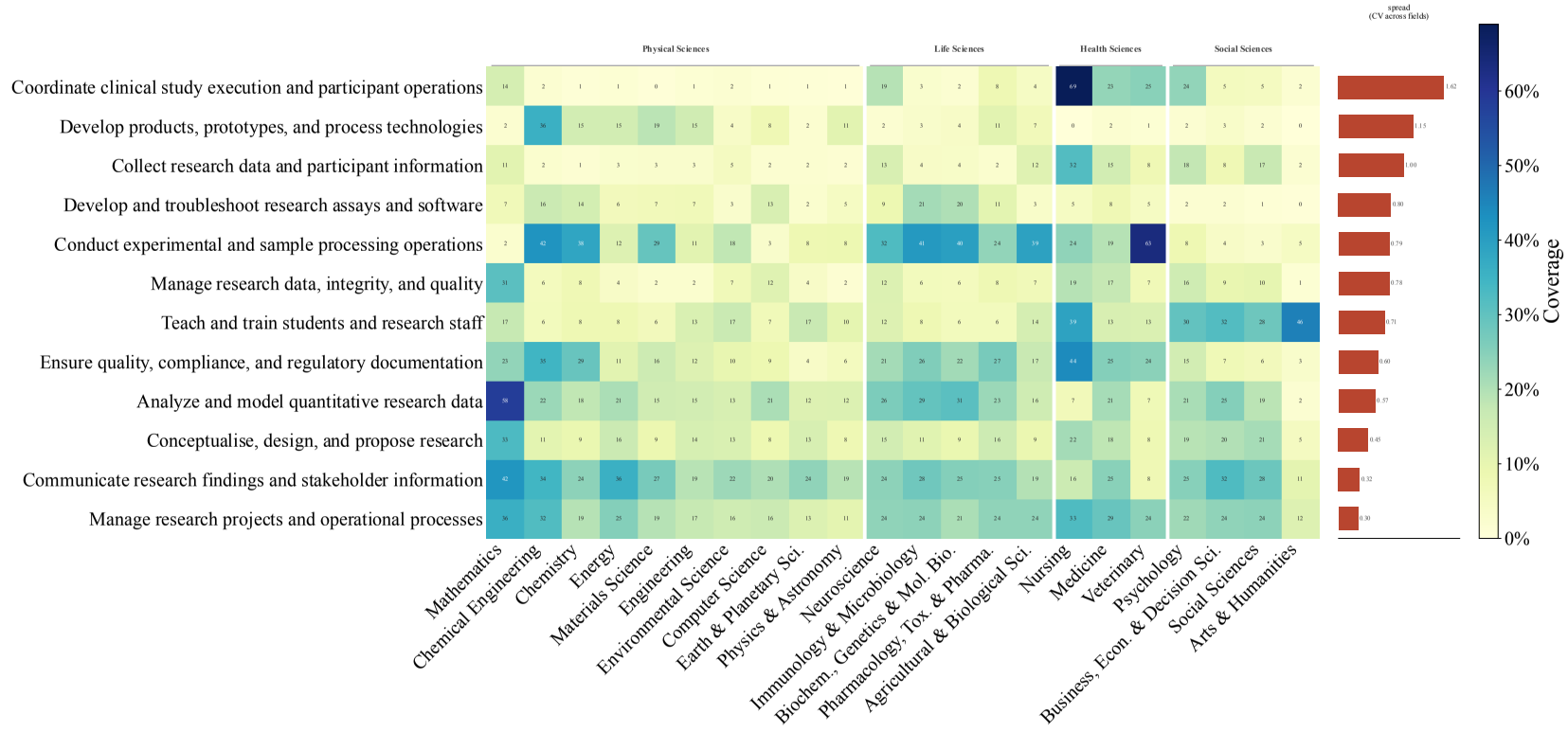
Task overlap regressed on citation proximity, by taxonomy level

	Level-1	Level-2	Level-3	Level-4
Citation proximity (percentile rank)	0.193***	0.193***	0.128***	0.072***
	(0.018)	(0.014)	(0.009)	(0.005)
Spearman ρ	0.352***	0.416***	0.425***	0.478***
R^2	0.130	0.186	0.191	0.215
Mean task overlap	0.445	0.288	0.163	0.055
Pairs	26,565	26,565	26,565	26,565

