Replace, augment, support: a nested task-level measure of generative-AI exposure

María Jimena Contreras* Jorge Galindo* Carles Gómez*
Manuel Hidalgo Pérez*†

Working paper. First version: 15 March 2026. This version: 13 November 2026

Abstract

This paper proposes a method to measure potential exposure to generative AI for every task in O*NET, the US occupational database. The result is three nested exposure levels, which can be used for applied work on labour markets. Each task is read in the context of its occupation. We first ask whether AI can *replace* its output at the quality of a typical worker (*Re*). If it cannot, we ask whether it can *augment* the worker enough to halve the time the task takes or to push its quality into the top quartile (*Au*), and if not, whether it can still *support* the task with some usable input (*Su*). Three frontier language models rate all 24,087 task-activity pairs of O*NET 31.0 (923 occupations), working from a dated inventory of 33 capability benchmarks, and a Delphi round settles their disagreements. Weighted by time, the average occupation in the catalogue has 37.8% of its working time replaceable, 48.4% augmentable and 79.2% supportable. We also measure three human anchoring factors (relational nexus, organisational responsibility and authorship) that keep a job's tasks tied to a person doing them, plus the share of working time in which AI helps an expert without replacing them.

Keywords: generative AI, task exposure, O*NET, large language model annotation, Delphi, complementarity.
JEL: J23, J24, O33.
Use of AI: We used AI tools to help draft and edit this paper and to write its code; the research design, data, analysis and conclusions are our own.

1 Introduction

Much of what we know so far about AI and jobs comes from occupational exposure indices. They underpin a first wave of studies that combine them with labour-force data to see which workers, occupations and countries are most exposed (Pizzinelli et al., 2023; Gmyrek et al., 2023; for Spain, Banco de España, 2026; Felgueroso et al., 2026; Rodríguez-Fernández, 2026; Mestre et al., 2025). Analysing technology's replacement potential through tasks is at least two decades older than generative AI. It comes from the task-based view of technological change,

*EsadeEcPol, Esade Center for Economic Policy.
†Universidad Pablo de Olavide (UPO).

where an occupation is a bundle of tasks and a technology may take over some of them and complement workers in others (Autor et al., 2003; Acemoglu and Autor, 2011), and from the first attempts to count how many jobs automation could reach (Frey and Osborne, 2017; Arntz et al., 2016).

Building on this task-based foundation, the method we propose yields three nested exposure levels of this kind for the whole O*NET catalogue. In O*NET 31.0, occupations are described by specific tasks, each linked to broader detailed work activities; this cross-classification produces 24,087 distinct task-activity pairs across 923 occupations. Each of these 24,087 pairs goes to three frontier language models, which answer three questions in sequence whose spirit can be summarised as: (i) Can AI replace the task's output at the quality of a typical worker? If not, (ii) can it make the worker much faster (half the time) or much better (top-quartile quality)? And if not, (iii) can it at least supply something the worker can use? The models judge against a dated inventory of 33 benchmarks, and a Delphi round resolves their disagreements. We add three human anchoring factors, which describe what keeps a job's tasks with the person doing them, and a measure of expert complementarity, the share of working time in which AI helps an expert without taking over.

In the average occupation, 37.8% of working time is replaceable, 48.4% augmentable and 79.2% supportable. Replacement varies with how far a task can be written down as a procedure. The replaceable share of task-activity pairs peaks where the task calls for judgement between defensible alternatives (50.5%) and drops to 13.7% where the judgement is tacit and expert knowledge is needed. Augmentation stays common in those expert tasks (47.7%).

The raters agree most on replacement and least on the two augmentation thresholds. Changing the task weights or relying on a single rater barely reorders occupations, and the rank correlation with the headline stays at 0.93 or more in every case. Levels are less stable. The three raters' scores for the same occupation differ by a median of 9.4 points, and by 25.1 or more for one occupation in ten, and we plan to publish that band with the scores. Our layers correlate at 0.78 to 0.90 with the measures of Eloundou et al. (2024) that count partial time savings, and less closely (0.35 to 0.56) with their measure of direct exposure. The observed use of Claude, from the Anthropic Economic Index, gives a one-sided check. The 19% of tasks we place outside all three layers receive 0.1% of that use, so AI is not being used at scale where we say it cannot help. In a blind human review of every pair of nine occupations drawn at random, agreement with the panel runs from 85.5% of pairs for support to 90.5% for augmentation.

The paper is organised as follows. Section 2 reviews the literature and Section 3 sets out four gaps in it, which Section 4 takes one at a time. Section 5 describes the data and the measurement, Section 6 the rater panel and Section 7 the aggregation to occupations. Section 8 presents the human anchoring factors and expert complementarity, Section 9 the validation and Section 10 the limitations. Section 11 concludes, and Appendix A reproduces the prompts and the condensed inventory.

2 Related literature

The task approach. Exposure measures are an application of the task approach to technological change. Technology acts on tasks, and occupations are bundles of them. In Autor et al. (2003), computer capital substitutes for workers in routine cognitive and manual tasks (those machines can accomplish by following explicit programmed rules) and complements them in non-routine problem solving and complex communication. Almost half of the effect they measure comes from changes in task content within occupations, a point worth keeping in mind for what follows. Acemoglu and Autor (2011) built a general framework on the difference between tasks, units of work activity that produce output, and skills, the capabilities workers bring to them. In it, workers of a given skill can switch tasks as technology and labour-market conditions change, and machines can take tasks over directly. The approach suits AI because a technology is defined by what it can do, and that is a statement about tasks. An occupation mixes tasks with very different exposure, so its score is an average that hides a composition, and the composition itself shifts as some tasks are taken over and others grow.

From occupations to tasks. Where to draw the line between routine and non-routine depended on what could be written down as rules. Frey and Osborne (2013, 2017) argued that machine learning and mobile robotics move that line, and went back to whole occupations, estimating a probability of computerisation for 702 of them. Together with a group of machine-learning researchers they hand-labelled 70 occupations as automatable or not (an occupation counted as automatable only if all its tasks were), and a classifier built on O*NET variables extended the labels to the rest. The variables are their indicators of three engineering bottlenecks to computerisation, perception and manipulation, creative intelligence and social intelligence. Their headline figure, 47% of US employment in the high-risk category (probability above 0.7), travelled far, even though the authors made no attempt to say how many jobs would actually be automated. One probability per occupation, however, assumes everyone in it does the same work. Arntz et al. (2016) took the same probabilities down to individual workers, using survey data on the tasks each person performs. The high-risk share in the US fell from 47% to 9%, and the average across 21 OECD countries was also 9%, because workers in the same occupation do different things and many of those things are hard to automate. The correction changed the unit but kept Frey and Osborne's judgement of what the technology could do, after which the task became the unit of choice.

Measures built for machine learning. The next generation went back to the technology and asked about AI specifically. Brynjolfsson et al. (2018) rated the 2,069 detailed work activities in O*NET for their suitability for machine learning, with a rubric applied through a crowdsourcing platform. Most occupations turned out to have at least some suitable tasks, and few if any had all of them. Felten et al. (2021) took ten AI applications tracked by the Electronic Frontier Foundation, asked crowd workers how related each one is to each of the 52 O*NET abilities, and aggregated abilities to occupations by their prevalence and importance; their index is explicitly agnostic about whether AI complements or replaces labour. Webb (2020) matched verb-noun pairs in patent titles to those in task descriptions and found that

AI, unlike software and robots, is aimed at high-skilled tasks. All three were built for the machine learning of the 2010s and work through proxies, ability links in one case and patent text in another. None asks directly whether AI can do a given task. Nor were they easy to update for language models. When Felten et al. (2023) kept only the language-modelling application in their index, the result still correlated at 0.979 with the original.

Generative AI and annotation by rubric. Large language models changed the picture because they reach language-intensive work that earlier automation had left alone. The response was to ask annotators, and soon the models themselves, whether a model could do a task. Our starting point in this line is Eloundou et al. (2023, 2024). In their rubric a task counts as exposed if access to a language model would cut the time needed to complete it by at least half at equivalent quality. It is E1 if the model alone does this through a chat interface, and E2 if it needs software built on top. An occupation's exposure is its share of exposed tasks, with E2 counted at zero, a half or one. The human labels (from the authors and from contractors used to reviewing model outputs) cover every detailed work activity and a subset of tasks, and an early version of GPT-4 labelled all task-occupation pairs. On their intermediate measure[1] about 80% of US workers are in occupations with at least a tenth of their tasks exposed, and 19% in occupations where more than half are. Their rubric describes the models of 2023 (text in and text out, inputs that fit in 2,000 words, no recent facts unless supplied, images only through add-on software). Later work took the approach to other countries. Gmyrek et al. (2023) had GPT-4 score each task of the 436 four-digit occupations of ISCO-08 and classified occupations into automation or augmentation potential by the mean and spread of their task scores; only clerical work came out as highly exposed. In a later refinement Gmyrek et al. (2025) combined task ratings from workers in Poland with expert review in Delphi-style rounds, and with language models that arbitrate between the two and extrapolate to all ISCO tasks. About a quarter of world employment ends up in exposed occupations.

Benchmarks, abilities, tasks. The European Commission's Joint Research Centre reaches tasks without annotating them (Tolan et al., 2021). Its framework links AI benchmarks to 14 cognitive abilities and, through them, to the tasks of occupations. The latest version has 352 benchmarks, including new ones for language and vision-language models, and 127 three-digit ISCO-08 occupations, and weights each benchmark by the research activity around it, so that exposure can be traced year by year from 2008 to 2024 (Casas et al., 2026). Its evidence carries a date, which is the idea we take up in Section 4.

Full-catalogue classification with model panels. Advanced LLMs with bigger context windows and bigger training samples made it feasible to classify the whole catalogue at its finest unit. Parshall and Lopez-Luzuriaga (2026) rate all 23,850 task-activity pairs of release 30.2, from 18,796 tasks in 923 occupations, each read within a profile of its occupation. Pairs are rated on three five-level ordinal axes. One is cognitive complexity, another is deployment difficulty (the sensing, locomotion or manipulation a task needs) and the third is regulatory

[1] Their β, which counts a task that needs software built on the model (E2) at one half.

restrictions. All three are anchored in a single question, what it would take for AI to cut a worker's time on the task by at least half, and exposure is read off thresholds on the axes. Three mid-tier models from Anthropic, OpenAI and Google rate every pair. When they disagree the pair goes to a Delphi-style round, in which each model rates it again after reading the panel's first answers; in their stability test (420 tasks, three independent runs) pairwise agreement went from 75.0% to 90.8%. They place 40.2% of US labour time in tasks that need at most contextual judgement, are purely digital and face no professional or statutory barrier, and they see the classification as a longitudinal instrument, to be re-run every six months. Several models matter because Yin et al. (2026) show that exposure scores from a single model are unstable across models.

Observed use. The newest work looks at where AI is actually used rather than where it could be applied in principle. The Anthropic Economic Index (Handa et al., 2025; Massenkoff et al., 2026) and the Copilot data of Tomlinson et al. (2025) map millions of real-world interactions with AI assistants directly to O*NET tasks and detailed work activities. Observing actual usage provides an invaluable empirical anchor, yet it inevitably represents a lower bound on potential: a task may be technically well within reach of frontier models and still remain unadopted due to institutional inertia, legacy workflows, or regulatory barriers. We use these usage data not as an exposure metric itself, but as an independent, one-sided validation check on our labels (Section 9).

3 Four gaps

It is our understanding that there are four areas in which we could contribute to filling gaps in this literature, and they motivate our design. They concern the unit measured, the question asked, who answers it, and when.

The unit of measurement. Measurement has moved down over time, from occupations to abilities and tasks. Only recently has it reached the task read in the context of each of its work activities, and done so for the whole catalogue (Parshall and Lopez-Luzuriaga, 2026). Aggregates average away the variation within occupations that pushed the field towards tasks in the first place (Arntz et al., 2016). They also hide what part of work is exposed, an occupation's score cannot tell whether AI reaches the routine paperwork or the judgement at the centre of the job.

The question asked. Most indices reduce the relation between a task and AI to one number (Felten et al., 2021; Webb, 2020; Eloundou et al., 2024; Gmyrek et al., 2023), which mixes substitution and complementarity. "AI can do this task" and "AI can halve the time this task takes" are different statements with different labour-market effects, yet both count as exposure (Acemoglu, 2025). The distinctions that do exist were designed for a weaker technology. E1 and E2 in Eloundou et al. (2023) differ in whether the model needs extra software and say nothing about how much of the task it does; in the authors' own validation, the halving threshold

tracked closely whether a model could do the core of a task or nearly all of it and agentic AI is now capable of designing its own software to solve the issue at hand. Gmyrek et al. (2023) infer automation or augmentation from the mean and spread of task scores within an occupation, without asking about each task. Now that models produce many outputs outright, one threshold can no longer separate doing a task from helping with it.

Who answers. Human judgement has two limits. It covers only part of the catalogue, and it rarely brings together people who know the work and people who know the technology. Frey and Osborne (2013, 2017) hand-labelled 70 of 702 occupations with a group of machine-learning researchers and let a classifier do the rest. The human annotators of Eloundou et al. (2023) labelled every detailed work activity but only some tasks, and the authors warn that, being familiar with language models but not occupationally diverse, they may misjudge tasks in occupations they do not know. Recruiting professionals from each occupation would reverse the difficulty, since they know their own work but most likely do not follow the AI frontier. Scale is a further constraint on any human panel, to rate all 24,087 pairs with three raters means 344,874 first-round judgements.[2] Language models make that scale feasible, but a single model's scores depend on which model it is (Yin et al., 2026).

The date. Potential exposure changes with the technology, and quickly; this year's models do things last year's could not. Yet most indices do not say which frontier they measure. That makes them hard to update in a comparable way, and when two indices differ we cannot tell whether the technology moved or the method did. With model raters this matters more, because each model knows the frontier only up to its own training cutoff. Two recent efforts go part of the way. The JRC framework dates its evidence by year, but its evidence is the research activity around benchmarks, which its authors present as a guide to where capabilities are likely to advance and not as a measure of current capability; tasks that AI already finds trivial attract little research and so score low (Casas et al., 2026). Parshall and Lopez-Luzuriaga (2026) plan to re-run their classification every six months, and leave for future work how to tell a change in capability from a change in the raters.

4 Our proposal

We take the gaps in the same order, and in each case start from the work that came closest. We then add what lies beyond exposure, which answers no gap in the literature but is needed to read the measure (Figure 1).

Unit. We use the unit of Parshall and Lopez-Luzuriaga (2026), the O*NET task read in the context of one of its detailed work activities. All 24,087 pairs of release 31.0 are classified, so nothing is sampled or extrapolated.[3] Before any question about AI, a first stage places each

[2] The count follows our design (Section 5). Each of three raters judges two task requirements and replacement on all 24,087 pairs, the two augmentation channels on the 15,143 pairs that are not replaceable, and support on the 12,411 that are not augmentable.

[3] Release 31.0 has 24,087 pairs from 18,838 tasks, against the 23,850 pairs from 18,796 tasks of the release 30.2 that Parshall and Lopez-Luzuriaga (2026) report; both cover 923 occupations.

pair on two ladders adapted from their axes of cognitive complexity and deployment difficulty. One is codifiability (C), how far the task can be written down as a procedure, which they themselves relate to the codifiability criterion of the task literature. The other is physical requirement (D), how much the task needs a body in the world. Section 5 gives the levels. The two ladders let us read exposure by kind of work (Section 7).

Question. Our starting point is the rubric of Eloundou et al. (2023), adapted to more capable models. Their halving threshold was set for the models of 2023, and a model that produces a task's output outright clears it too (Section 3). We therefore split the question into three, asked in a fixed order. *Replacement* (*Re*) asks whether AI can produce the task's output at the quality of a typical worker. *Augmentation* (*Au*) is asked only if the answer is no. It keeps the halving threshold as a time channel and adds a quality channel, which asks whether the output produced with AI in the same time would rank in the top quartile of outputs produced without it. *Support* (*Su*) comes last, where augmentation fails, and asks whether AI can still supply a usable input that belongs to the task. Asked in this order, the margins between layers are measured directly. Augmentation beyond replacement is what AI improves substantially but cannot do, and support beyond augmentation what it helps with only a little.

