ANTHROP\C

Appendix to “Labor market impacts of AI”

March 2026

Defining exposure

We begin at the task level. We only consider a task as covered if it sees sufficient traffic in our Economic Index samples (Handa et al. 2025), with more weight given to API usage (indicating deeper integration in production systems):

$$\underbrace{WorkUsage_t}_{\text{Weighted work-related count for task } t} = \underbrace{ClaudeWorkUsage_t}_{\text{Observed work-related usage in Claude.ai}} + \underbrace{APIUsage_t}_{\text{Observed usage in 1P API}}$$

$ClaudeWorkUsage_t$ is the count of tasks t in Claude.ai that were classified as work-related. We rely on the use case primitive from Appel et al. (2026) to restrict to work-related transcripts, as opposed to educational and personal use cases.[1] Restricting the counts to work-related uses seems to better capture the scope for labor market impacts. For example, using AI to explain a science lecture (coursework) or to get advice on treating an injury (personal use) seem fairly removed from automating the work of teaching or nursing. The final term, *APIUsaget*, counts all 1P API traffic of task *t*. We do not differentiate between API calls that are work related, since API calls typically signal integration into production workflows.[2]

The strict gate we impose is that $WorkUsage_t$ must be 100 or 0.0025% of traffic.[3] The task counts we observe have a long tail of low-count uses that likely reflect uncommon behavior, testing, or classifier errors. The exact cutoff has little impact on the job rankings. Tasks that do not meet the gate of $WorkUsage_t \geqslant 100$ are assigned an exposure of 0.

Some occupations share tasks, for example the task "Observe and evaluate students' performance, behavior, social development, and physical health" appears for ten types of K-12 teachers. Similarly, several tasks differ by just a word or comma. Since we do not record the contextual information to assign these to one job or another, we group these identical and highly similar tasks and allocate the task count equally across jobs according to employment shares.

This measure is meant to identify initial steps toward implementation and automation. It does not measure the intensive margin of AI use (increased

adoption in debugging vs. email drafting, for example), because the connection with usage and work performed is fuzzy and highly dependent on the task. At the level of a job, however, coverage will increase as more of its tasks are observed and automated.

Exposure of a task t is then given by $\tilde{r}_t$:

$$\tilde{r}_t = \mathbb{1}\{WorkUsage_t \geq 100\} \times \mathbb{1}\{\beta_t \geq 0.5\} \times \alpha_t$$

where β_t is the task's exposure from Eloundou et al. (2023), set to 1 if β is above zero. This amounts to ounting any tasks that are theoretically doable with an LLM or an LLM plus tools like websearch and image recognition.[5] We upgrade the tasks with a β of 0.5 because many LLMs today will have such capabilities, and the extent to which they are not actually helpful will be captured in the usage and automation measures.

The term α_t is a factor meant to upweight tasks seeing relatively more automative uses. It is defined as follows:

$$\alpha_t = \frac{1}{2} + \frac{1}{2} \times \underbrace{\frac{ClaudeWorkUsage_t \times AutoShare_t + APIUsage_t}{ClaudeWorkUsage_t + APIUsage_t}}_{\text{Automation share}},$$

where $AutoShare_t$ is the share of Claude.ai usage that is automative. A task that sees only augmentative uses and does not appear in the API transcripts would have α_t equal to 0.5 (because the Automation share gets zeroed out). A task with only automative uses would have $\alpha_t = 1$. The goal here is to have coverage scale with the degree of automation, but for augmentative uses to still count towards coverage.

In this framework, the jobs that have high exposure will be those whose tasks have at least some usage, are theoretically doable with an LLM, and are mostly used in automative patterns. For example, "Identify, compile, abstract, and code patient data, using standard classification systems" is a key task of Health

Information Technologists, estimated to account for 13 percent of their time (Tamkin and McCrory, 2025). With over 2,000 observations, it passes the count gate. Eloundou et al. (2023) give it a β value of 0.5, and the automation factor is 0.96 since over 90 percent of the usage is in the 1P API data. This contributes to high coverage for Health Information Technologists.