Notes: OLS of the weighted Jaccard (Ružička) task overlap between subfield pairs on the percentile rank of their citation proximity, so the coefficient is the fitted change in overlap from the least to the most scientifically proximate pair. Standard errors in parentheses and 95% intervals in brackets come from a dyadic cluster bootstrap over subfields (2,000 draws), the same procedure that draws the confidence band in Figure A1.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Figure A2

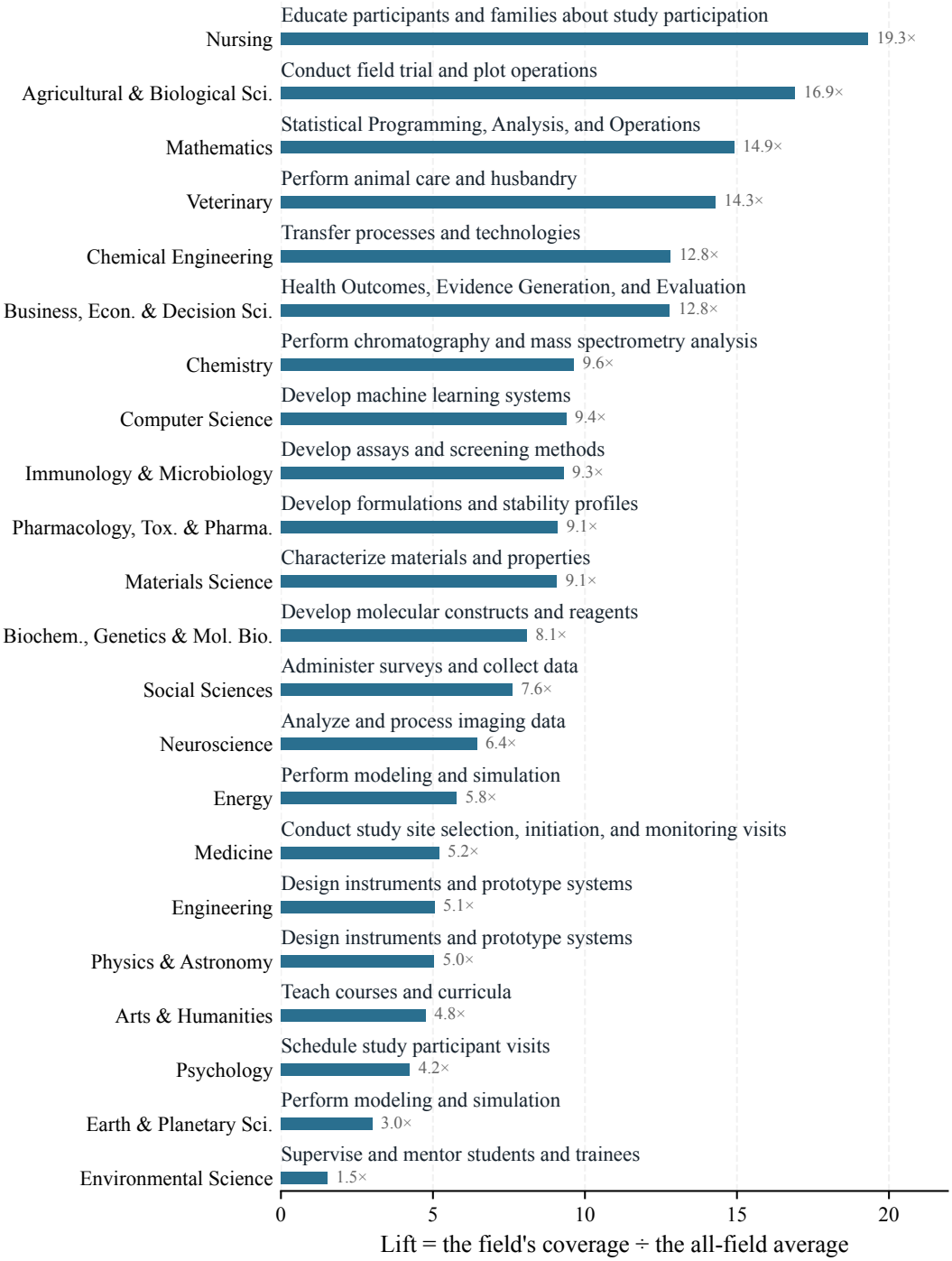
Level-1 task areas by broad field



Notes: Level-1 task areas × the 22 OpenAlex broad fields. Each cell is the field's advert-weighted share of job adverts mentioning at least one task under the node, measured with soft occupation membership and union coverage. Rows are ordered by the coefficient of variation and the columns are grouped by OpenAlex domain. The results are aggregated over both sectors and all corpus years.