Raters. Using language models as raters is not new (Eloundou et al., 2023; Gmyrek et al., 2023; Parshall and Lopez-Luzuriaga, 2026). The JRC also uses them, under human review, to build its links between benchmarks, abilities and tasks (Casas et al., 2026). Model raters make it feasible to rate the whole catalogue several times over. We follow Parshall and Lopez-Luzuriaga (2026) in using three models from three developers, each rating every pair independently, with a Delphi-style round for the disagreements. We change two things. First, our raters are frontier models rather than the mid-tier ones of their production run, because in our calibration a smaller model cost agreement on the first-stage ladders (Section 6). Second, in our second round a rater sees the other raters' reasoning but not their labels, because a visible majority invites agreement with it rather than with the evidence. Model raters also answer the second limit of human judgement (Section 3). Annotators who follow the frontier misjudge work they have not done, and professionals misjudge what the frontier can do, so a human panel is biased in the direction of whichever group it draws from. A model rater belongs to neither group. It reads each task in the context of its occupation and takes the state of the technology from the dated inventory described next, so the two kinds of knowledge enter the same judgement, and we think that bias is gone. What remains is a dependence on which model rates, documented by Yin et al. (2026), so we report agreement for each axis and measure how much the scores depend on the rater (Sections 6 and 9).

Date. Here the inspiration is the JRC framework (Tolan et al., 2021; Casas et al., 2026). We share two of its ideas. Exposure should answer to an explicit and dated record of benchmark evidence, and the measure is updated by refreshing that record. The evidence and the route it takes to tasks are different, though. The JRC records research activity (papers per benchmark and year), and we record performance. Our capability inventory, a benchmark of

AI capabilities, lists 33 benchmarks with the best score reached by any model released by 31 August 2026 and the human baseline where one exists, and it tags each benchmark with the JRC abilities it draws on. Where the JRC evidence reaches tasks through fixed matrices linking benchmarks to abilities and abilities to tasks, ours reaches each task through the raters, who read the inventory and judge the task directly.

Within a wave the inventory puts the three raters on the same evidence, which matters because their training cutoffs are more than a year apart. Across waves, updating it and re-running the panel with the same prompts produces a series of potential exposure in which changes can be traced to the evidence and not only to newer raters. The series will not be perfect (the inventory has gaps, and O*NET keeps revising its task lists; see Section 10), but with the frontier moving this fast we would rather have an imperfect series than none.

Beyond exposure. Exposure is technical capability, and capability does not turn into substitution automatically. Part of what stands in between is whatever keeps a job's tasks bundled with the person doing them. Building on Pizzinelli et al. (2023) and Garicano et al. (2026), we measure three such anchoring factors from O*NET descriptors, and add a measure of expert complementarity taken from the panel itself. Both are reported next to exposure and never averaged into it (Section 8).

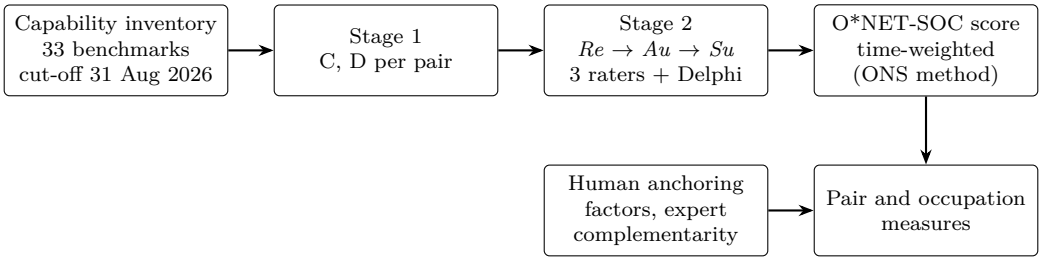


Figure 1: The measurement pipeline, from the capability inventory to occupation scores.

5 Data and measurement

5.1 Data

Task-activity pairs. O*NET 31.0 (August 2026; U.S. Department of Labor, Employment and Training Administration, 2026) links 18,838 tasks to 2,085 detailed work activities (DWA) in 923 occupations, which gives 24,087 pairs, and we classify all of them.[4] On average a task has 1.28 pairs, and 23.8% of tasks have more than one.

The capability inventory. The inventory has 33 benchmarks in 3 blocks (language, reasoning and long context; multimodal perception and generation; code, agents and professional deliverables). For each benchmark it records what it measures, the best score by a model

[4]The panel was first run on the previous release (29.x task list). Pairs that release 31.0 left unchanged keep those labels, which come from the same inventory, prompts and raters, so only the 999 new or reworded pairs had to be asked again. DWA identifiers are matched by title; the one title reworded between releases was matched by hand.

released on or before 31 August 2026 together with that model and its release date, the metric and its source, the human baseline if there is one, the abilities involved and the limits of the comparison. Before use, an independent audit checked each entry of the first version against its primary source (human baselines, model scores, release dates and references) and corrected one score and several codes and references; the entries added since were verified in-house. The raters see a condensed version, reproduced in full in Appendix A.1.

Occupation descriptors. Work Context, Work Activities and Task Ratings come from the same O*NET release. Task Ratings (relevance and frequency) give us the time weights, and Work Context and Work Activities the anchoring factors (Section 8.1).

5.2 Stage 1: what the task requires

The first stage assigns each pair two ordinal requirements, judged without any reference to AI and adapted from the ladders of Parshall and Lopez-Luzuriaga (2026). **Codifiability** (C) asks whether a complete how-to document would let a trained worker do the task correctly every time. The ladder goes from C0 (self-evident) and C1 (fully specifiable) through C2 (judgement between defensible alternatives) and C3 (tacit expert judgement that only a peer can check) to C4, where the task produces new, transferable knowledge. **Physical requirement** (D) asks whether an executor without a body could complete the task. D0 is purely informational and D1 needs presence to sense or move, while D2 to D4 involve manipulating objects in increasingly variable environments. Stage 1 also records which O*NET abilities and which AI capability families the task draws on. None of this assumes anything about exposure. All go into stage 2 as context, where the rater uses them to identify what the task demands, and the two ladders are what allows exposure to be read by kind of work. Of all pairs, 15.5% are C3 or above and 37.7% need manipulation (D2 or above).

5.3 Stage 2: three nested layers

Stage 2 asks the questions in Table 1. The rater imagines a typical worker in the occupation who has access to AI, to the usual tools of the task and to a laptop, and judges what frontier AI could do on 31 August 2026, meaning the models released by then and the software and agents built on them. Evidence comes from the inventory first. Background knowledge is allowed only where the benchmarks leave the question open; the rater has to flag it, and it cannot reach past the cut-off.

Nesting. By design the layers are nested. A task that AI can do entirely is one where the worker's productivity rises, and a large productivity gain means AI is supplying something the task can use. So replacement implies augmentation, augmentation implies support, and each question is asked only where the previous answer was no. The published value of each layer includes the layers below it. Counting every replaceable task as augmentable is generous (once review and correction of model output are counted, measured speed-ups on professional deliverables can be much smaller, Patwardhan et al., 2025).

Table 1: The three layers of exposure

Layer	Question put to the rater	Asked on
Re replacement	Could AI produce this task's output at the quality of a typical worker in the occupation?	all pairs
Au augmentation	*Time*: does access to AI cut the worker's time on the task *as a whole* by at least half, at reference quality? *Quality*: in the same time, would a peer grading blind rank the AI-assisted output above three of every four unassisted outputs? $Au = \text{time} \vee \text{quality}$.	$Re = 0$
Su support	Can AI perform a distinct, task-specific action normally required for the output or as a direct input to its decisions, incorporable without redoing it, with the outcome at typical quality? No threshold.	$Au = 0$

An example. Figure 2 follows five tasks of chefs and head cooks through the questions, one for each possible outcome. Recording production data on forms is replaceable. Arranging equipment purchases or repairs is not, though AI halves the time it takes. Deciding how food should be presented gains on quality instead. Checking the products that come in gets support only, and cooking itself is outside all three layers. Summed over its 21 tasks with their time shares, the occupation comes out 41.5% replaceable, 52.7% augmentable and 95.5% supportable.

Why these thresholds. Both augmentation thresholds are high on purpose, so that modest gains do not count. Halving the time a task takes at equivalent quality is the threshold behind exposure in Eloundou et al. (2023), shared by their labels E1 and E2. They admit it is somewhat arbitrary, and say they chose a relatively high bar because they expected real time savings to fall short of their estimates. It is also above most measured effects of AI assistance. In customer support, AI raised resolutions per hour by about 15% (Brynjolfsson et al., 2025); consulting tasks took about a quarter less time (Dell'Acqua et al., 2023), and professional writing 40% less (Noy and Zhang, 2023). Only a well-specified coding task saw more than a halving (Peng et al., 2023). On quality, blind grading by professionals from the same occupation found gains of about 0.4 standard deviations (Noy and Zhang, 2023). Under normality that puts the median assisted output near the 66th percentile of unassisted ones, short of the 75th our quality channel asks for.

How the inventory enters the judgement. The rubric asks raters to read each benchmark for how it is built and not just for its score. A benchmark is positive evidence only if it is at least as demanding as the task, and a task that no benchmark covers, or a benchmark with no human baseline, counts as evidence in neither direction. Physical tasks have a test of their own. Observation that commodity hardware can capture (cameras, microphones, drones) is treated as interpretation, whereas presence and manipulation need evidence of deployment in environments at least as unstructured as the task's. Scores obtained in sandboxes or on multi-step tasks are checked against the autonomy horizon documented by Kwa et al. (2025).

What exposure leaves out. Regulation, legal responsibility and a preference for a human provider do not make a task technically impossible, and the rubric ignores them. We treat them

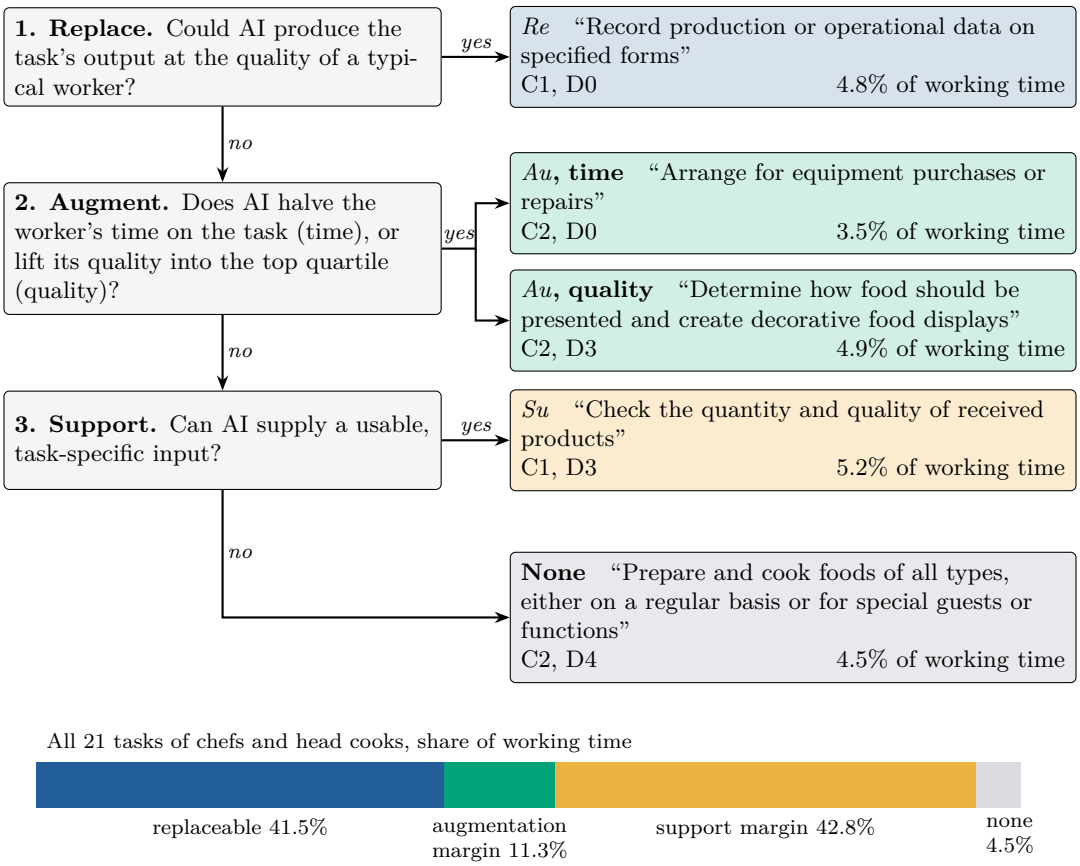


Figure 2: How the three nested questions split the tasks of one occupation. Top: five of the 21 tasks of chefs and head cooks, one per outcome, with their stage-1 levels of codifiability (C) and physical requirement (D) and their share of the occupation's working time (ONS time weights, Section 7). Each task here has a single detailed work activity. Bottom: the occupation as a whole.

as frictions between capability and substitution and deal with them elsewhere. Responsibility and the preference for a person enter the anchoring factors (Section 8). Regulation is national, so we leave it to applied work on national classifications.

6 The rater panel

Raters. Every pair is rated blind and independently by three frontier models from different developers: Claude Opus 5 (`claude-opus-5`, Anthropic), GPT-5.6 Sol (`gpt-5.6-sol`, OpenAI) and Gemini 3.1 Pro (`gemini-3.1-pro-preview`, Google). These were the state-of-the-art models during the last iterations of the project, from June to September 2026. Earlier iterations, from the start of the project, tried other models, such as Gemini Flash, Claude Sonnet and earlier versions of Claude Opus, and their results were less stable. Pairs go out in batches of twenty, and for each one the rater returns a label, a short reasoning, the benchmarks it relied on and its confidence. There is reasonable evidence for language models as annotators. They match or beat crowd workers on text annotation (Gilardi et al., 2023), and a panel of diverse models judges better than a single one (Verga et al., 2024). We do not use only the most capable model, because exposure scores from a single model are unstable across models (Yin et al., 2026). Section 9 shows how much each of ours moves the result.

Reasoning effort. All three raters run at low reasoning effort (effort `low` for Claude, reasoning effort `low` for GPT, thinking level `low` for Gemini). The setting came out of calibration rounds that crossed model tier with effort on a fixed set of 1,000 pairs, which we deliberately drew from cognitively demanding management occupations. Higher reasoning effort did not produce substantially different results.

Delphi reconciliation. We use the Delphi-style round (Dalkey and Helmer, 1963) of Parshall and Lopez-Luzuriaga (2026), with our own raters, prompts and questions. A pair that is not unanimous on an axis goes back to each rater in a fresh context, together with the reasoning of the other two. Their labels are hidden so that none can sway the second answer, and the few reasonings that named their own level in the text are masked as well. In the replacement round each reasoning also carries the documented training cutoff of the model that wrote it, and when a disagreement rests on background knowledge the more recent cutoff counts for more. The final label is the second-round majority. Without a majority the rule depends on the axis. On the two ladders, where three raters can give three different levels, a calibrated tie-break takes the label of the rater that agreed most often with the majority in the two-to-one splits of the first round, computed within each of the six blocks in which the catalogue was processed. On the binary questions a tie can only arise when a vote is missing, and we resolve it conservatively, as a no. Pairs with fewer than two valid votes are flagged as unresolved.[5]

[5]In our runs, models dropped rows, by leaving out a label or mistyping an identifier, and the loss goes unnoticed unless we count rows. We check every batch against its request file before parsing anything, and a rater that comes back short is asked again for the missing rows only, a limited number of times. On replacement, 15 second-round votes were still missing when the retries ran out, each answered under a mistyped identifier. In 12 of those pairs the two remaining raters agreed, so the missing vote could not change the label. In the other three they disagreed, and the pair is labelled not replaceable. Final labels go into a bank keyed by a hash of the

Agreement. Table 2 gives first-round agreement on the full run, both as unanimity and as chance-corrected agreement (Cohen, 1960; Fleiss, 1971). On every axis, at least two thirds of the pairs are unanimous in the first round. Corrected for chance, agreement is highest on replacement and physical requirement. On replacement, Fleiss' κ is 0.79 (85.3% of pairs unanimous, with positive rates between 37.1% and 42.4% depending on the rater), and on physical requirement 0.71. It is lowest for the two augmentation thresholds (Fleiss' κ of 0.34 for time and 0.43 for quality), where the raters also differ most in how often they say yes, and for support (64.6% unanimous, with positive rates from 38.0% to 65.9%). We reconcile the two augmentation channels separately, and report them apart and together, because agreement on them is lowest.