Then our job-level measure is calculated as:

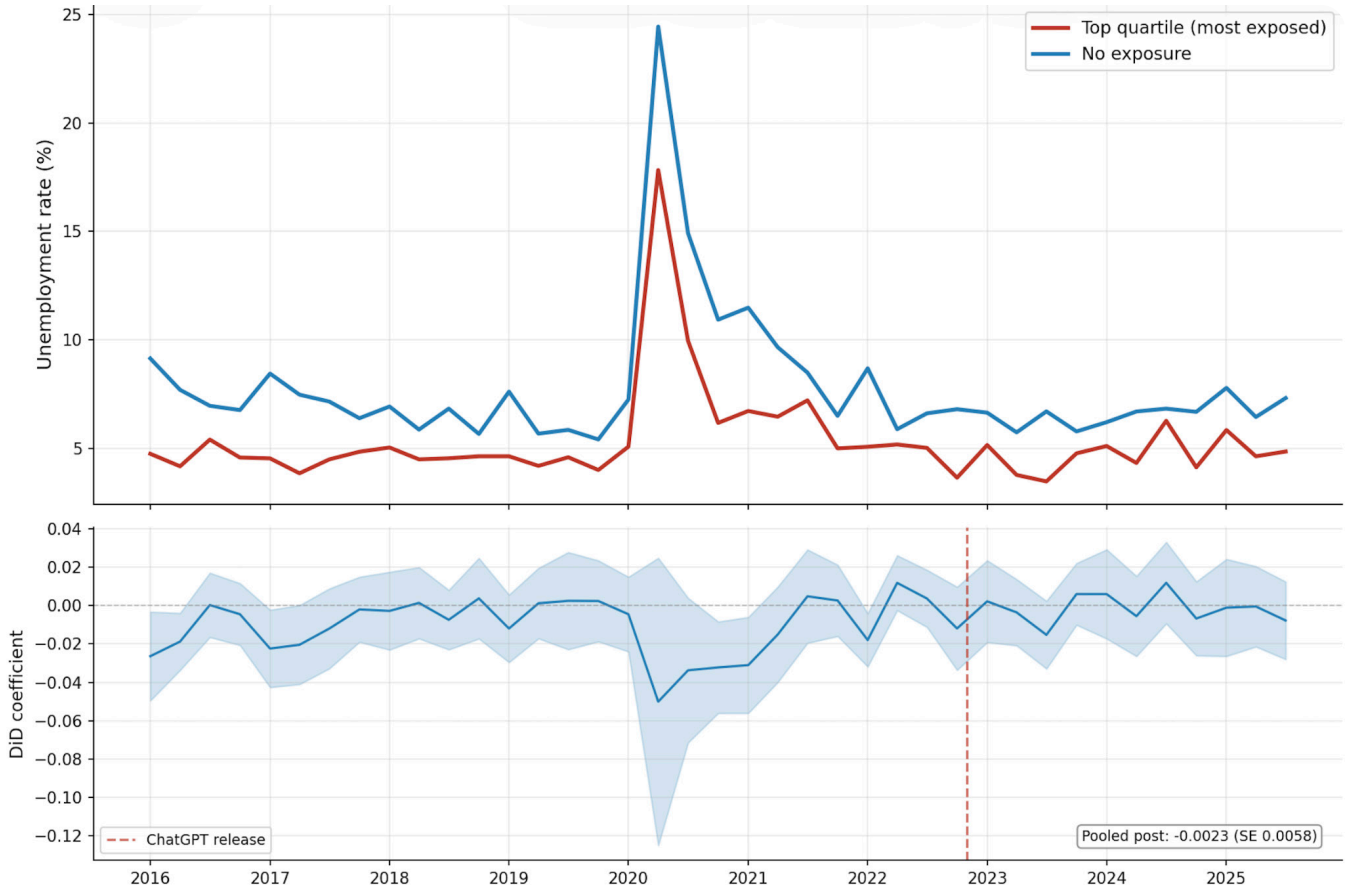
$$R_o = \frac{\sum_{t \in \mathcal{T}_o} w_t \cdot \widetilde{r}_t}{\sum_{t \in \mathcal{T}_o} w_t}$$

where T_o is the set of tasks for job o, w_t is the fraction of time spent on task *t* (Tamkin and McCrory, 2025), and $\tilde{r}_t$ is the task-level exposure. This measure can be loosely interpreted as the share of a job being performed or accelerated by an LLM, though the automation weighting means it is not a pure percentage of task coverage. Comparing two jobs that differ by 0.10 on the measure, the higher-coverage job could have a 10 percentage point higher share of their day covered by AI, or a 20 percentage point higher automation share in our data.