Figure A3

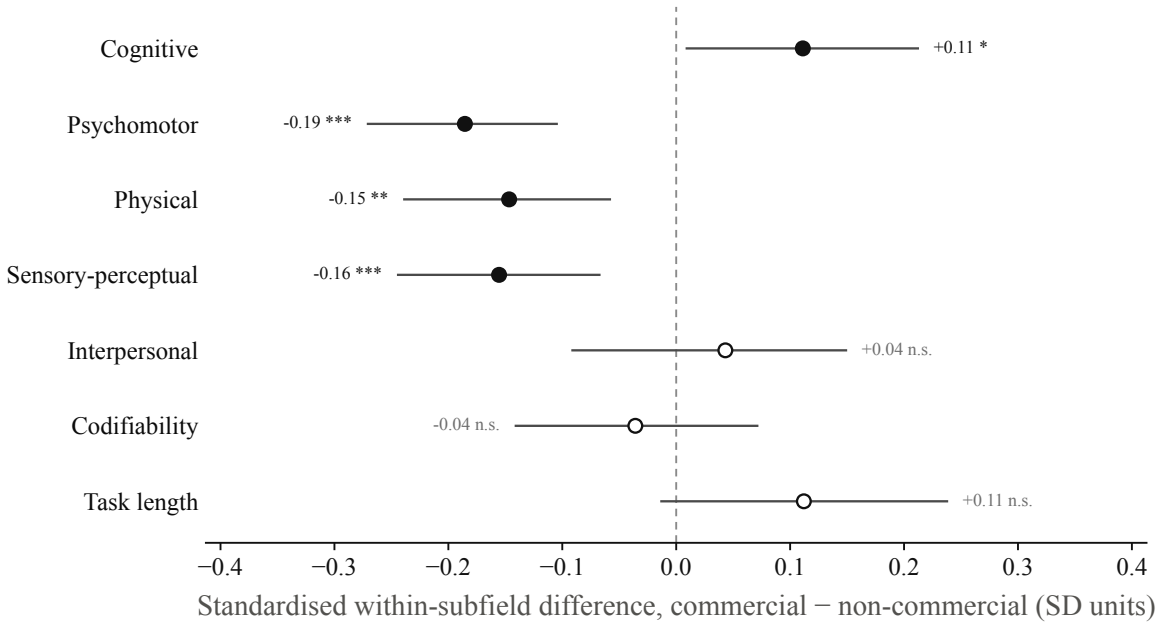
Signature tasks by field



Notes: Each field's most distinctive Level-2 task, measured as lift — the field's coverage divided by the average field's coverage of that task — among tasks covering at least 5% of the field's adverts.

Figure A4

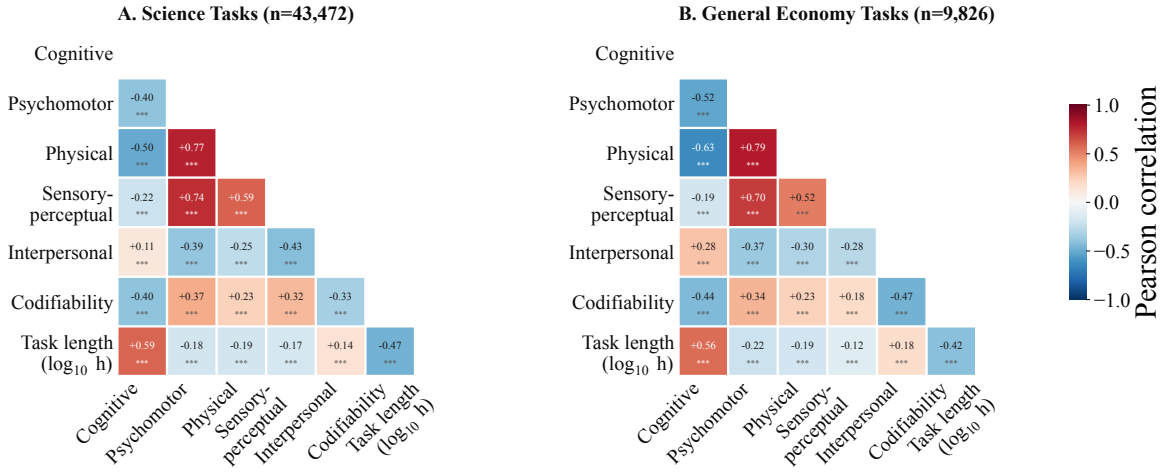
Commercial vs non-commercial science tasks