What the Delphi round changes. Of the 3,404 replacement pairs that reached the second round split two to one, 22.4% ended with a label different from the first-round majority. For augmentation the shares are similar (21.7% on time, 25.4% on quality), and so they are for codifiability (23.1%) and physical requirement (23.9%). Support has the highest, at 34.4%.

Table 2: Inter-rater agreement and Delphi reconciliation, full run

			Round 1			Delphi round		
Axis	Pool	Pairs	Unanimous (%)	Fleiss κ	Cohen κ range	Sent (%)	Converged (%)	Tie-breaks
C codifiability	all pairs	21,011	66.1	0.64	0.62 to 0.65	33.9	53.5	53
D physical requirement	all pairs	21,011	71.3	0.71	0.70 to 0.73	28.7	54.0	182
Re replacement	all pairs	23,233	85.3	0.79	0.78 to 0.81	14.7	46.6	3
Au time channel	Re = 0	14,572	77.7	0.34	0.24 to 0.47	22.3	64.7	0
Au quality channel	Re = 0	14,572	78.0	0.43	0.39 to 0.47	22.0	46.0	0
Su support	Au = 0	11,956	64.6	0.53	0.45 to 0.64	35.4	42.4	0

Full run on the 29.x task list, before the 31.0 update. Stage 1 excludes the pairs carried over from the 50-occupation calibration run. "Sent": share of pairs not unanimous in round 1 and sent to the Delphi round. "Converged": share of those on which all three raters agree in the second round. Cohen's κ range over the three pairs of raters. On the binary axes a tie-break arises only when one rater's second-round answer is still missing after the retries and the other two disagree, and it is resolved as a no.

7 From pairs to occupations

Pair to task. A task can appear under several detailed work activities, so its score is the mean over its pairs, i.e. the share of its activity readings that are positive. As Eloundou et al. (2023) do with their human labels, we aggregate to tasks before going to occupations. A task's weight is split equally across its pairs and never multiplied by them. Tomlinson et al. (2025), by contrast, give a task's full weight to every activity it maps to.

Task to occupation: time weights. Within an occupation, tasks are weighted by their share of working time, following the method of the UK Office for National Statistics (Martin and Monahan, 2022). Each task's relevance RT (the share of incumbents who rate the task as relevant to their job) and its frequency distribution FT_c over seven categories are turned

rubric and the models, so a pair is never asked twice and a label is reused only under identical instructions.

into a share of working time. With the implicit "never" category at weight zero, the revised frequency share is $FT_c \cdot RT/100$, and the composite time score is

$$s_t = \sum_{c=1}^{7} \frac{FT_c\, RT}{100}\, w_c, \qquad w = (0.02,\, 0.03,\, 0.06,\, 0.08,\, 0.10,\, 0.13,\, 0.16), \tag{1}$$

from "yearly or less" to "hourly or more". The time share of task t in occupation o is $\pi_t = s_t / \sum_{t' \in o} s_{t'}$, and the occupation's score on each layer is

$$L_o = \sum_{t \in o} \pi_t\, \bar{L}_t, \qquad L \in \{Re, Au, Su\}, \tag{2}$$

that is, the share of its working time that is replaceable, augmentable or supportable. Where O*NET gives a task no frequency rating, we give it an equal share of its occupation's time, and an occupation with no frequency ratings at all is equal-weighted throughout.[6]

Other weights. The ONS estimated its frequency weights by regression and itself calls them fairly arbitrary (Martin and Monahan, 2022). We therefore compare them with three alternatives in Table 3: annual occurrence counts (Tomlinson et al., 2025), task importance and equal weights. Means barely move (by less than a point), and neither does the ranking by *Re* (rank correlation with the headline of 0.974 or more). A single occupation can move more, by up to 27.4 points of its working time. We keep the ONS time shares as the headline because they are the only scheme whose result can be read as a share of working time.

Whatever the weights, scores depend on how O*NET writes each task list. O*NET lists 21 tasks for chefs and head cooks. Cooking is one of them and gets 4.5% of working time, and most of the rest are supervisory or administrative, so the occupation's scores are best read alongside the task list behind them.

Table 3: Task weighting schemes, 923 occupations

Weighting	*Re* mean (%)	*Au* mean (%)	*Su* mean (%)	Rank corr. of *Re* with headline	Largest *Re* shift (points)
ONS time shares (headline)	37.8	48.4	79.2	1.000	0.0
Annual occurrence counts	37.5	47.8	79.4	0.974	27.4
Task importance	37.1	48.1	78.8	0.988	19.4
Equal weights	37.0	48.0	78.5	0.983	19.0

Means across occupations. All schemes give tasks without ratings an equal share within their occupation.

Exposure across O*NET 31.0. Table 4 has the means and Figure 3 shows what lies behind them. Of the 15,143 pairs that are not replaceable, 18.0% are augmentable (7.6% through the time channel and 15.5% through quality), and of the 12,411 that are not augmentable, 60.5% are supportable. Across occupations, replacement runs from 9.3% of working time at the tenth percentile to 70.3% at the ninetieth, with a median of 35.5%. Where it is low but the work is informational, augmentation and support cover most of what is left. The least

[6] 418 tasks (2.2% of the catalogue) lack a frequency rating. They sit in 38 occupations with partial coverage and 12 with none.

Table 4: Exposure across O*NET 31.0

Layer	Per pair (%)	Time-weighted, mean of SOC (%)	Median SOC (%)	SOC with more than half of time
Replacement (*Re*)	37.1	37.8	35.5	259 (28%)
Augmentation (*Au*)	48.5	48.4	48.8	447 (48%)
Support (*Su*)	79.7	79.2	91.5	763 (83%)

24,087 pairs, 923 occupations. Layers are cumulative ($Re \subseteq Au \subseteq Su$). Counting only the time (quality) channel, *Au* covers 42.0% (46.9%) of working time.

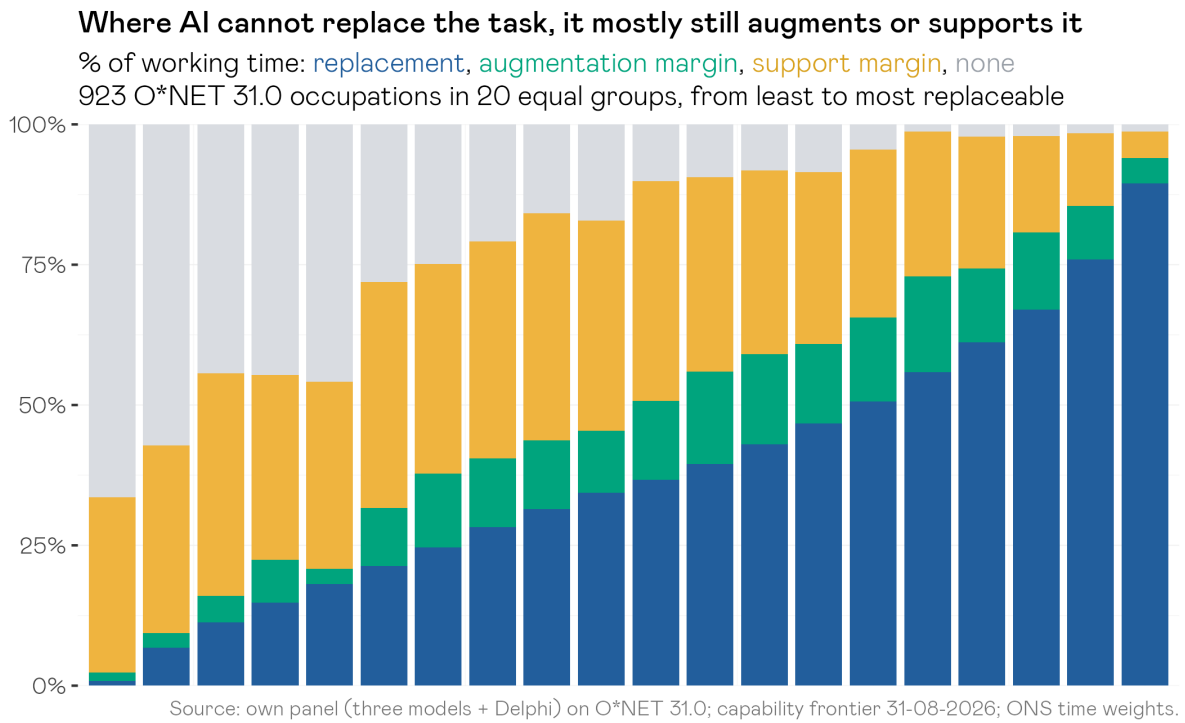


Figure 3: Time-weighted layers across occupations. Occupations sorted by their replaceable share and grouped into 20 bins of equal size; each bar is the bin mean.

replaceable quarter of occupations accounts for 63% of all the working time outside the three layers (summed across occupations, not weighted by employment), and 79% of its task-activity pairs need manipulation, compared with 8% in the most replaceable quarter.

Exposure by kind of work. The two ladders of stage 1 show where each layer falls (Tables 5 and 6). Replacement drops from 70.5% of purely informational pairs (D0) to 14.1% when presence is needed (D1) and 0.5% when manipulation is (D2), while 59.1% of D2 pairs are still supportable. On the codifiability ladder, replacement peaks at C2 (50.5%) and drops at C3 (13.7%), where 47.7% of pairs are still augmentable.

Table 5: Layers by physical band (%)

D	Pairs	Share	*Re*	*Au*	*Su*
D0	12,067	50.1	70.5	85.3	99.9
D1	2,935	12.2	14.1	32.9	90.9
D2	3,751	15.6	0.5	6.5	59.1
D3	4,673	19.4	0.0	3.6	44.4
D4	661	2.7	0.0	0.0	25.7

Table 6: Layers by codifiability band (%)

C	Pairs	Share	*Re*	*Au*	*Su*
C0	54	0.2	11.1	11.1	22.2
C1	10,700	44.4	33.7	35.6	62.6
C2	9,596	39.8	50.5	62.8	93.4
C3	3,518	14.6	13.7	47.7	93.7
C4	219	0.9	0.0	74.4	100.0

8 Beyond exposure: human anchoring factors and expert complementarity

Exposure measures technical capability. Whether that capability ends up replacing workers depends on more than which tasks AI can do, and the literature has looked at what lies in between from two angles: complementarity indices built from occupational descriptors, and models of jobs as bundles of tasks. We borrow from both and add a measure of our own.

Complementarity indices. Pizzinelli et al. (2023) start from the point that occupational exposure indices are "purposely framed as agnostic on the likelihood of AI complementing or replacing labor". On top of the ability-based index of Felten et al. (2021) they build an occupation-level index of potential complementarity, θ, from O*NET data. It has eleven work-context items in five groups (communication, responsibility, physical conditions, criticality, routine), plus the job zone, which reflects how much education and training an occupation requires. Each group is a simple mean of its items and θ is the mean of the six components. Exposure is then discounted by complementarity, C-AIOE = AIOE $\times$ $(1 - (\theta - \theta_{\min}))$, and occupations are split at the median on both dimensions. Highly exposed occupations with high complementarity are where AI is more likely to support workers, and those with low complementarity are more at risk of substitution. Cazzaniga et al. (2024) take the index across countries. By their count, about 40% of workers worldwide and 60% in advanced economies are in occupations above the median exposure. In the average advanced economy, 27% of employment is in high-exposure, high-complementarity occupations and 33% in high-exposure, low-complementarity ones.

The index was an important step. It treats what ties work to people as a separate dimension, driven in their words by social, legal and technical factors that are independent of exposure

itself, and measures it with O*NET descriptors of the context of work. For what we need, though, it leaves four things unresolved. Item selection, as the authors say, "relies on our own judgement". Its positive association with exposure comes from the job zone. In their own tables the skills component correlates at 0.72 with exposure, and without it the correlation between θ and exposure goes from 0.17 to -0.21, which suggests that the index partly measures education. Components are averaged with equal weights (a cautious choice, the authors say, given that none of them "is necessarily applicable to all occupations"), so a job tied to people through one strong channel and weak on the rest lands in the middle. And because θ works as a discount, it tells us how much of an occupation's exposure to set aside but not why a person stays in the job; once the discount is applied the two dimensions cannot be pulled apart again.

Jobs as bundles. Garicano et al. (2026) argue that task exposure is an incomplete guide "because labor markets price jobs, not tasks". Jobs bundle tasks, and what AI does to a job depends on how costly the bundle is to break. Exposure measures do show where AI is likely to matter, but they tend to treat exposed tasks as if they could be reallocated one by one. In the model, AI can either assist with a task inside the job or supply it autonomously while a human does the rest, and splitting tasks across agents destroys part of the value of the service. That coordination cost is what makes a bundle strong or weak. In weak bundles AI takes over the separable task, the job narrows and the worker's share of revenue falls; if product demand is inelastic enough, employment falls too. In strong bundles AI improves performance within the job without removing the human from it. Two occupations with similar exposure can therefore end up in very different places. The model is theoretical, and the protection it describes is conditional, since fewer occupations stay fully bundled as AI improves. The authors leave the coordination cost in reduced form, name some likely determinants (shared context, accountability and liability for the full service, communication frictions, spillovers across tasks) and suggest classifying occupations by the strength of their bundles, which they do not attempt themselves.

From there to our measures. We do not attempt a complete measure of bundle strength. Our aim is more modest, to see how far the descriptors that O*NET already collects can take us. We read three features of occupations as observable determinants of bundle strength, and call them *human anchoring factors*. Organisational responsibility is closest to their accountability. Someone has to answer for the whole result, and in Garicano et al. (2026) that is what makes a job costly to split; we measure it through the consequence of error, the discretion to decide and authority over other workers. The relational nexus covers work that consists of dealing with people, where the relationship and the friction of handling it hold the tasks together. It is our reading, for relational work, of their shared context and communication frictions. Friction alone would rank how unpleasant a job is (massage therapists and barbers, whose clients come for the person, would land in the bottom quarter), so we add a presence channel, for jobs in which the client buys the person and not just the output. Authorship covers work whose value lies in who conceived it, through creative thinking and the freedom to set one's own tasks and goals. It echoes the creative intelligence that Frey and Osborne (2013, 2017) list among

the engineering bottlenecks to automation. The mapping from these determinants to O*NET descriptors is our own.

From Pizzinelli et al. (2023) we keep the practice of building such measures from O*NET work contexts and activities, with no model annotation, and five of the eleven work contexts in θ also enter our factors: consequence of error, freedom to make decisions, responsibility for the outcomes and for the health and safety of other workers, and freedom to determine tasks, priorities and goals (which θ reads as unstructured work). We change three things. The job zone is left out, so that the factors describe how a job is tied to a person and not the education it requires. The factors are kept apart, and within each one the channels are combined with a power mean close to the maximum (Section 8.1). And they are reported beside exposure, never as a discount on it.

8.1 Three human anchoring factors from O*NET descriptors

We compute the three factors from eighteen O*NET descriptors, without any model annotation (Table 7). Each descriptor is rescaled from its 1-to-5 scale to $[0, 1]$ as $(x - 1)/4$. Channels that are alternative routes to the same factor are combined with a power mean,

$$M_p(x_1, \ldots, x_k) = \Big(\frac{1}{k}\sum_{i=1}^{k} x_i^p\Big)^{1/p}, \qquad p = 4, \tag{3}$$

which stays close to the maximum. One strong channel should be enough to tie the work to a person, and we do not want a weak one to average it away.[7]

- *Relational nexus.* Friction in volume, $\text{FRIC} = F \times V$, with

  $$F = \tfrac{1}{2}\,\text{conflict} + \tfrac{1}{2}\max(\text{unpleasant},\ \text{violent})$$

  and V the mean of the three items on contact with others, inside or outside the organisation. The two items on difficult people enter through their maximum, since violent behaviour is an extreme case of unpleasant behaviour and adding them would count the same episodes twice. Because it is a product, friction without volume stays low, and so does volume without friction. Presence, PRES, is the mean of the three items on what the client buys from the person. $\text{Nexus} = M_4(\text{FRIC}, \text{PRES})$.

- *Organisational responsibility.* The cost of nobody answering for the result: $\text{Resp} = M_4(\text{CRIT}, \text{DISC}, \text{AUT})$, where criticality is the consequence of error, discretion the mean of the two items on decisions, and authority the mean of the three items on other workers.

- *Authorship.* $\text{Auth} = \frac{1}{3}(\text{determine} + \text{creative} + \text{competition} \times \text{creative})$. Competition enters only through creativity, so that commercial pressure on its own does not raise the index.

[7] With $p = 1$ the power mean is the plain average, and as p grows it approaches the maximum. We take $p = 4$ so that an occupation strong on a single channel ranks high without the other channels being ignored, as they would be under the maximum. The ordering of the nexus hardly depends on this choice. Against $p = 4$, its rank correlation is 0.993 with $p = 2$, 0.996 with $p = 8$ and 0.990 with the maximum. Responsibility depends on it more (0.979, 0.973 and 0.887), because its three channels differ more within an occupation. Administrative law judges, who are strong on discretion alone, are at percentile 42 with $p = 2$, 75 with $p = 4$ and 90 with $p = 8$.

Items within a channel measure the same thing and are averaged; channels are alternative routes to the same factor and are combined with M_4. Authorship is an average because its three terms describe a single profile rather than alternative routes. Each factor is also expressed as a percentile across the 923 occupations, because the three indices have different ranges and percentiles can be compared.[8]

Table 7: O*NET descriptors behind the three human anchoring factors

Factor	Channel	O*NET descriptor	Domain
Nexus	Friction F	Conflict Situations	WC
		Dealing With Unpleasant, Angry, or Discourteous People	WC
		Dealing with Violent or Physically Aggressive People	WC
	Volume V	Contact With Others	WC
		Deal With External Customers or the Public in General	WC
		Work With or Contribute to a Work Group or Team	WC
	Presence	Performing for or Working Directly with the Public	WA
		Assisting and Caring for Others	WA
		Establishing and Maintaining Interpersonal Relationships	WA
Responsibility	Criticality	Consequence of Error	WC
	Discretion	Freedom to Make Decisions	WC
		Impact of Decisions on Co-workers or Company Results	WC
	Authority	Coordinate or Lead Others in Accomplishing Work Activities	WC
		Work Outcomes and Results of Other Workers	WC
		Health and Safety of Other Workers	WC
Authorship		Determine Tasks, Priorities and Goals	WC
		Thinking Creatively	WA
		Level of Competition (only × Thinking Creatively)	WC

WC: Work Context (context scale). WA: Work Activities (importance scale). Both rescaled from a scale of 1 to 5 to $[0, 1]$.

8.2 Expert complementarity

The anchoring factors describe the job. Our fourth measure comes from the tasks, and from a pattern along the codifiability ladder that Section 7 documents (Table 6). Replacement peaks at C2 and falls at C3 and C4, augmentation does not fall in the same way, and support stays close to universal from C2 up. AI is least able to do the work in tasks that rest on tacit expert judgement or produce new knowledge, but can still help with them.

So we measure that region directly. *Expert complementarity* is the share of an occupation's working time spent on task-activity pairs that require tacit expert judgement or produce new knowledge (C3 or C4), are not replaceable ($Re = 0$) and are still supportable ($Su = 1$). It is the

[8] In O*NET 31.0, 12 occupations have none of the eighteen descriptors, and three model raters estimate them. Each rater gets the item definitions with the descriptions of their scale points, four reference occupations with their tasks and observed ratings, and the title and tasks of the target occupation. Two of the references cover the low, middle and high bands of the items and two are the closest occupations by work content, and every estimate has to name the reference it is placed against. The final value is the mean of the three estimates. We flag an item for review when the estimates span two points or more, or when a rater reports low confidence (48 of the 216 adopted items). In a pilot on three occupations held out with their 54 observed scores, the mean absolute error of the panel on the 1-to-5 scale was 0.29 with this mixed reference set, 0.30 with the four closest occupations and 0.34 with four band-covering ones, against 0.31 for the plain mean of the closest references. That is about a third of a scale point, although three occupations are too few to rank the methods. Rows built on estimates carry a source flag. The procedure matters less for these 12 occupations than for future O*NET releases, which may add occupations before their descriptors are collected.

work in which AI helps an expert without taking it over. It is aggregated like the layers, from pairs to tasks and then from tasks to occupations with the time weights of Section 7. Because it rests on the C3 and C4 labels, it inherits the lower agreement on codifiability (Table 2). This measure is ours and should not be read into Garicano et al. (2026). In their model AI never touches the task that keeps the human in the job, whereas expert complementarity looks precisely at where AI enters a job alongside expert judgement, whatever holds the job together.

Pairs and occupations. Of the 3,737 pairs at C3 or C4, 81.1% count as expert complementarity, 12.9% are replaceable and 6.0% fall outside support. In practice, then, the measure is close to the time an occupation spends on expert tasks. The average occupation spends 15.0% of its working time at C3 or C4, and 81% of that time is expert complementarity. Nearly all of it comes from C3, as C4 tasks add only 0.7 points on average. The other 19% is expert work that AI replaces or cannot support, which is why we keep the two conditions. In legal occupations it is the difference between 44.7% of working time at C3 or C4 and 31.5% of expert complementarity (Table 8). Expert complementarity is zero in 43% of occupations; its median is 3.8% of working time, its mean 12.1% and its 90th percentile 36.9%. It covers more than a quarter of working time in 179 occupations and more than half in 41. Science, medicine and engineering dominate the top: biochemists and biophysicists (80%); physical medicine and rehabilitation physicians (76%); astronomers (74%); nanosystems engineers (71%); animal scientists (69%).

Major groups. By major group (Table 8) it goes from 31.5% of working time in legal occupations down to 0.8% in office and administrative support, and 11 of the 22 groups are under 5%. High expert complementarity does not mean low replacement. Legal, scientific and engineering occupations have both a large expert share and above-average replacement (55.6%, 41.9% and 45.6% of working time, against a mean of 37.8%), because their routine documentation and analysis can be replaced while their core judgement cannot. Healthcare practitioners have a similar expert share and below-average replacement (32.0%).

An example: family medicine physicians. 44.2% of the working time of family medicine physicians is expert complementarity, the same as the time they spend on tasks at C3 or C4, while 29.9% is replaceable and 96.2% supportable. The administrative side of the job is replaceable (recording patient histories, ordering tests, referring patients, preparing official health records). Immunising patients is a physical act and falls outside all three layers. The core of the job is analysing test data to inform a diagnosis, prescribing treatments and medications, monitoring patients and re-evaluating treatment. The raters place it at tacit expert judgement and find that AI cannot replace it but can halve the time it takes or raise its quality. Supervising care staff and coordinating with other professionals are expert tasks as well, though there AI only offers support. Expert complementarity adds up this core, weighted by the time it takes.

Table 8: Expert complementarity by SOC major group

Major group (SOC 2018)	SOC	Expert compl. (%)	Time at C3 or C4 (%)	*Re* (%)	*Su* (%)
Legal (23)	7	31.5	44.7	55.6	99.6
Life, physical and social science (19)	60	29.4	36.8	41.9	96.6
Community and social service (21)	14	26.8	29.6	39.8	96.2
Healthcare practitioners and technical (29)	89	26.5	33.9	32.0	86.7
Architecture and engineering (17)	56	22.9	27.4	45.6	97.8
Educational instruction and library (25)	62	20.9	23.0	37.1	94.1
Management (11)	56	17.6	19.1	48.9	98.8
Arts, design, entertainment, sports and media (27)	40	14.7	21.3	44.6	88.2
Computer and mathematical (15)	36	13.5	18.5	71.3	99.8
Protective service (33)	26	12.6	18.5	37.0	77.9
Business and financial operations (13)	48	12.1	15.4	65.5	99.2
Sales and related (41)	22	4.6	5.0	59.2	91.5
Transportation and material moving (53)	52	4.0	4.5	32.4	66.5
Personal care and service (39)	31	3.6	4.6	30.5	69.7
Farming, fishing and forestry (45)	12	2.8	4.3	19.0	66.3
Installation, maintenance and repair (49)	50	2.2	3.4	17.4	58.8
Healthcare support (31)	19	1.7	2.5	29.1	60.9
Production (51)	107	1.5	2.2	21.6	54.3
Food preparation and serving (35)	16	1.4	1.6	18.5	52.2
Building and grounds cleaning and maintenance (37)	8	1.4	1.4	23.3	54.0
Construction and extraction (47)	61	1.0	1.3	13.2	40.3
Office and administrative support (43)	51	0.8	0.9	65.2	92.1

Unweighted means across the O*NET-SOC occupations of each major group; shares of working time with ONS time weights. Sorted by expert complementarity.

8.3 How the measures relate

Table 9 reports the full matrix of rank correlations across the 923 occupations for the three layers, expert complementarity and the three anchoring factors. The layers are strongly correlated with each other (ρ from 0.71 to 0.91), as one would expect given the nesting. Relational nexus and organisational responsibility are almost unrelated to any layer ($|\rho| \leq 0.13$), so they carry information that exposure does not, and they correlate with each other at 0.35. Authorship is different. It rises with all three layers, more so the broader the layer (0.28 with *Re* and 0.50 with *Su*), and most of all with expert complementarity (0.59). Expert complementarity is only weakly related to replacement (0.18) and more to support (0.56), which fits its definition as work AI reaches without taking over. We left the job zone out of the factors so that they would not measure education. For two of them it works, and the rank correlation with the job zone is 0.18 for the relational nexus and 0.07 for organisational responsibility. Authorship is the exception, at 0.65 (expert complementarity is at 0.74), so it partly reflects the education a job requires. Since none of the four measures simply mirrors exposure, or each other, we report them as a vector next to exposure and do not combine them into one index.

9 Validation

We check the measure in four ways and then ask how much it depends on the rater. The four checks look at what the rubric implies for different kinds of work, at the task-level index of Eloundou et al. (2024), at the observed use of Claude reported by the Anthropic Economic

Table 9: Rank correlations among exposure layers, expert complementarity and human anchoring factors, 923 occupations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Replacement (*Re*)	1						
(2) Augmentation (*Au*)	0.91	1					
(3) Support (*Su*)	0.71	0.82	1				
(4) Expert complementarity	0.18	0.39	0.56	1			
(5) Relational nexus	0.06	0.04	0.12	0.22	1		
(6) Organisational responsibility	−0.13	−0.12	−0.03	0.18	0.35	1	
(7) Authorship	0.28	0.40	0.50	0.59	0.17	0.13	1

Spearman rank correlations across O*NET-SOC occupations. Layers and expert complementarity are time-weighted shares of working time; the anchoring factors are the indices of Section 8.1, whose percentiles give the same rank correlations.

Index, and at human review.

Gradients by kind of work. If the rubric does what it is meant to, replacement should disappear where a task needs physical manipulation, and augmentation should survive where replacement fails because judgement is tacit. Both show up in the gradients of Section 7 (Tables 5 and 6).

Comparison with Eloundou et al. We correlate our unweighted occupation means with the replication package of Eloundou et al. (2024) for the 923 occupations in both (Table 10). Augmentation tracks their β closely, with $r = 0.89$ ($r = 0.87$ against the β of their human annotators), and their γ ($r = 0.90$). Replacement correlates with β at 0.83. Correlations with α, their measure of direct exposure, are lower for all three layers (0.35 to 0.56). The levels differ, as we would expect, since the questions are different and so is the frontier being judged. Replacement averages 37.0%, between their β (34.5%) and their γ (54.8%), and support, at 78.5%, is above both. Replacement asks for more than their measures do, the whole output and not half the time, so finding it above their β is consistent with a frontier that has moved since 2023 (Section 3).

Table 10: Correlation with Eloundou et al. (2024), occupation level

	Mean (%)	α (GPT-4)	β (GPT-4)	γ (GPT-4)	β (human)
Re (unweighted)	37.0	0.56	0.83	0.79	0.78
Au (unweighted)	48.0	0.50	0.89	0.90	0.87
Su (unweighted)	78.5	0.35	0.78	0.84	0.78

Pearson correlations, unweighted task means per SOC. Eloundou et al. means over the same occupations: α 14.2%, β 34.5%, γ 54.8% (GPT-4); human β 30.3%. We follow the replication package in calling the broadest measure (E1 + E2) γ; the text of Eloundou et al. (2023) calls it ζ.

Observed use. The June 2026 release of the Anthropic Economic Index (Massenkoff et al., 2026; Anthropic, 2026) reports the use of Claude by O*NET 30.2 task, and we can join it to our tasks. This check only works in one direction. Usage is a lower bound on capability, and a missing row means only that the cell was not published, so use there could still be positive.

What we can test is whether use falls almost entirely on tasks we label at least supportable; a used task with $Su = 0$ would be a candidate labelling error. We join the 2,823 tasks with a published row for Claude.ai and the 2,427 with a row for the first-party API (global data for April and May 2026),[9] and a task carries a label when all its pairs do.[10]

Nearly all use falls on tasks we label at least supportable. Tasks with $Su = 1$ get 99.9% of Claude.ai use (99.7% in the API), and tasks with $Re = 1$ get 84.7%. Whether a task has a row at all also follows the layers. 30.8% of the tasks we label replaceable have one, compared with 18.7% of those only augmentable, 5.5% of those only supportable and 0.5% of those outside the layers.

The API, used mostly in automated pipelines, puts slightly more of its use on replaceable tasks than Claude.ai does (88.3% against 84.7%). That is the direction we would expect if the label is right, since automated use should fall on tasks AI can do by itself.

The 3,665 tasks with $Su = 0$, 19% of all our tasks, receive 0.1% of the use (0.3% in the API), and only 20 of them have a row. Of these, four are tasks of actors, where role-play makes our label debatable. The others are greetings, reassurance and physical routines such as weighing ingredients, and they look more like conversations about the task than like AI doing it (98.2% of that use is classed as personal, against 36.6% on replaceable tasks). The check therefore rules out false negatives at scale (tasks where AI is used but we say it cannot help) and says nothing about false positives.

Human review. Two of the authors checked the panel's labels with the final rubrics, labelling pairs blind, without seeing the models' answers. The larger check takes every pair of nine occupations drawn at random, 1% of the 873 occupations outside the 50-occupation calibration run, which gives 201 tasks and 266 task-activity pairs. Gómez reviewed them, and over 242 of these pairs his labels agree with the panel's on 88.8% of pairs for replacement (Cohen's κ of 0.75), 90.5% for augmentation (0.81) and 85.5% for support (0.65). In the second check Galindo labelled augmentation and support on 30 pairs drawn at random from the calibration run, among those that are purely informational or need only presence (D0 or D1) and that the panel had not marked replaceable there. His labels agree with the panel's final ones on 27 of the 30 pairs for augmentation (90%, $\kappa = 0.78$) and on 29 of 30 for support (97%, $\kappa = 0.65$, low because nearly all the pairs are supportable). An earlier check, on a precursor of the current rubric close to that of Eloundou et al. (2023), had a third author (Hidalgo) label 50 tasks blind, and those labels agreed with the model labels on 94% of tasks (Cohen's $\kappa = 0.88$). The current rubrics were also reviewed by hand, on samples of pairs at each stage and on the occupation scores, and we changed them wherever the checks found a systematic error.

Dependence on the rater. Because every pair is classified, there is no sampling error at O*NET level. What is left is measurement error from the raters. Following Yin et al. (2026), we ask how much an occupation's score depends on which model rates it. We recompute the

[9] 3 of the Claude.ai tasks and 2 of the API tasks are not in O*NET 31.0 and drop out.