Notes: The estimate shown is the coefficient on a commercial dummy from a weighted least squares fit, with subfield fixed effects, on a z-standardised outcome so the seven dimensions are comparable and task length can share the axis. This regression is identified off the 217 of 232 subfields carrying both sectors (99.6% of volume). Controlling for subfield matters because commercial share ranges from 0% of a subfield to 92%, so a raw pooled gap is partly composition. Confidence intervals are 95% bootstrap intervals resampling Level-3 nodes over 2,000 draws. A hollow marker identifies that the interval covers zero. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

FIGURE A5

CROSS-CORRELATION OF THE TASK CHARACTERISTICS



Notes: Pearson correlations among the seven task characteristics, for the science sample (n = 43,472 raw task instances) and the general-economy control (n = 9,826). The six scores enter on the raw 0–4 scale and duration as $\log_{10}$ *hours.* $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

B LLM Prompts

Prompts are reproduced verbatim, with em dashes rendered as hyphens. Each note states the model, the decoding settings, and the shape of the input and output.

B.1 Phase 1 — Science-job classification

Model: `gpt-5.4` (OpenAI Batch API).
Input: system message below; the user message is the job-advert body, truncated to the first 10,000 characters (no title or other metadata).
Output: a single token, `Yes` or `No` (`is_scientific_research_job` $\in \{0, 1\}$; unlabelled rows default to 0).

```
Task. You are classifying whether a job advertisement is a scientific
    research job.

A scientific research job is one where the primary responsibility is to
    generate new scientific knowledge through systematic investigation,
    experimentation, or theoretical work.
This includes both academic industry, non-profit and government research
    roles.

Important: Focus on the responsibilities of the advertised role itself,
    not the wider purpose of the employing organisation. For example, an
    HR manager or accountant at a university is not a scientific research
    job, even though the university conducts research.

In making your judgement, consider the following indicators:
- Research-oriented job titles (e.g. research scientist, postdoctoral
    fellow, bench scientist, investigator, R&D roles).
- Explicit descriptions of conducting, designing, or leading research
    projects or programmes.
- References to scientific fields, experimental methods, study design, or
     methodology.
- Mentions of research outputs such as publications, journals, or
    conference presentations.
- Employment by universities, research institutes, or dedicated R&D units
```

```
    (supporting evidence, not sufficient alone).
- Include research operations and support roles that directly enable
    scientific research (e.g. research project managers, lab technicians,
    equipment maintenance specialists, research data managers, research
    procurement/purchasing staff, grant managers).

Exclude roles where "research" refers primarily to market, user, policy,
    financial, or investment analysis, or where research activities are
    incidental to commercial, sales, marketing, or client-facing
    responsibilities.

Question. Is this job a scientific research job?

Answer with only "Yes" or "No".
```

B.2 Phase 2 — Broad-field classification

Model: `gpt-5.4` (pilot sync; boost via the OpenAI Batch API).
Input: the system message below; the user message is `"Classify this job advertisement:" + body`.
Output: 0, 1, or 2 broad-field names ("|"-separated), or `Unclassified`. The `FIELDS` block injects the 22 broad fields of Table 1, each with its OpenAlex description (merged description for the merged fields); it is abbreviated below for space. Non-science rows are auto-labelled `Unclassified` without an API call.

```
You are a classifier that categorizes scientific job advertisements into
    OpenAlex broad fields.

Given a job advertisement, classify it into 0, 1, or 2 of the following
    22 broad fields. Return "Unclassified" if the job doesn't clearly fit
    any listed field or is not a scientific/research position -- do not
    force a pick.

FIELDS:
[ one line per field -- "- <Field name>: <OpenAlex field description>" --
  for the 22 broad fields listed in the fields table; omitted here for
    space ]
- Unclassified: The job does not fit any of the above scientific fields,
    or is not a scientific/research position.

INSTRUCTIONS:
1. Read the job advertisement carefully.
2. Identify the primary broad field based on the role's research area and
     responsibilities.
3. Return 1 field if the job clearly belongs to one field.
4. Return 2 fields (separated by " | ") only if the job equally straddles
     two broad fields.
5. Never return more than 2 fields.
6. Return "Unclassified" if NO listed field is a clear fit, or if the
    role is not scientific/research. Forced picks add noise; an honest "
    Unclassified" is more useful.
```

```
RESPONSE FORMAT:
- Single field: "Medicine"
- Two fields: "Chemistry | Biochemistry, Genetics and Molecular Biology"
- No fit / not scientific: "Unclassified"

Return ONLY the field name(s) or "Unclassified", nothing else.
```