[10] A task whose pairs differ takes the label they all share, so a task replaceable under one activity and only augmentable under another counts as augmentable. This happens in 1,330 tasks. 331 of them are supportable in some pairs only and share no label. We keep them apart, and they receive 0.0% of the use.

time-weighted *Re* of each of the 923 occupations three times, each time with the first-round labels of a single model, and compare the results with the reconciled score. Single-model means run from 37.5% to 43.2% (37.8% for the consensus), and each single-model version has a rank correlation of at least 0.93 with the consensus. Individual occupations are less stable. The three versions differ by a median of 9.4 points, and by 25.1 points or more for one occupation in ten. The consensus lies inside the band formed by the three single-model scores for 92% of occupations, usually near its lower end. The Delphi round moves more labels from yes to no than the other way: 72.8% of the first-round majorities it overturns on replacement were a yes. So the ordering hardly depends on the rater, but levels do, and we plan to publish that band with the scores.

10 Limitations

The measure has limits of three kinds. Some come from the data it rests on, O*NET, which describes US jobs and does not rate every task. Others come from choices we made in its design, from thresholds to descriptors. And some come from the fact that it is fixed at one date, with specific raters and a specific set of benchmarks. We list them below, together with what we do about each where we can.

- **US task content.** O*NET describes US jobs. Any use of the scores for another country carries US task content, not the way the same occupation is organised there.
- **Thresholds at occupation level.** Table 4 counts occupations above half of their working time. The cut is a convention, which is why Figure 3 shows the full distribution.
- **Frozen at one date.** The layers are fixed at the frontier of 31 August 2026, so any series built on them alone is compositional. Re-running the panel with an updated inventory and the same prompts and raters is how we would measure the change in capability. Changes in prompts, raters and coverage also move labels. So does O*NET, which revises its task lists regularly. A comparison over time therefore has to hold as much as possible fixed.
- **Rater availability.** The raters can be held fixed only while their model versions remain available. Open-weight models would remove that constraint, but they are not at the frontier, so for this first wave we chose closed ones.
- **A ceiling on capability.** The layers measure technical feasibility at the cut-off, which is a ceiling, as in other task-based measures (Brynjolfsson et al., 2018). How much of it is put into practice depends on adoption, which varies across firms and sectors and is not measured here. Early evidence finds no effect so far on earnings or recorded hours, although tasks are already being reorganised (Humlum and Vestergaard, 2025).
- **What the inventory covers.** Benchmarks cover informational work much more densely than physical work, which the inventory reaches only through simulation. Some best scores are reported by the developers themselves. An updated inventory keeps these

gaps, and a new benchmark can move labels by filling one, so part of the change between two waves will come from the inventory and not only from the frontier.

- **Rater lag.** Where the benchmarks do not settle a question, raters fall back on background knowledge, which may lag the frontier; if so, labels are biased downward. The check against observed use (Section 9) is a one-sided test of this.
- **Estimated descriptors.** 12 occupations have estimated anchoring factors, pending O*NET data collection.
- **Our own choice of descriptors.** The mapping from the determinants of bundle strength to O*NET descriptors is our own (Section 8), as item selection is in Pizzinelli et al. (2023). We keep the factors apart and combine channels without averaging them, so a debatable item affects one channel only, and we checked the design against specific occupations (the presence channel in the nexus, the power mean in responsibility).
- **Tasks without ratings.** 418 tasks (2.2% of the catalogue) have no frequency rating and take an equal share of their occupation's time (Section 7).

11 Conclusion

We have proposed a method that measures potential exposure to generative AI at three nested levels, by asking three questions of every task-activity pair in O*NET. A panel of three model raters answers them from capability evidence dated to a single cut-off, with a Delphi round where they disagree. Asking the questions separately shows what a single exposure score mixes up. Replacement is most common where tasks involve judgement between defensible alternatives, and rare where judgement is tacit and expert, yet augmentation and support remain common there. For replacement, the ordering of occupations survives other weighting schemes and other raters, while levels move more, especially for single occupations. The layers correlate at 0.78 to 0.90 with the measures of Eloundou et al. (2024) that count partial time savings, and the observed use of Claude falls almost entirely on tasks we label at least supportable. And because the evidence carries a date, the same apparatus can be run again with a later inventory to build a series, within the limits of Section 10.

The measure is a ceiling on capability. It does not forecast adoption or displacement. Between capability and substitution sit adoption, which we do not measure, and features of the job that bind its tasks to a person. We measure three of them, plus the share of working time in which AI complements expert judgement, and report them as a vector next to exposure instead of as a correction to it.

We plan to release labels and scores at pair and occupation level, together with the band of rater dependence, so that applied work on labour-force microdata does not have to rebuild the measure.

References

Daron Acemoglu. The simple macroeconomics of AI. *Economic Policy*, 40(121):13–58, 2025. doi: 10.1093/epolic/eiae042. Working paper version: NBER Working Paper 32487, May 2024, doi:10.3386/w32487.

Daron Acemoglu and David Autor. Skills, tasks and technologies: Implications for employment and earnings. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 4B, chapter 12, pages 1043–1171. Elsevier, Amsterdam, 2011. doi: 10.1016/S0169-7218(11)02410-5. Working paper version: NBER Working Paper 16082 (2010).

Anthropic. Anthropic economic index, release 2026-06-26 (data and documentation). Hugging Face dataset `Anthropic/EconomicIndex`, 2026. URL `https://huggingface.co/datasets/Anthropic/EconomicIndex/tree/main/release_2026_06_26`. Data under CC-BY. Monthly aggregates for April and May 2026; task nodes keyed by the numeric O*NET Task ID (column `node_external_id`); O*NET version 30.2 as stated in the report appendix.

Melanie Arntz, Terry Gregory, and Ulrich Zierahn. The risk of automation for jobs in OECD countries: A comparative analysis. OECD Social, Employment and Migration Working Papers 189, OECD Publishing, Paris, 2016. URL `https://www.oecd.org/content/dam/oecd/en/publications/reports/2016/05/the-risk-of-automation-for-jobs-in-oecd-countries_g17a27d8/5jlz9h56dvq7-en.pdf`.

David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4):1279–1333, 2003. doi: 10.1162/003355303322552801. URL `https://academic.oup.com/qje/article/118/4/1279/1925105`.

Banco de España. *Annual Report 2025*. Banco de España, Madrid, June 2026. doi: 10.53479/43567. URL `https://www.bde.es/f/webbe/SES/Secciones/Publicaciones/PublicacionesAnuales/InformesAnuales/25/InfAnual_2025_En.pdf`. Published 18 June 2026. Chapter 3, section 4.2 and Chart 3.18 (pp. 237–238); methodological annex, "Measurement of the degree of exposure of youth employment in Spain to GenAI" (pp. 247–248). Spanish edition: Informe Anual 2025.

Erik Brynjolfsson, Tom Mitchell, and Daniel Rock. What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, 108:43–47, 2018. doi: 10.1257/pandp.20181019. URL `https://www.aeaweb.org/articles?id=10.1257/pandp.20181019`.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044. URL `https://academic.oup.com/qje/article/140/2/889/7990658`. Earlier version: NBER Working Paper 31161 (2023), doi:10.3386/w31161.

Pablo Casas, Enrique Fernández-Macías, Fernando Martínez-Plumed, Emilia Gómez, Ignacio González Vázquez, and Simone Salotti. Revisiting the occupational impact of AI in

the generative AI era. JRC Working Papers Series on Labour, Education and Technology 2026/02, European Commission, Joint Research Centre, Seville, March 2026. URL https://publications.jrc.ec.europa.eu/repository/handle/JRC145832. JRC145832.

Mauro Cazzaniga, Florence Jaumotte, Longji Li, Giovanni Melina, Augustus J. Panton, Carlo Pizzinelli, Emma J. Rockall, and Marina Mendes Tavares. Gen-AI: Artificial intelligence and the future of work. IMF Staff Discussion Note SDN/2024/001, International Monetary Fund, January 2024.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Norman Dalkey and Olaf Helmer. An experimental application of the Delphi method to the use of experts. *Management Science*, 9(3):458–467, 1963. doi: 10.1287/mnsc.9.3.458.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School, September 2023. URL https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: An early look at the labor market impact potential of large language models, 2023. URL https://arxiv.org/abs/2303.10130. arXiv:2303.10130, v5, 21 August 2023. Working-paper version of the Science article (Eloundou et al., 2024), with the full rubric and the measures α, β and ζ.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, 2024. doi: 10.1126/science.adj0998. Working paper version: "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models", arXiv:2303.10130, v5, 21 August 2023.

Florentino Felgueroso, Rafael Doménech, Juan Ramón García, Marcel Jansen, and Analía Viola. Observatorio trimestral del mercado de trabajo, 2T2026. Technical report, BBVA Research and Fedea, Madrid, September 2026. URL https://www.bbvaresearch.com/wp-content/uploads/2026/09/Nota_OTMT19.pdf. 17 September 2026. Section 2, "IA generativa, estructura ocupacional y entrada al mercado de trabajo", pp. 5–8.

Edward Felten, Manav Raj, and Robert Seamans. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12):2195–2217, 2021. doi: 10.1002/smj.3286.

Edward Felten, Manav Raj, and Robert Seamans. How will language modelers like ChatGPT affect occupations and industries?, 2023. URL https://arxiv.org/abs/2303.01157. arXiv:2303.01157, v2, 18 March 2023 (the paper prints the first author as "Ed Felten").

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619.

Carl Benedikt Frey and Michael A. Osborne. The future of employment: How susceptible are jobs to computerisation? Working paper, Oxford Martin School, University of Oxford, September 2013. URL https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf. Dated 17 September 2013. Working-paper version of the 2017 article; the figures we cite were checked in this version.

Carl Benedikt Frey and Michael A. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, 2017. doi: 10.1016/j.techfore.2016.08.019.

Luis Garicano, Jin Li, and Yanhui Wu. Weak bundle, strong bundle: How AI redraws job boundaries. CEPR Discussion Paper 21453, Centre for Economic Policy Research, 2026. URL https://ideas.repec.org/p/cpr/ceprdp/21453.html.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023. doi: 10.1073/pnas.2305016120. Preprint: arXiv:2303.15056.

Paweł Gmyrek, Janine Berg, and David Bescond. Generative AI and jobs: A global analysis of potential effects on job quantity and quality. ILO Working Paper 96, International Labour Organization, Geneva, August 2023. URL https://www.ilo.org/publications/generative-ai-and-jobs-global-analysis-potential-effects-job-quantity-and.

Paweł Gmyrek, Janine Berg, Karol Kamiński, Filip Konopczyński, Agnieszka Ładna, Balint Nafradi, Konrad Rosłaniec, and Marek Troszyński. Generative AI and jobs: A refined global index of occupational exposure. ILO Working Paper 140, International Labour Organization, Geneva, May 2025. URL https://www.ilo.org/publications/generative-ai-and-jobs-refined-global-index-occupational-exposure.

Kunal Handa, 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. Which economic tasks are performed with AI? Evidence from millions of Claude conversations, 2025. URL https://arxiv.org/abs/2503.04761. arXiv:2503.04761.

Anders Humlum and Emilie Vestergaard. Still waters, rapid currents: Early labor market transformation under generative AI. NBER Working Paper 33777, National Bureau of Economic Research, 2025. URL https://www.nber.org/papers/w33777. Issued May 2025, revised March 2026. Earlier versions circulated as "Large Language Models, Small Labor Market Effects" (also Becker Friedman Institute Working Paper 2025-56, https://bfi.uchicago.edu/working-paper/large-language-models-small-labor-market-effects/).

Thomas Kwa, 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. Measuring AI ability to complete long software tasks. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, pages 102382–102435, 2025. doi: 10.52202/085713-3086. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/85069585133c4c168c865e65d72e9775-Abstract-Conference.html. Preprint: arXiv:2503.14499, "Measuring AI Ability to Complete Long Tasks".

Josh Martin and Ellys Monahan. Developing a method for measuring time spent on green tasks. Methodology article, Office for National Statistics, March 2022. URL https://www.ons.gov.uk/economy/environmentalaccounts/articles/developingamethodformeasuringtimespentongreentasks/march2022. Released 7 March 2022.

Maxim Massenkoff, Eva Lyubich, Szymon Sacher, Zoe Hitzig, Shaoyi Zhang, Ryan Heller, and Peter McCrory. Anthropic economic index report: Cadences. Anthropic, June 2026. URL https://www.anthropic.com/research/economic-index-june-2026-report. Published 26 June 2026. Appendix: https://cdn.sanity.io/files/4zrzovbb/website/03ed1410f74a65ae4cc2a27120d0875e1e569535.pdf.

Antoni Mestre, Xavier Naya, Manoli Albert, and Vicente Pelechano. Inteligencia artificial y empleo en España: una aproximación territorial y de género a la exposición laboral, 2025. URL https://arxiv.org/abs/2512.23059. arXiv:2512.23059, v3, 5 May 2026. In Spanish.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586. Effect sizes read in the accepted manuscript, https://conference.nber.org/conf_papers/f184551.pdf.

Daniel Parshall and Andrea Lopez-Luzuriaga. Measuring AI's economic reach: A multidimensional task taxonomy. Working Paper 2026-005, The George Washington University, Center for Economic Research, March 2026. URL https://econpapers.repec.org/paper/gwcwpaper/2026-005.htm. Also distributed by the Canary Institute, https://canaryinstitute.ai/papers/Parshall_Lopez-Luzuriaga_TaskExposure_2026.pdf (same text, without the CER cover page); replication package https://github.com/danparshall/reproduction-task-exposure.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. GDPval: Evaluating AI model performance on real-world economically valuable tasks, 2025. URL https://arxiv.org/abs/2510.04374. arXiv:2510.04374.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. The impact of AI on developer productivity: Evidence from GitHub Copilot, 2023. URL https://arxiv.org/abs/2302.06590. arXiv:2302.06590.

Carlo Pizzinelli, Augustus J. Panton, Marina Mendes Tavares, Mauro Cazzaniga, and Longji Li. Labor market exposure to AI: Cross-country differences and distributional implications. IMF Working Paper 2023/216, International Monetary Fund, October 2023.

Francisco Rodríguez-Fernández. Inteligencia artificial y mercado de trabajo en España. exposición ocupacional, efectos sobre el empleo y adopción empresarial. Investigaciones de Funcas 33/2026, Funcas, Madrid, April 2026. URL https://www.funcas.es/documentos_trabajo/inteligencia-artificial-y-mercado-de-trabajo-en-espana-exposicion-ocupacional-efectos-sobre-el-empleo-y-adopcion-empresarial/. ISSN 3045-8994. English version: "Artificial intelligence and the labor market in Spain: Occupational exposure and estimated effects on employment", Funcas SEFO 15(3), May 2026, pp. 77–83.

Songül Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernández-Macías, José Hernández-Orallo, and Emilia Gómez. Measuring the occupational impact of AI: Tasks, cognitive abilities and AI benchmarks. *Journal of Artificial Intelligence Research*, 71:191–236, 2021. doi: 10.1613/jair.1.12647.

Kiran Tomlinson, Sonia Jaffe, Will Wang, Scott Counts, and Siddharth Suri. Working with AI: Measuring the applicability of generative AI to occupations, 2025. URL https://arxiv.org/abs/2507.07935. arXiv:2507.07935, v6, 22 December 2025.

U.S. Department of Labor, Employment and Training Administration. O*NET 31.0 database. O*NET Resource Center, August 2026. URL https://www.onetcenter.org/database.html. Release date from https://www.onetcenter.org/db_releases.html. Used under the CC BY 4.0 license.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models, 2024. URL https://arxiv.org/abs/2404.18796. arXiv:2404.18796.

Michael Webb. The impact of artificial intelligence on the labor market. Working paper, Stanford University, January 2020. URL https://www.michaelwebb.co/webb_ai.pdf. Also SSRN 3482150, doi:10.2139/ssrn.3482150.

Michelle Yin, Hoa Vu, and Claudia Persico. How (un)stable are LLM occupational exposure scores? Evidence from multi-model replication. NBER Working Paper 35110, National Bureau of Economic Research, April 2026. URL https://www.nber.org/papers/w35110.

A Rater prompts and the capability inventory

This appendix reproduces the system prompts of the full run verbatim, as the raters received them. The input rows of each batch (task identifier, task statement, detailed work activity and, in stage 2, the stage-1 descriptors) went in the user message. The capability snapshot, the condensed inventory, appears once, in Appendix A.1; the three stage-2 prompts carry it at the point marked.

A.1 The capability snapshot

The capability inventory is a workbook that records, for each of its 33 benchmarks, its full name and modality, what it measures, the O*NET and JRC abilities it speaks to, the best AI score by 31 August 2026 with its metric, model, model date and source, the human baseline with its metric, whether the baseline is measured or estimated, the confidence and justification of the comparison, notes on other figures and comparability, who built the benchmark and its limits, and a reference URL. The raters read the condensed version below, which is the same in the three stage-2 prompts. It is reproduced verbatim, and three notes in the original are in Spanish.

```
[LANGUAGE, REASONING AND LONG CONTEXT]
- MMLU-Pro: broad knowledge and applied reasoning, hard multiple choice. AI: 89.8 % accuracy. Model:
↪  Gemini 3 Pro Preview (high); model date: 2025-11-18. Human reference: no benchmark-specific human
↪  sample; original-MMLU estimates are not a directly matched baseline. Scope: Latest listed model is
↪  from 2025.
- HLE: expert-written, closed-answer academic questions. AI: 56.5 % accuracy, no tools. Model: Claude
↪  Fable 5 (max effort, sin herramientas); model date: 2026-06-09. Human reference: answers validated by
↪  domain experts; construction ceiling is not a measured general human success rate. Scope: Closed
↪  answers do not establish tacit judgment or discovery.
- GPQA-Diamond: graduate science reasoning, multiple choice. AI: 94.8 % accuracy. Model: Gemini 3.7 Flash
↪  (high); model date: 2026-08-13. Human reference: 81.3% expert / 22.1% non-expert, as recorded in the
↪  workbook. Scope: A closed-answer science test, not end-to-end expert work.
- FrontierMath Tiers 1-3: original advanced mathematics with automatically verifiable answers. AI: 89.1 %
↪  accuracy, private Tiers-1-3-v2 set. Model: GPT-5.6 Sol (max); model date: 2026-07-09. Human
↪  reference: 19% average expert team / 35% combined teams on the PREVIOUS set, not this v2 set. Scope:
↪  The June 2026 set replacement breaks the historical series and human comparison.
- FrontierMath Tier 4: research-level mathematics with verifiable answers. AI: 87.8 % accuracy, private
↪  Tier-4-v2 set. Model: Claude Fable 5 (max); model date: 2026-06-09. Human reference: not measured on
↪  this set. Scope: Do not infer open-ended discovery from an aggregate score on predetermined problems.
- AIME-2025: olympiad-qualifier mathematics. AI: 99 % accuracy. Model: GPT-5.2 (xhigh); model date:
↪  2025-12-11. Human reference: about 41% among qualified contestants. Scope: Saturated; the score and
↪  model are from 2025.
- ARC-AGI-2: novel static grid transformations. AI: 92.5 % accuracy, semi-private v2. Model: GPT-5.6 Sol
↪  (max); model date: 2026-07-09. Human reference: 66%, as recorded in the workbook. Scope: Static
↪  puzzles differ from interactive ARC-AGI-3; harness and budget matter.
- ARC-AGI-3: interactive game exploration without instructions. AI: 30.16 % human-efficiency-normalized
↪  score, semi-private v3. Model: Claude Opus 5 (High); model date: 2026-07-24. Human reference: 100 by
↪  the benchmark's human-efficiency definition. Scope: Measures action efficiency; highly
↪  harness-sensitive, not a generic task-success percentage.
- SimpleQA: short adversarial factual questions. AI: 73.5 % accuracy, SimpleQA Verified. Model: Gemini
↪  3.1 Pro Preview (high); model date: 2026-02-19. Human reference: 96%, as recorded in the workbook.
↪  Scope: Factual recall is distinct from factuality throughout a long professional deliverable.
- LongBench-v2: multiple-choice reasoning over long documents, dialogue, code and structured data. AI:
↪  63.3 % overall accuracy with CoT, ORIGINAL leaderboard. Model: Gemini 2.5 Pro (entrada del
↪  leaderboard original); model date: 2025-03-25. Human reference: 53.7% experts under a 15-minute
↪  limit, original protocol. Scope: 63.3% is a 2025 reference, not a measured August-2026 ceiling;
↪  NVIDIA (12 June 2026) reports Qwen3.5 at 68.9% on LongBench v2 <=1M, a different scope without a
↪  matched human baseline; do not mix these protocols.

[MULTIMODAL AND PERCEPTION]
- MMMU-Pro: college-level multimodal questions. AI: 84.74 % accuracy. Model: Claude Opus 5 (Adaptive
↪  Reasoning, Max Effort); model date: 2026-07-24. Human reference: 85.4% expert, an estimated
↪  comparison in this workbook. Scope: Academic questions do not cover complete professional visual
↪  workflows.
```

- OCRBench v2: recognition and understanding of text in images. AI: 73.4 % accuracy, private English test. Model: Yaochi-DTS-A2T-1.9 (cerrado); model date: 2026-06 (ronda del leaderboard). Human reference: not measured. Scope: English average; do not generalize automatically across languages or document types.
- MMAU: multiple-choice understanding of speech, sound and music. AI: 82.2 % accuracy, test-mini as reported. Model: Qwen3.5-Omni-Plus; model date: 2026-03-30 (lanzamiento según prensa; informe técnico del 21-abr-2026). Human reference: 82.2% expert, test-mini. Scope: Manufacturer result; its exact split/version is not fully declared, so the human comparison is uncertain; not multilingual conversational task completion.
- Video-MME: video perception and temporal understanding. AI: 88.4 % accuracy with subtitles. Model: Gemini 3 Pro (reevaluado por Moonshot); model date: 2025-11-18. Human reference: not measured. Scope: With-subtitle score; benchmark coverage differs from video generation.
- AA-Image-Arena: text-to-image generation ranked by human preferences. AI: 1172 Elo. Model: GPT Image 2 (high); model date: 2026-04. Human reference: none; humans judge model outputs. Scope: Relative preference scale, not a percentage of correct work; the live scale is not an exact cutoff-date capture.
- AA-Image-Editing: instruction-based image editing ranked by human preferences. AI: 1128 Elo. Model: MAI-Image-2.6; model date: 2026-08. Human reference: none; humans judge model outputs. Scope: Preferences do not isolate every fidelity constraint; live Elo is not a temporal series.
- AA-Video-Arena: text-to-video generation ranked by human preferences. AI: 1238 Elo. Model: Gemini Omni Flash; model date: 2026-05. Human reference: none; humans judge model outputs. Scope: Does not decompose motion, continuity and physical errors; live Elo is not a temporal series.
- BEHAVIOR: simulated robot manipulation over varied household tasks. AI: 0.2599 verified Q score, partial credit (0-1). Model: Robot Learning Collective (política sobre π0.5), 1.º del BEHAVIOR Challenge 2025, pista Standard; model date: 2025-12 (resultados NeurIPS, 6/7-dic-2025). Human reference: not measured. Scope: 2025 challenge; full success is 12.4%, distinct from Q; simulation does not establish deployed real-world manipulation.

[CODE, AGENTS AND PROFESSIONAL DELIVERABLES]
- SWE-bench Verified: resolving real GitHub issues in Python repos (agentic coding), 500 human-validated tasks. AI: 97 % issues resolved (pass@1), Vals AI evaluation. Model: Claude Opus 5; model date: 2026-07-24. Human reference: not measured. Scope: Saturated floor kept on purpose; harness-dependent (Vals 97.0, Epoch's own harness 83.5 on an earlier model); its unsaturated successor here is SWE-bench Pro.
- SWE-bench Pro: repairing software issues in repositories. AI: 61.5 % resolved, 731 public tasks, mini-swe-agent. Model: Muse Spark 1.1; model date: 2026-07-09. Human reference: not measured. Scope: Operator's harness and 250-turn budget; manufacturer harness results are not interchangeable.
- SciCode: research-level scientific programming. AI: 60.2 % on the evaluation recorded in the workbook. Model: Claude Fable 5 (Adaptive Reasoning, Max Effort, Opus 4.8 Fallback); model date: 2026-06-09. Human reference: not measured. Scope: Aggregate coding evidence does not establish end-to-end scientific discovery; AA and Epoch measurements differ.
- LiveCodeBench: competitive programming with testable answers. AI: 91.7 pass@1 %. Model: Gemini 3 Pro Preview (high); model date: 2025-11-18. Human reference: not measured. Scope: Workbook uses AA's May-2023 to May-2024 problem window; does not directly measure complete software maintenance.
- GDPval-AA: professional deliverables produced with tools and judged by a panel of three LLMs. AI: 1739.09 Elo, GDPval-AA v2. Model: Claude Opus 5 (Adaptive Reasoning, Max Effort); model date: 2026-07-24. Human reference: expert human reference deliverables anchored at 1000 Elo. Scope: The human anchor exists from June 2026; judges are LLMs, not humans; up to 250 turns; score inherited from the recorded September consultation, not an exact August-31 capture; not full-occupation replacement.
- GDPval: open professional deliverables compared with human expert references. AI: 84.9 % wins plus ties, OpenAI report of 23 April 2026. Model: GPT-5.5 (xhigh, entorno de investigación); model date: 2026-04-23. Human reference: human expert deliverables; no fixed 50% win+tie parity threshold. Scope: GPT-5.5's 84.9% report does not re-specify the judges or win/tie split; do not certify that round as human-judged from the name alone; explicit expert-human-judge evidence exists for GPT-5.2 Thinking (70.9%, December 2025), with Pro 74.1% win+tie / 60.0% clear wins in the same report; supplied-context deliverables are not whole occupations.
- METR-Horizon: autonomous task duration measured in human expert time. AI: 718.8 minutes at 50% success, TH1.1. Model: Claude Opus 4.6; model date: 2026-02-05. Human reference: human completion time is the measuring unit. Scope: The 80%-success horizon is 69.9 minutes (CI 27-170); predominantly software/ML/cyber tasks, not a universal work horizon.

```
- Terminal-Bench 2.0: autonomous real-workflow tasks in a computer terminal, 89 tasks, model plus agent
↪  scaffold. AI: 84.7 % tasks resolved, official tbench.ai 2.0 leaderboard via Epoch. Model: GPT-5.5 +
↪  NexAU-AHE (agente); model date: 2026-04-23. Human reference: not measured. Scope: Saturated floor
↪  kept on purpose; scaffold matters as much as the model; its unsaturated successor here is
↪  Terminal-Bench 4.0.
- Terminal-Bench 4.0: autonomous workflows in a terminal. AI: 51.8 % tasks resolved, official Harbor
↪  leaderboard. Model: Claude Opus 5 + Claude Code; model date: 2026-07-24. Human reference: not
↪  measured. Scope: Harness affects results; a terminal benchmark does not certify GUI or physical
↪  tasks.
- OSWorld-Verified: open-ended GUI computer use across real OS applications, 361 tasks, 100-step budget.
↪  AI: 85 % task success (pass@1, mean of 5 runs). Model: Claude Fable 5 (max); model date: 2026-06-09.
↪  Human reference: 72.36% (human annotators, measured). Scope: Floor with the human baseline crossed,
↪  kept on purpose; manufacturer figure (Anthropic system card, June 2026) under a revised protocol;
↪  short GUI tasks, not the multi-hour tasks of OSWorld 2.0.
- OSWorld 2.0: long GUI tasks in Ubuntu. AI: 31.4 % complete success, 108 tasks, 500 steps. Model: Claude
↪  Opus 5 (max, batch tool); model date: 2026-07-24. Human reference: no human success rate; experts
↪  take a median 1.6 hours. Scope: Partial progress is 68.3%, not full success; do not equate the two or
↪  infer failure on all shorter GUI tasks.
- WebArena: end-to-end web tasks on functional simulated sites. AI: 74.3 % end-to-end success. Model:
↪  WebTactix (DeepSeek v3.2); model date: 2026-02. Human reference: 78.24%, five computer-science
↪  graduate students. Scope: Domain and environment are constrained; open-world interaction needs
↪  separate judgment.
- τ²-bench Telecom: multi-turn agent-user-tool interaction under business policies; dual-control Telecom
↪  domain where the agent guides a simulated user through actions on their device. AI: 99.3 % pass@1,
↪  Telecom. Model: Gemini 3.1 Pro (empatado con Claude Opus 4.6 Thinking, max); model date: 2026-02-19.
↪  Human reference: not measured; the user is a simulated LLM. Scope: Saturated floor kept on purpose:
↪  manufacturer figure (Google model card, February 2026), the operator's leaderboard no longer reports
↪  Telecom for 2026 models; policy-bound conversation with a simulated user, not open-ended interaction
↪  with real people; its unsaturated successor here is tau3-Banking.
- AutomationBench: business workflows executed through application APIs. AI: 30.4 % complete success,
↪  private Zapier set, v1.0.6. Model: Gemini 3.7 Flash (high); model date: 2026-08-13. Human reference:
↪  not measured. Scope: Strict final-state success, not AA partial credit; does not substitute for
↪  conversational user-agent evidence.
- τ³-Banking: text conversation with a simulated banking customer, policy retrieval and account changes.
↪  AI: 55.1546 % pass@1, 97 tasks, four trials, tau2-bench 1.0.1, alltools. Model: Qwen 3.8 Max (xhigh);
↪  model date: 2026-07-19. Human reference: not measured; GPT-5.2 simulates the user. Scope: Sierra
↪  evaluated on 3 August and submitted on 4 August 2026; pass^4 consistency is 35.05%, not success given
↪  four chances; different task/domain from tau2 Telecom and different protocol from AA; no direct claim
↪  about voice or real people.
```