B.3 Phase 2 — Subfield classification

Model: `gpt-5.4` (pilot sync; boost via the OpenAI Batch API).
Input: the system message below; the user message is `"Classify this job advertisement:" + body`. The prompt is *instantiated once per broad field* (21 variants): `<broad field>` and its OpenAlex `<broad field context>` are filled in, and the subfield list is that field's children (see Table 1), abbreviated below.
Output: 0, 1, or 2 subfield names ("|"-separated), or `Unclassified`.

```
You are a classifier that categorizes scientific job advertisements into
    OpenAlex subfields within the broad field "<broad field>".

BROAD FIELD CONTEXT:
<OpenAlex description of the broad field (merged description for merged
    fields)>

Given the job advertisement below, classify it into 0, 1, or 2 of the
    following subfields under "<broad field>":
- <subfield 1>
- <subfield 2>
[ ... the child subfields of this broad field; omitted here for space ...
    ]

INSTRUCTIONS:
1. The broad field is fixed; choose the best-fitting subfield(s) only
    from the list above.
2. Return 1 subfield if the role clearly belongs to one.
3. Return 2 subfields (separated by " | ") only if the role equally
    straddles two of them.
4. Never return more than 2 subfields.
5. Return "Unclassified" if NO listed subfield is a clear fit -- do not
    force a pick. Forced picks add noise; an honest "Unclassified" is more
    useful.

RESPONSE FORMAT:
- Single subfield: "Subfield A"
- Two subfields: "Subfield A | Subfield B"
```

```
- No subfield fits: "Unclassified"

Return ONLY the subfield name(s) or "Unclassified", nothing else.
```

B.4 Stage 1 — Task extraction (science filter, title, tasks, sector)

Model: `gpt-5-mini`, reasoning effort `low` (OpenAI Batch API).
Input: the prompt below; the job title, listed company, staffing flag, and job body (truncated to 12,000 characters) are interpolated at the placeholders at the foot of the prompt.
Output: a single JSON object carrying the science flag, standardised title, extracted task list, extracted employer, and sector.

```
You are analyzing a scientific job advertisement.

## STEP 1: Science Research Job Filter
A scientific research job is one where the primary responsibility is to
    generate new scientific knowledge through systematic investigation,
    experimentation, or theoretical work.
This includes both academic, industry, non-profit and government research
     roles.

Important: Focus on the responsibilities of the advertised role itself,
    not the wider purpose of the employing organisation. For example, an
    HR manager or accountant at a university is not a scientific research
    job, even though the university conducts research.

In making your judgement, consider the following indicators:
- Research-oriented job titles (e.g. research scientist, postdoctoral
    fellow, bench scientist, investigator, R&D roles).
- Explicit descriptions of conducting, designing, or leading research
    projects or programmes.
- References to scientific fields, experimental methods, study design, or
     methodology.
- Mentions of research outputs such as publications, journals, or
    conference presentations.
- Employment by universities, research institutes, or dedicated R&D units
     (supporting evidence, not sufficient alone).
- Include research operations and support roles that directly enable
    scientific research (e.g. research project managers, lab technicians,
    equipment maintenance specialists, research data managers, research
    procurement/purchasing staff, grant managers).
```

```
- Exclude roles where "research" refers primarily to market, user, policy,
     financial, or investment analysis, or where research activities are
    incidental to commercial, sales, marketing, or client-facing
    responsibilities.

If this is NOT a science research job, return exactly:
{"is_science_research": false}

If this IS a science research job, continue to the remaining steps and
    return the full JSON.

## STEP 2: Title Standardisation
Produce a standardised occupation title comparable across scientific
    disciplines and sectors.

Assign one `title_status`:

1. `llm_generated` -- the provided title has an error. Set `
    title_error_type` and generate a short generic title from the body.
   - misclassification: title names a different occupation than the body
    describes (title "Archaeologist", body describes a "Sr. Rheologist").
   - multi_level: title spans 3+ seniority levels or fundamentally
    different career stages ("Assistant/Associate/Full Professor", "
    Scientist I/II/III/IV"). Two adjacent levels are NOT an error, keep as
    -is ("Research Associate III/IV", "Junior/Senior Scientist").

2. `standardized_from_body` -- the provided title is valid but the body
    names a more generic, cross-discipline title. Set `title_error_type =
    null` and use it, stripping field-specific suffixes ("Research
    Associate III/IV - Analytical Chemistry" -> "Research Associate III/IV
    ").

3. `validated` -- the provided title is acceptable and the body offers
    nothing more generic. Set `title_error_type = null` and `final_title`
    identical to the provided title.

`final_title` rules (llm_generated / standardized_from_body only):
```

```
- SHORT generic role, 2-4 words, matching the body's seniority and
    function (e.g. "Senior Scientist", "Postdoctoral Researcher", "Lab
    Technician", "R&D Manager").
- Strip ALL field/discipline and application terms (e.g. "Chemistry", "
    Oncology", "Nuclear"). "Thermal Hydraulics Engineer" -> "Engineer". A
    single seniority level ("III", "Senior") is fine; no departments or
    company names.

## STEP 3: Task Extraction
Extract specific, atomic tasks that the position will personally perform.

TASK STATEMENT STRUCTURE (in order):
ACTION + OBJECT + [ENABLER] + [CONTEXT]

Component definitions:
- ACTION: concrete verb in base form (lemmatized), e.g., perform, analyze,
     design, coordinate, develop, maintain.
- OBJECT: direct object noun phrase; include essential noun modifiers
    when needed for meaning.
- ENABLER: how the task is done, with what tools/methods/sources (often "
    using...", "via...", "by...", "through...", "with...", "following...").

- CONTEXT: where/when/with whom/for whom; often trailing prepositional or
     clause modifiers.

Ordering and inclusion rules:
- For each task string, keep components in this order: ACTION, OBJECT,
    ENABLER, CONTEXT.
- Include ENABLER and CONTEXT only if supported by the job text.
- Each task must begin with the ACTION verb.

Extraction rules (apply all):
- Extract only duties the applicant/position will personally perform (e.g
    ., "the position will...", "responsible for...", "duties include...").
- Extract a task only if it is explicitly stated in the job advertisement;
     do not infer, assume, or add unstated duties.
- Exclude requirements/qualifications/preferences/skills framing (e.g., "
```

```
    knowledge of", "ability to", "preferred", "required", "seeks
    applicants with interests in...").
- Exclude mission statements, organisational goals, or broad purpose
    statements that describe the employer's aims rather than the position'
    s concrete duties (e.g., "further the mission of discovering
    therapeutics", "advance patient care", "support the strategic academic
     mission").
- Do not output composite tasks. A composite task is one that contains
    more than one distinct action or more than one distinct object. Split
    as follows:
  - If one sentence has multiple actions joined by "and" or listed in
    sequence, split into one task per action.
  - If one action governs multiple coordinated objects (a list), repeat
    the action and split by object.
  - If a task takes the form "do X to do Y" (action-purpose construction),
     always split into two separate tasks: "do X" and "do Y".
- Component propagation on splits:
  - When splitting a composite sentence, copy shared ENABLER and CONTEXT
    to each resulting task unless clearly item-specific.
- If both acronym and full-form are provided in the advertisement, use
    the full-form only in extracted tasks. Do not expand acronyms that
    appear only in abbreviated form in the advertisement.
- Do not use semicolons in task strings.
- Remove duplicates.

Atomicity examples:
- "design and analyze experiments" -> "design experiments", "analyze
    experiments"
- "perform PCR, reverse transcriptase PCR, and real-time PCR" -> "perform
     PCR", "perform reverse transcriptase PCR", "perform real-time PCR"
- "analyze patient outcomes to evaluate treatment effectiveness, using R,
     under IRB-approved protocols" -> "analyze patient outcomes using R
    under IRB-approved protocols", "evaluate treatment effectiveness using
     R under IRB-approved protocols"
  [always split on do-X-to-do-Y; ENABLER and CONTEXT propagated to both
    tasks]
- "document results to support chemical development" -> "document results
```

```
    ", "support chemical development"
- "plan, coordinate and report on market assessments" -> "plan market
    assessments", "coordinate market assessments", "report on market
    assessments"

## STEP 4: Sector Classification

A metadata block is provided below the job body as a starting hypothesis:
- `listed_company`: upstream-extracted employer name. May be exact, a
    parent/sub-campus, a recruitment agency, "Unclassified", or empty.
- `is_staffing_agency`: TRUE if upstream flagged this posting as agency-
    posted. Advisory signal -- your judgement from the body takes
    precedence.

**Step 4a: Verify the listed company**

Does `listed_company` accurately reflect the institution offering the job?


It does NOT accurately reflect the institution if any of:
- `listed_company` is a recruitment / staffing / contracting agency
    posting on behalf of a client.
- `listed_company` is "Unclassified", empty, or missing.
- The body explicitly names a different actual employer (e.g. `
    listed_company` is a parent corporation but the body names a specific
    subsidiary doing the hiring).

- If it DOES accurately reflect the institution: set `extracted_employer
    = null` (i.e. accept `listed_company` as-is). Proceed to Step 4b.
- If it does NOT: extract the actual employer from the body and put it in
     `extracted_employer`. If the actual employer is not named in the body
     (e.g. agency posting for "a leading pharmaceutical company"), set `
    extracted_employer = null` and proceed to Step 4b -- sector may still
    be derivable from context.

**Step 4b: Classify the actual employer into one of four sectors**
```

```
The "actual employer" is `extracted_employer` if non-null, otherwise `
    listed_company`.

- `academia` -- degree-granting institutions (universities, colleges,
    university medical centers).
- `government` -- federal/state/local agencies (NIH, NASA, FDA, DOE),
    national labs (incl. FFRDCs operated under contract), military
    research.
- `non_profit` -- non-degree-granting research institutes (Broad, Salk,
    HHMI/Janelia, foundation-funded research orgs).
- `industry` -- for-profit companies of any size.
- `unclassified` -- body genuinely lacks the information to judge between
     the four. Use sparingly.

## OUTPUT FORMAT
Return valid JSON only:
{
  "is_science_research": true,
  "final_title": "<standardised title>",
  "title_status": "validated" or "standardized_from_body" or "
    llm_generated",
  "title_error_type": "misclassification" or "multi_level" or null,
  "tasks": [
    "<task1>",
    "<task2>",
    ...
  ],
  "extracted_employer": "<actual employer per body, or null if
    listed_company is accepted as-is>",
  "sector": "academia" or "industry" or "government" or "non_profit" or "
    unclassified"
}

Notes:
- If no tasks can be extracted, use "tasks": []

---
```

```
Job Title: {title_name}
Listed Company: {company_name}
Is Staffing Agency: {company_is_staffing}

Job Body:
{job_body}
```