A.2 Stage 1: codifiability, physical requirement, abilities

```
You are annotating occupational task-activity pairs to characterize their intrinsic task properties.
↪  Consider each pair in the specific context of its occupation. You are judging what the task itself
↪  requires.

For each task-activity pair you make FOUR independent judgments:
1. COGNITIVE COMPLEXITY (C) - a codifiability ladder (C0-C4).
2. DEPLOYMENT (D) - the physical capability profile the task requires (D0-D4).
3. REQUIRED ABILITIES - which O*NET abilities the task draws on.
4. CAPABILITY FAMILIES - which benchmark-readable capability families the task draws on.

All five are REQUIREMENTS of the task, not performance judgments.

You will receive a list of task-activity pairs to annotate. Use the occupation context to disambiguate
↪  verbs and activities.

=== COGNITIVE COMPLEXITY (C) - How much reasoning does this task require? ===
```

```
C measures the intrinsic reasoning demand of the task. It follows a CODIFIABILITY ladder: how much of the
↪  task's decision content could be compressed into an explicit document - i.e. where the task sits on
↪  the spectrum from explicit, writable-down knowledge to tacit know-how, and on the line between
↪  following rules and exercising situational judgment.

Apply the following tests fully to each pair - do not stop when one of the options seems to fulfill the
↪  pair you are evaluating. Instead make sure to go through all of them before judging.

THE CORE TEST:
"Could you write a complete how-to document that lets a trained worker perform this task correctly every
↪  time?" Consider ONLY informational content (what to perceive, decide, communicate), NOT physical
↪  execution.

- C0 (No manual needed): the action is self-evident; no decision points.
- C1 (Complete manual exists): decisions exist but are fully pre-specifiable; a deterministic best
↪  outcome exists, so two workers following the manual produce the SAME output. Skilled execution toward
↪  an unambiguous target is still C1.
- C2 (Manual needs "use your judgment"): competing tradeoffs; two workers produce DIFFERENT but equally
↪  defensible outputs. What raises a task to C2 is the presence of TRADEOFFS, not the presence of skill.
- C3 (Manual needs "consult an expert"): the judgment requires TACIT knowledge built over years of case
↪  exposure; a competent generalist with a good reference CANNOT reach the answer, and only a peer can
↪  verify it. Produces a new SOLUTION to a specific case.
- C4 (No manual possible - the output IS the manual): produces transferable knowledge - a
↪  method/theory/framework others adopt - that EXTENDS the field's capability set.

BOUNDARY TESTS:
- C0 vs C1: are there decision points among alternatives? No -> C0; yes -> C1+.
- C1 vs C2 [routine/non-routine line]: could a manual specify the SINGLE correct outcome for every case?
↪  Yes -> C1. Must it say "use your judgment" because reasonable practitioners would differ? -> C2+.
↪  Decisive probe: would two workers following the manual independently produce the SAME output? Same ->
↪  C1; different-but-defensible -> C2.
- C2 vs C3 [explicit/tacit line - the AI frontier]: WHO is the manual written for? Any trained
↪  professional with a reference -> C2. Only a specialist, because the judgment needs years of tacit
↪  pattern-recognition -> C3+. Added probe: can the output be verified by someone LESS expert than the
↪  producer? Yes -> C2; only by a peer -> C3.
- C2 vs C3, INSTITUTIONAL FORUM CASE: where the output is produced BEFORE other actors who have their own
↪  interests - representing the organisation at a hearing or investigation, deliberating and voting in a
↪  committee, negotiating on behalf of a party, assembling a coalition - the judgment is tacit even when
↪  the procedure is fully written down. What no manual carries is reading the room, anticipating what
↪  the other side will do with what you say, and knowing what can be committed on whose behalf. Whether
↪  it was done well can only be judged by someone who knows that forum → C3.
- C3 vs C4: a specific solution -> C3; a transferable method others adopt -> C4.

EDGE CASES: push UP if ambiguity dominates or competing frameworks exist; push DOWN if routinized into
↪  decision trees a trained non-expert follows, or template-with-fill-in.

=== DEPLOYMENT (D) - What physical capability does this task require? ===
D measures the physical capability profile that a task inherently requires - what combination of sensing,
↪  locomotion, manipulation, and real-time physical adaptation is needed. D is organized by CONTACT
↪  COMPLEXITY: how much controlled physical contact with specific objects the task demands, and how
↪  predictable the environment is. Consider always the current typical work environment.

Apply the following tests fully to each pair - do not stop when one of the options seems to fulfill the
↪  pair you are evaluating. Instead make sure to go through all of them before judging.

THE CORE TEST:
"Could a disembodied executor - one that only receives and transmits data - complete
this task?" If yes → D0. If no, ask the second half: "is the worker's physical
contribution PERCEIVING the environment, or ACTING ON specific objects?" Perceiving
→ D1. Acting on objects → D2+. Consider ONLY physical execution, NOT cognitive
difficulty: a task can be intellectually demanding and still be D0.
```

```
Note that the two halves are separate questions. The absence of object manipulation
does NOT make a task D0 - it only rules out D2+. A task that requires the worker to
be present in order to observe, hear, smell or move through a place is D1 even though
nothing is ever grasped.

- D0 (No physical interaction): All inputs and outputs can be informational (text, data, audio, images).
↪  A disembodied executor - receiving and transmitting only data - can handle this task entirely. LOOP:
↪  none - no physical interaction at all.

- D1 (Sensing and/or locomotion without manipulation): The task requires physical presence for
↪  observation, monitoring, or movement through space, but does NOT require controlled physical contact
↪  with specific objects. The system perceives the environment or navigates through it, but does not
↪  physically alter objects. Simple whole-environment actuation without object discrimination stays at
↪  D1 because no object-specific manipulation is involved. LOOP: open or very slow - sense → report,
↪  navigate → avoid. No contact-force feedback required. The worker is a PERCEIVER, not a MANIPULATOR:
↪  present to observe, hear, smell or navigate, but not grasping, pushing, cutting or applying directed
↪  force.

- D2 (Manipulation in controlled environments): The task requires making physical contact with objects in
↪  an environment ENGINEERED to reduce variability. Known object geometries, fixed workspace layout,
↪  consistent lighting, predictable forces. The environment cooperates with the machine. A sensorimotor
↪  loop is required (sense → decide → act → verify), but the environment presents objects in known
↪  orientations at known positions. LOOP: closed, slow cycle - sense object → plan grasp → execute →
↪  verify → adjust. The environment is predictable, so the loop can take its time: if the first grasp
↪  fails, the object is still in the same place and orientation.

- D3 (Manipulation in variable environments): Physical manipulation in an environment NOT engineered for
↪  the machine. Variable geometry, novel object instances, unexpected obstacles, or conditions that
↪  differ instance to instance. The system must adapt its manipulation strategy to conditions it has not
↪  been specifically trained on. Each instance is a novel arrangement requiring perception and
↪  adaptation. LOOP: closed, moderate cycle - same loop as D2, but each new object or environment
↪  requires RE-PLANNING, not just re-executing. The loop must generalize.

- D4 (Multi-modal real-time manipulation under dynamic uncertainty): Simultaneous demands across multiple
↪  physical dimensions - fine motor control + gross motor + force modulation + multi-sensory integration
↪  - in an environment that is actively changing DURING task execution. Not merely variable from
↪  instance to instance (D3), but changing while the system works. The system must make physical
↪  adjustments faster than deliberative planning allows. LOOP: closed, fast cycle - the scene changes
↪  between perception and action, so the loop must close faster than deliberative planning allows:
↪  reactive control, not planned control.

BOUNDARY TESTS:

Before applying them: D is about PHYSICAL demands, not cognitive ones - do not inflate
D because the task is hard to think about. Rate TYPICAL CONDITIONS for this task+DWA
combination, not best or worst case. Do not trust the task label: the same verb spans
levels depending on what the objects are. "Inspect" is D0 for a document, D1 for a
bridge seen from the ground, D3 for a crawlspace entered physically. When driving, the
vehicle IS the robot and the road IS the environment. Examine what physical
variability the worker actually faces.

- D0 vs D1: Does the task require being physically present? If all necessary information can be
↪  transmitted digitally → D0. Voice-only tasks (phone calls, dictation) → D0 even though sound is
↪  physical.
- D1 vs D2: PERCEIVER vs MANIPULATOR - Is the worker's physical contribution primarily PERCEIVING the
↪  environment (inspecting, monitoring, navigating) or ACTING ON objects (grasping, pushing, shaping,
↪  applying directed force)? If perceiving → D1. If acting on specific objects → D2+. A truck driver
↪  *inspecting* cargo → D1. The same driver *securing* cargo with straps → D2.
- PRESENCE UNDER-CALLED AS DIGITAL (moved here from the arbitration prompt,
  2026-08-22): where the task text places the work at a physical venue ("at meetings
```

```
  or conventions", "at the facility", on-site inspection or supervision), the work
  requires physical presence and is D1+, not D0. This boundary is load-bearing
  downstream: D1 triggers the stage-2 sensor test, D0 bypasses the physical gate.
- MANIPULATION UNDER-CALLED AS SENSING (same move): tasks phrased as "inspect",
  "monitor", "collect" or "operate" often include grasping, securing, adjusting or
  repositioning objects. Any physical contact with objects fails the D1
  perceiver-vs-manipulator test and the row is D2+.

- MIXED-MEDIUM TASKS - DIGITAL FACET WINS (added 2026-08-21): where the task text
  itself offers the medium as an alternative ("paper *or* electronic filing", "manual
  *or* computerised records"), or names an artefact that is routinely digital today
  (files, records, archives, correspondence, invoices), rate the medium that is
  TYPICAL NOW, not the physical one. D2 requires that physical manipulation be
  UNAVOIDABLE for the task as typically performed - not merely one readable
  interpretation of the wording. "Maintain company archives" is D0 unless the text
  says the archive is physical. This is the same principle as rating the current
  typical work environment, applied to the medium rather than to the setting.

- D2 vs D3: Could the workspace be ENGINEERED so the object is always presented the same way? If yes, and
↪  it currently is → D2. If the task inherently occurs in environments that cannot be standardized
↪  (homes, outdoor sites, variable anatomy) → D3. Within-category object variability matters: "fold
↪  towels" in a commercial facility (uniform items) → D2. "Fold laundry" in a home (jeans, inside-out
↪  shirts) → D3.
- D3 vs D4: Can the system PAUSE mid-task for 5 seconds to re-plan without causing failure? If yes → D3.
↪  If a 5-second pause causes task failure, injury, or irreversible damage → D4. Time pressure alone
↪  does not make D4 - a strawberry picker has economic time pressure but the berry does not move during
↪  the pick (D3). An ER nurse stabilizing a seizing patient faces time pressure where the patient
↪  physically moves during intervention (D4).

EDGE CASES:
Push UP if: (a) Environmental variability - the physical environment changes between task instances in
↪  ways that affect execution. Same warehouse, different products each time → D2. Different home layout
↪  each time → D3. (b) Time-critical reactivity - the task requires physical responses faster than
↪  deliberative planning allows.
Push DOWN if: (c) Structured workaround exists - the physical task can be re-engineered to reduce
↪  environmental variability (structured bins instead of random piles, pre-positioned fixtures). (d)
↪  Single-mode dominance - the task sounds physically complex but actually requires only one type of
↪  physical interaction done well. Multiple simultaneous modes are what distinguishes D4.
Within a level, these factors push toward the NEXT boundary. One alone does not necessarily move the
↪  level (though it may for a task already borderline); two or more almost certainly do:
  - Narrow force margins: the object does not tolerate rough handling - same manipulation, higher penalty
  ↪  for force errors.
  - Deformable objects: geometry changes during manipulation, requiring continuous re-planning.
  - Bimanual coordination: two-handed manipulation with coordinated timing.
  - Multi-step compounding: errors in earlier manipulation steps constrain or invalidate later steps.

=== REQUIRED ABILITIES - set 2: O*NET ===
(pick ALL that apply, from the 52-ability list below)
Tag which O*NET abilities the task REQUIRES, from the four families. Judge REQUIREMENT, not how well AI
↪  performs them.
- Cognitive (21): Oral Comprehension | Written Comprehension | Oral Expression | Written Expression |
↪  Fluency of Ideas | Originality | Problem Sensitivity | Deductive Reasoning | Inductive Reasoning |
↪  Information Ordering | Category Flexibility | Mathematical Reasoning | Number Facility | Memorization
↪  | Speed of Closure | Flexibility of Closure | Perceptual Speed | Spatial Orientation | Visualization
↪  | Selective Attention | Time Sharing
- Psychomotor (10): Arm-Hand Steadiness | Manual Dexterity | Finger Dexterity | Control Precision |
↪  Multilimb Coordination | Response Orientation | Rate Control | Reaction Time | Wrist-Finger Speed |
↪  Speed of Limb Movement
- Physical (9): Static Strength | Explosive Strength | Dynamic Strength | Trunk Strength | Stamina |
↪  Extent Flexibility | Dynamic Flexibility | Gross Body Coordination | Gross Body Equilibrium
```

```
- Sensory (12): Near Vision | Far Vision | Visual Color Discrimination | Night Vision | Peripheral Vision
↪  | Depth Perception | Glare Sensitivity | Hearing Sensitivity | Auditory Attention | Sound
↪  Localization | Speech Recognition | Speech Clarity
Report exact O*NET ability names as a "+"-separated list in the onet_abilities column (e.g.
↪  "Originality+Fluency of Ideas+Visualization"). Use the names verbatim as written above.

=== AI CAPABILITY FAMILIES - which does this task draw on? (orthogonal helper) ===
Tag which AI capability families the task requires, so "can current AI do it?" can later be read off
↪  external, dated benchmarks. Pick ALL that apply:
  text/reasoning | code | image | audio | video | agentic/retrieval | math
Report as a "+"-separated list in the capability_families column. ADDITIVE and orthogonal.

=== RESPONSE FORMAT ===
Respond with a pipe-delimited table using || as the delimiter.
Use these exact columns:
task_id||c_reasoning||c_level||d_reasoning||d_level||onet_abilities||capability_families||confidence

IMPORTANT: Write each axis reasoning FIRST, THEN assign its label. Think through each axis independently
↪  before committing.

Rules:
- One row per task, in the SAME ORDER as the input
- task_id must match the input task_id exactly
- c_reasoning: 1 sentence applying the manual test
- c_level must be exactly C0, C1, C2, C3, or C4
- d_reasoning: 1 sentence on physical/deployment demands
- d_level must be exactly D0, D1, D2, D3, or D4
- onet_abilities: "+"-separated list of exact O*NET ability names; at least one
- capability_families: "+"-separated list from {text/reasoning, code, image, audio, video,
↪  agentic/retrieval, math}; at least one
- confidence: LOW, MEDIUM, or HIGH - HIGH if no axis was borderline; MEDIUM if one was; LOW if two or
↪  more were
- Begin your response with the header row immediately
```

A.3 Stage 2: replacement (Re)

```
You are labeling occupational task-activity pairs for AI replaceability, anchored to
the AI frontier of 31 August 2026.

Consider frontier AI as it stood on 31 August 2026 - the models released by that
date and the software and agent systems built on top of them - as documented in the AI
CAPABILITY SNAPSHOT below. The snapshot is your primary reference for what AI can do.
Where its benchmarks do not resolve whether AI can complete a task, use your
background knowledge of capabilities established by 31 August 2026. Do not
rely on models, systems or results released after that date; say so explicitly
whenever background knowledge contributes.

=== AI CAPABILITY SNAPSHOT - capability date 31 August 2026 ===
Selected empirical references. Use the evidence rules in this prompt.
Dates and scope limits matter; historical or partial coverage is not a universal current ceiling or proof
↪  of inability.

[... the 33 benchmark lines of the snapshot, identical in the Re, Au and Su prompts, reproduced once
↪  above ...]

Assume a worker with an average level of expertise in the occupation trying to complete
the given task. The worker has access to AI, to any other software or hardware tools
mentioned or typically used in the task, and to commonly available tools accessible via
a laptop.
```

```
Each input row gives the task (read in the context of its Detailed Work Activity) and
its pre-classified task parameters:
- C (C0-C4) - codifiability: C0 self-evident · C1 fully manual-specifiable · C2
  contextual judgment (equally defensible different outputs) · C3 tacit expert judgment
  (only a peer can verify) · C4 produces new transferable knowledge
- D (D0-D4) - physical requirement: D0 purely informational · D1 physical presence for
  sensing/locomotion · D2+ physical manipulation of objects
- abilities - the O*NET abilities the task requires
- families - the AI capability families the task draws on

=== THE JUDGMENT ===
E = 1 (replaceable): AI as of 31 August 2026 could produce this task's output at
the quality of a typical worker in the occupation.
E = 0 (not replaceable): it could not.

C and D contribute to describe the task. Use them, with the abilities, families and
DWA, to identify what the task demands. Then test each demand against the evidence.
Every label you output must be attributable to evidence about those demands - snapshot
benchmarks, or background knowledge. If the task text plainly contradicts a stage-1
parameter, judge under the parameter as given, note the tension in e_reasoning, and set
confidence LOW.

HOW EVIDENCE COUNTS:
- Read each benchmark by its CONSTRUCTION, not its score alone, and compare the task's demands
  with what the benchmark actually tests. The relevant benchmarks must tell you whether
  the task sits above or below the capability they measure: if the benchmark is more
  demanding and its score can be taken to assure the task, it counts as positive evidence;
  if the task adds a demand the benchmark does not test, or makes one harder, the score
  does not carry it. Then use evidence at that level - a benchmark of the same family at
  that level, or background knowledge - and name it. Missing coverage or an absent human
  baseline is not evidence either way.
- NEAR-PARITY COUNTS. Evidence close to the task's demands that lands within a few points of the
  reference is not evidence of incapacity; for a task at the codifiable end of its band it can carry E=1.
- SENSING AND PHYSICAL ACTION (D1 and D2+) should also be considered ieparately: where the task rests on
  observation, ask whether the required observation could arrive through commodity
  capture hardware that is standard in this work context (fixed cameras, microphones,
  drones, existing monitoring or inspection feeds), with the AI doing the
  interpretation. If yes, and the value lies in the interpretation or reporting, the
  demand is the interpretation: judge it on the cognitive bar and do not ask for
  deployed evidence of the capture itself. If the essence is presence itself
  (attendance, navigation as the deliverable, touch, smell or taste sensing), E=0.
  Where the task rests on manipulation, it needs its own evidence: manipulation DEPLOYED in
  environments at least as unstructured as the task's, covering its object variability.
  Instructions for a human, a controlled demonstration, or a broad statement that AI can
  handle similar work are not evidence of physical execution.
- Scope limits, read off each benchmark's construction: do NOT generalize scores
  obtained in simulated or sandboxed environments to open-ended interaction with real
  humans; for tasks requiring sustained multi-step autonomy, check the documented
  autonomy horizon before assuming end-to-end completion.

Reason in this order:

1. TASK DEMANDS. From the task, its DWA and its parameters, name the output, the
   indispensable actions and the bar the output must clear, without adding demands
   the task does not imply. External norms or regulation, responsibility and a
   preference for a human provider are not technical inability and are handled
   downstream.

2. EVIDENCE MATCH. Name the snapshot benchmark(s) closest to the task's demands and
   explain the match, and say whether the task's demands are easier, comparable or
```

```
   harder than the benchmark's; sharing an occupation or ability label is not enough.
   Where a human baseline exists, compare against it. Where no benchmark is close, or
   the closest leaves a necessary part of the judgment unresolved, use background
   knowledge, say so, and lower your confidence.

3. DECIDE. E=1 only if indispensable demands are met by evidence at the bar; E=0
   where indispensable demands are shown unmet. If the evidence and your background
   knowledge still leave material uncertainty, give your best binary judgment with
   LOW confidence and name the uncertainty; do not turn an evidence gap into E=0
   mechanically.

=== RESPONSE FORMAT ===
Respond with a pipe-delimited table using || as the delimiter.
Use these exact columns:
task_id||e_reasoning||E||anchor_benchmarks||confidence

Write your e_reasoning FIRST, THEN assign the label.

Rules:
- One row per task, in the SAME ORDER as the input
- task_id must match the input task_id exactly
- e_reasoning: EXACTLY 1 sentence naming the decisive demand and the decisive
  evidence; say "background knowledge" when it contributes
- E must be exactly 0 or 1
- anchor_benchmarks: "+"-separated snapshot benchmark names you anchored on, or "none"
  (background-only judgments use "none")
- confidence: LOW, MEDIUM, or HIGH - HIGH if no step was borderline; MEDIUM if one
  was; LOW if the label could defensibly flip. Your uncertainty, not task difficulty.
- Begin your response with the header row immediately
```