B.5 Task requirements — cognitive, psychomotor, physical, sensory-perceptual, interpersonal

Model: `gpt-5.4-mini`, reasoning effort `low` (OpenAI Batch API), three votes per task.
Input: the system message below; the user message is the task statement alone, as `Task: "<task statement>"`.
Output: a JSON object carrying a one-sentence rationale and five integer scores in $\{0, \ldots, 4\}$, one per dimension. The score used is the mean across the three votes.

```
You are an expert in occupational task analysis. You will be given a
    single work task statement. Score what performing this task, as it is
    typically carried out in a work context by a competent worker,
    requires of the person performing it.

Score the task on five requirement dimensions. Score each dimension
    independently: a task can score high on several dimensions, or on none.
     The example requirements listed for each dimension are illustrative,
    not exhaustive.

1. Cognitive requirement - mental and verbal information processing. For
    example: oral comprehension or expression, written comprehension or
    expression, mathematical reasoning, inductive reasoning, deductive
    reasoning.

2. Psychomotor requirement - coordinated movement and motor control. For
    example: manual dexterity, finger dexterity, reaction time, control
    precision.

3. Physical requirement - bodily strength, endurance, and gross movement.
     For example: static strength, dynamic strength, stamina, gross body
    coordination, body equilibrium.

4. Sensory-perceptual requirement - perceiving and discriminating stimuli
     in the environment. For example: far or near vision, night vision,
    sound recognition, depth perception, spatial awareness.

5. Interpersonal requirement - engaging with other people. For example:
```

```
    coordinating with other people, speaking publicly, managing other
    people's needs, negotiating.

Score each of the five dimensions on this scale:
0 = Not required
1 = Minimally required
2 = Moderately required
3 = Strongly required
4 = Essential or defining - this requirement is central to what the task
    is

Respond in JSON: first a one-sentence rationale, then the five integer
    scores.
```

B.6 Task requirements — codifiability

Model: `gpt-5.4-mini`, reasoning effort `low` (OpenAI Batch API), three votes per task.
Input: the system message below; the user message is the task statement alone, as `Task: "<task statement>"`. Codifiability is scored in a separate call from the five requirement dimensions.
Output: a JSON object carrying a one-sentence rationale and one integer score in $\{0, \ldots, 4\}$. The score used is the mean across the three votes.

```
You are an expert in occupational task analysis. You will be given a
    single work task statement. Score how CODIFIABLE the task is: the
    degree to which the method of performing the task could be written
    down, in advance, as explicit, ordered rules and instructions that
    cover the situations a competent worker would typically encounter.

The test: could a competent worker with the relevant general skills, who
    has never performed this particular task before, carry it out
    correctly guided only by the written rules and instructions? If yes
    for essentially all situations the task presents, the task is fully
    codifiable - routine. If correct performance instead depends on tacit
    knowledge - context-dependent judgment, adaptation to variable or
    unpredictable situations, expertise that cannot be fully articulated
    in writing - the task is non-codifiable - non-routine.

Codifiability concerns whether the method can be fully specified in
    advance. It is not about how frequently the task is performed, and not
     about how much skill or effort execution demands.

Score codifiability on this scale:
0 = Fundamentally non-codifiable - performance rests almost entirely on
    tacit knowledge, judgment, or creativity; no written procedure would
    get a newcomer to correct performance.
1 = Mostly non-codifiable - general guidelines could be written, but most
     situations require judgment beyond the rules.
2 = Partly codifiable - substantial parts follow explicit procedure, but
    key steps depend on tacit judgment.
3 = Mostly codifiable - explicit procedure covers standard cases;
```

```
    judgment is needed only for occasional exceptions.
4 = Fully codifiable - complete, explicit, ordered rules cover
    essentially all situations encountered in typical performance.

Respond in JSON: first a one-sentence rationale, then the integer score.
```

B.7 Human task duration

Model: `gpt-5.4-mini`, reasoning effort `low`, temperature 1.0 (OpenAI Batch API), two draws per task.

Input: the prompt below. `{examples}` is the anchor block — 100 survey-measured tasks per prompt, listed in ascending order of duration in the form `- "<task statement>" (<occupation>) - <hours> hours` under the heading reproduced at the foot of this note. `{context}` is the work-context line (the occupation title for O*NET tasks, the parent Level-1 and Level-2 areas for a science representative task) and `{target}` is the task being estimated.

Output: a single number, the estimated total active working time in hours. Draws are combined by geometric mean and then corrected onto the human scale.

```
Your task
You are an occupational analyst with experience conducting job analyses
    and time-and-motion studies across a wide range of industries and
    occupations. Your task is to estimate how long a particular work task
    takes to complete for the typical worker who performs it in the stated
     work context.

Instructions
Estimate the time required to complete one instance of this task from
    start to finish, observing the following rules:
- Assume an average worker with mid-level experience in the role (5-9
    years).
- Assume a typical instance of the task.
- Use the specific work context stated below.
- If the task is a composite (which may, but not always, be indicated by
    the word AND), give the total time to complete all components.
- If the task description mentions multiple examples of outputs (
    indicated by OR), provide the time to complete the most typical one.
- The task must be completed to a level of quality where it would be
    accepted by a manager or supervisor.
- Report the total active working time required, not calendar time. For
    example, if a worker spends 3 hours per day over 5 days on a report,
    the working time is 15 hours, not 1 week. Do not include time spent
    waiting for responses, handling interruptions, taking breaks, or
    working on other tasks.
```

```
- The allowed range is 1 minute to 4 weeks.
If the task describes an ongoing responsibility or recurring activity,
    estimate the time for one episode or one deliverable.
{examples}
Task to estimate
Work context: {context}
"{target}"

Respond with ONLY a single number: your estimate of the total active
    working time in hours. No units, no explanation, just the number.
```

The {examples} block is inserted verbatim as:

```

Examples
Here are work tasks with completion times measured from surveys of
    workers who perform them:
- "<task statement>" (<occupation>) - <hours> hours
...
```