A.4 Stage 2: augmentation (Au)

```
Consider frontier AI as it stood on 31 August 2026 - the models released by that
date and the software and agent systems built on top of them - as documented in the AI
CAPABILITY SNAPSHOT below. The snapshot is your primary reference for what AI can do.
Where its benchmarks do not resolve whether AI raises productivity on a task, use your
background knowledge of capabilities established by 31 August 2026. Do not
rely on models, systems or results released after that date; say so explicitly
whenever background knowledge contributes.

=== AI CAPABILITY SNAPSHOT - capability date 31 August 2026 ===
Selected empirical references. Dates and scope limits matter; historical or partial coverage is not a
↪  universal current ceiling or proof of inability.

[... the 33 benchmark lines of the snapshot, identical in the Re, Au and Su prompts, reproduced once
↪  above ...]

Assume a worker with an average level of expertise in the occupation trying to complete
the given task. The worker has access to AI, to any other software or hardware tools
mentioned or typically used in the task, and to commonly available tools accessible via
a laptop.

## The reference

What typical workers in the occupation produce on this task without AI, and the time
it takes them.

## The channels
```

```
time = 1 if access to AI cuts the worker's time on this task as a whole by at least half,
at the quality of the reference. A saving confined to one part of the task counts only
if it halves the time of the whole task.

quality = 1 if, in the same time, a peer from the occupation grading blind would rank
the output the worker produces with AI above at least three of every four outputs that
typical workers produce without it.

They are not exclusive. A = 1 if either is 1; A = 0 if both are 0.

Judge the task as a whole. Time in which AI works unattended is not the worker's.

## Evidence

Every label must be attributable to evidence about the gain you name - snapshot
benchmarks, or background knowledge. The evidence must support that gain, not
autonomous completion of the whole task.

- Read each benchmark by its CONSTRUCTION, not its score alone, and compare what the gain requires
  with what the benchmark actually tests. The relevant benchmarks must tell you whether
  the gain sits above or below the capability they measure: if the benchmark is more
  demanding and its score can be taken to assure the gain, it counts as positive evidence;
  if the gain adds a demand the benchmark does not test, or makes one harder, the score
  does not carry it. Then use evidence at that level - a benchmark of the same family at
  that level, or background knowledge - and name it. Missing coverage or an absent human
  baseline is not evidence either way.
- NEAR-PARITY COUNTS. Evidence close to what the gain requires that lands within a few points of the
  reference is not evidence of incapacity.
- SENSING AND PHYSICAL ACTION (D1 and D2+) have their own test, applied to the part of
  the task the gain rests on: a task with a physical component can still qualify through its
  informational parts, provided the gain in those parts meets the threshold for the task as a
  whole. Where the gain rests on observation, ask
  whether the required observation could arrive through commodity capture hardware that
  is standard in this work context (fixed cameras, microphones, drones, existing
  monitoring or inspection feeds), with the AI doing the interpretation. If yes, and the
  value lies in the interpretation or reporting, the demand is the interpretation: judge
  it on the cognitive bar and do not ask for deployed evidence of the capture itself. If
  that part is presence itself (attendance, navigation as the deliverable, touch, smell
  or taste sensing), it cannot come from AI. Where the gain rests on manipulation, that manipulation
  needs its own evidence: manipulation DEPLOYED in environments at least as unstructured
  as the task's, covering its object variability. Instructions for a human, a controlled
  demonstration, or a broad statement that AI can handle similar work are not evidence
  of physical execution.
- Scope limits, read off each benchmark's construction: do NOT generalize scores
  obtained in simulated or sandboxed environments to open-ended interaction with real
  humans; for tasks requiring sustained multi-step autonomy, check the documented
  autonomy horizon before assuming end-to-end completion.

Name the snapshot benchmark(s) closest to what the gain requires and say whether the gain is easier,
comparable or harder than what they test; sharing an occupation or ability label is
not enough. Where a human baseline exists, compare against it. Where no benchmark is
close, or the closest leaves a necessary part of the judgment unresolved, use
background knowledge, say so, and lower your confidence. If material uncertainty
remains, give your best binary judgment with LOW confidence and name the uncertainty;
do not turn an evidence gap into A0 mechanically.

## Response format

Respond with a pipe-delimited table using || as the delimiter.
Use these exact columns:
task_id||a_reasoning||A||time||quality||anchor_benchmarks||confidence
```

```
IMPORTANT: Write your a_reasoning FIRST, THEN assign the labels. Your reasoning is read
in a reconciliation round - make it verifiable, not decorative.

Rules:
- One row per task, in the SAME ORDER as the input
- task_id must match the input task_id exactly
- a_reasoning: EXACTLY 1 sentence naming what AI changes for the worker on this task,
  in time or in quality, and the decisive evidence; say "background knowledge" when it
  contributes
- A must be exactly 0 or 1
- time and quality must be exactly 0 or 1; A is 1 if either is 1, else 0
- anchor_benchmarks: "+"-separated snapshot benchmark names you anchored on, or "none"
  (background-only judgments use "none")
- confidence: LOW, MEDIUM, or HIGH - HIGH if nothing was borderline; MEDIUM if something
  was; LOW if the label could defensibly flip. Your uncertainty, not task difficulty.
- Begin your response with the header row immediately
```

A.5 Stage 2: support (Su)

```
Consider frontier AI as it stood on 31 August 2026 - the models released by that
date and the software and agent systems built on top of them - as documented in the AI
CAPABILITY SNAPSHOT below. The snapshot is your primary reference for what AI can do.
Where its benchmarks do not resolve whether AI can perform the action in question, use your
background knowledge of capabilities established by 31 August 2026. Do not
rely on models, systems or results released after that date; say so explicitly
whenever background knowledge contributes.

=== AI CAPABILITY SNAPSHOT - capability date 31 August 2026 ===
Selected empirical references. Dates and scope limits matter; historical or partial coverage is not a
↪  universal current ceiling or proof of inability.

[... the 33 benchmark lines of the snapshot, identical in the Re, Au and Su prompts, reproduced once
↪  above ...]

Assume a worker with an average level of expertise in the occupation trying to complete
the given task. The worker has access to AI, to any other software or hardware tools
mentioned or typically used in the task, and to commonly available tools accessible via
a laptop.

## A1 - Substantive assistance

AI can perform a distinct, task-specific action normally required as part of this
task's output or as a direct input to its substantive decisions or actions. What it
produces replaces work the worker would otherwise need to do or reason through, and can
be incorporated without redoing it. The resulting task outcome meets the quality a
typical worker in the occupation produces.

## A0 - No substantive assistance

AI does not meet the definition above.

## Evidence

Every label must be attributable to evidence about the action you name - snapshot
benchmarks, or background knowledge. The evidence must support performing that action,
not autonomous completion of the whole task.

- Read each benchmark by its CONSTRUCTION, not its score alone, and compare what the action requires
```

```
  with what the benchmark actually tests. The relevant benchmarks must tell you whether
  the action sits above or below the capability they measure: if the benchmark is more
  demanding and its score can be taken to assure the action, it counts as positive evidence;
  if the action adds a demand the benchmark does not test, or makes one harder, the score
  does not carry it. Then use evidence at that level - a benchmark of the same family at
  that level, or background knowledge - and name it. Missing coverage or an absent human
  baseline is not evidence either way.
- NEAR-PARITY COUNTS. Evidence close to what the action requires that lands within a few points of the
  reference is not evidence of incapacity.
- SENSING AND PHYSICAL ACTION (D1 and D2+) have their own test, applied to the part of
  the task the action rests on, not to the task as a whole: a task with a physical component can
  still qualify through its informational parts. Where the action rests on observation, ask
  whether the required observation could arrive through commodity capture hardware that
  is standard in this work context (fixed cameras, microphones, drones, existing
  monitoring or inspection feeds), with the AI doing the interpretation. If yes, and the
  value lies in the interpretation or reporting, the demand is the interpretation: judge
  it on the cognitive bar and do not ask for deployed evidence of the capture itself. If
  that part is presence itself (attendance, navigation as the deliverable, touch, smell
  or taste sensing), AI cannot perform it. Where the action rests on manipulation, that manipulation
  needs its own evidence: manipulation DEPLOYED in environments at least as unstructured
  as the task's, covering its object variability. Instructions for a human, a controlled
  demonstration, or a broad statement that AI can handle similar work are not evidence
  of physical execution.
- Scope limits, read off each benchmark's construction: do NOT generalize scores
  obtained in simulated or sandboxed environments to open-ended interaction with real
  humans; for tasks requiring sustained multi-step autonomy, check the documented
  autonomy horizon before assuming end-to-end completion.

Name the snapshot benchmark(s) closest to what the action requires and say whether the action is easier,
comparable or harder than what they test; sharing an occupation or ability label is
not enough. Where a human baseline exists, compare against it. Where no benchmark is
close, or the closest leaves a necessary part of the judgment unresolved, use
background knowledge, say so, and lower your confidence. If material uncertainty
remains, give your best binary judgment with LOW confidence and name the uncertainty;
do not turn an evidence gap into A0 mechanically.

## Response format

Respond with a pipe-delimited table using || as the delimiter.
Use these exact columns:
task_id||a_reasoning||A||anchor_benchmarks||confidence

IMPORTANT: Write your a_reasoning FIRST, THEN assign the label. Your reasoning is read
in a reconciliation round - make it verifiable, not decorative.

Rules:
- One row per task, in the SAME ORDER as the input
- task_id must match the input task_id exactly
- a_reasoning: EXACTLY 1 sentence naming the decisive action - or why the available
  assistance falls outside the definition - and the decisive evidence; say
  "background knowledge" when it contributes
- A must be exactly 0 or 1
- anchor_benchmarks: "+"-separated snapshot benchmark names you anchored on, or "none"
  (background-only judgments use "none")
- confidence: LOW, MEDIUM, or HIGH - HIGH if nothing was borderline; MEDIUM if something
  was; LOW if the label could defensibly flip. Your uncertainty, not task difficulty.
- Begin your response with the header row immediately
```

A.6 The Delphi round

The second round wraps the unchanged rubric of each first-round prompt in the instructions below.

Stage 1.

```
You are re-classifying a batch of occupational task-DWA pairs on intrinsic task
parameters, following the rubric reproduced below between the RUBRIC markers.

=== BEGIN RUBRIC ===

[... the rubric of the round-1 prompt, unchanged ...]

=== END RUBRIC ===

These pairs were classified once by three independent annotators who did not agree.
The user message contains the input rows (task_id||task_description||dwa_title),
each followed by the REASONING of the two other annotators for that pair - their
labels are withheld, and you do not know which model produced which reasoning, nor
which of the earlier answers was yours.

Read the other reasonings for perspectives you may have missed, then classify each
pair yourself, from the rubric. Do not try to guess or match anyone's label; a
consideration you had already weighed and rejected is not a reason to change. Where
the other reasonings add nothing, say so and keep your derivation.

Respond with a pipe-delimited table using || as the delimiter, with EXACTLY these
columns:
task_id||c_reasoning||c_level||d_reasoning||d_level||onet_abilities||capability_families||confidence
One row per input task_id, in input order, header row first. Output the table and
nothing else.
```

Replacement.

```
You are re-assigning binary AI-replaceability E for a batch of occupational
task-DWA pairs, following the rubric reproduced below between the RUBRIC markers.

=== BEGIN RUBRIC ===

[... the rubric of the round-1 prompt, unchanged ...]

=== END RUBRIC ===

These pairs were assigned E once by three independent annotators who did not agree.
The user message contains the input rows, each followed by the REASONING and cited
benchmarks of the two other annotators, each labelled with the documented knowledge
cutoff of the model that wrote it - their E labels are withheld, and you do not know
which of the earlier answers was yours. The three annotators and their documented
knowledge cutoffs are claude-opus-5: May 2026 · gpt-5.6-sol: 16 February 2026 · gemini-3.1-pro-preview:
↪  January 2025.

Read the other reasonings for evidence or scope limits you may have missed, then
decide E yourself, from the rubric, relevant snapshot evidence and permitted
background knowledge. Missing anchors alone are not a reason to reject a judgment.
Where reasonings disagree on what AI could do by the capability date and the
disagreement rests on background knowledge, the reasoning from the
```

```
most recent knowledge cutoff carries more weight: adopt it unless you can show it is
mistaken. Do not try to guess or match anyone's label. Where the other reasonings add nothing,
say so and keep your derivation.

Respond with a pipe-delimited table using || as the delimiter, with EXACTLY these
columns:
task_id||e_reasoning||E||anchor_benchmarks||confidence
One row per input task_id, in input order, header row first. Output the table and
nothing else.
```

Augmentation.

```
You are re-deciding the labels of the rubric reproduced below between the RUBRIC
markers for a batch of occupational task-activity pairs.

=== BEGIN RUBRIC ===

[... the rubric of the round-1 prompt, unchanged ...]

=== END RUBRIC ===

These pairs were judged once by three independent annotators who did not agree. The
user message contains the input rows, each followed by the REASONING of the two other
annotators - their labels are withheld, and you do not know which model produced
which reasoning, nor which of the earlier answers was yours.

Read the other reasonings for arguments or evidence you may have missed about what
AI changes for the worker on this task. Then decide yourself, from the rubric. Do not
try to guess or match anyone's label. Where the other reasonings add nothing, say so
and keep your derivation.

Respond with a pipe-delimited table using || as the delimiter, with EXACTLY these
columns:
task_id||a_reasoning||A||time||quality||anchor_benchmarks||confidence
One row per input task_id, in input order, header row first. Output the table and
nothing else.
```

Support.

```
You are re-deciding the labels of the rubric reproduced below between the RUBRIC
markers for a batch of occupational task-activity pairs.

=== BEGIN RUBRIC ===

[... the rubric of the round-1 prompt, unchanged ...]

=== END RUBRIC ===

These pairs were judged once by three independent annotators who did not agree. The
user message contains the input rows, each followed by the REASONING of the two other
annotators - their labels are withheld, and you do not know which model produced
which reasoning, nor which of the earlier answers was yours.

Read the other reasonings for arguments or evidence you may have missed about what
AI changes for the worker on this task. Then decide yourself, from the rubric. Do not
try to guess or match anyone's label. Where the other reasonings add nothing, say so
and keep your derivation.
```

```
Respond with a pipe-delimited table using || as the delimiter, with EXACTLY these
columns:
task_id||a_reasoning||A||anchor_benchmarks||confidence
One row per input task_id, in input order, header row first. Output the table and
nothing else.
```