Additional results on employment impacts

Below, we show what happens when we look at the unemployment rates of 22-25 year old workers compared to similarly-aged workers in less exposed occupations. Prior to 2022, the unemployment rate is consistently lower for the more exposed group. If unemployment risk were increasing for young workers in exposed jobs, we would expect the gap between the two series to shrink and possibly reverse following the release of ChatGPT.

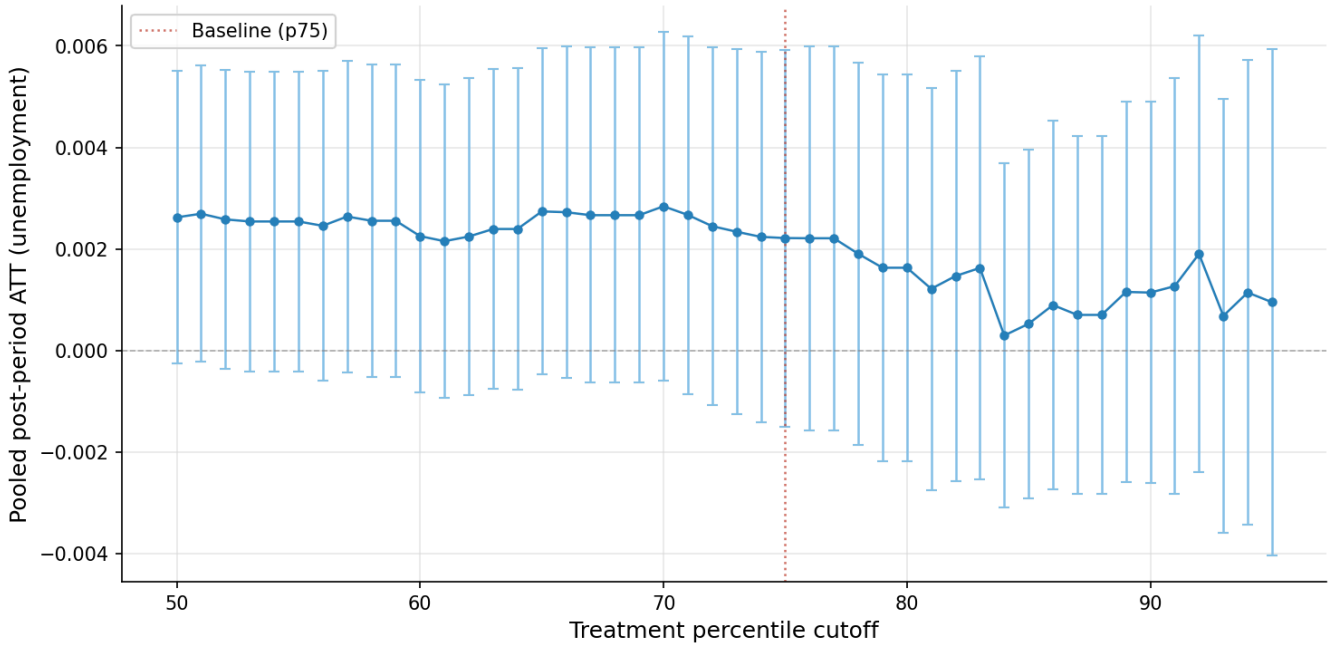
Instead, the gap has remained roughly constant. The difference-in-differences regression in the lower panel reflects this, with estimates close to zero throughout the post-treatment period. The pooled estimate suggests that overall change in unemployment for the exposed younger workers compared to the less exposed workers is negative and indistinguishable from zero.



Appendix Figure 1: Trends in the unemployment rate for young (22-25 year old) workers with high AI exposure and no AI exposure, Current Population Survey

In the main text, we designate the top quartile workers as high exposure. Below, we vary the threshold of exposure used to define workers as high exposure. We estimate pooled effects using the same difference-in-differences regression and control group as before, except every point estimate uses a different treated group. As we increase the threshold, zooming in on increasingly exposed occupations, the impact on unemployment remains small and insignificant.

The CPS is a survey, which makes it more prone to mismeasurement than administrative data. As one check on the results, we study what happens when we use data from unemployment insurance recipients, who report their former occupation when initially claiming benefits and may lose benefits if they respond inaccurately. These responses are aggregated at the state-quarter level by the Department of Labor into major SOC groups, so our measurement of exposure is less precise.[6]

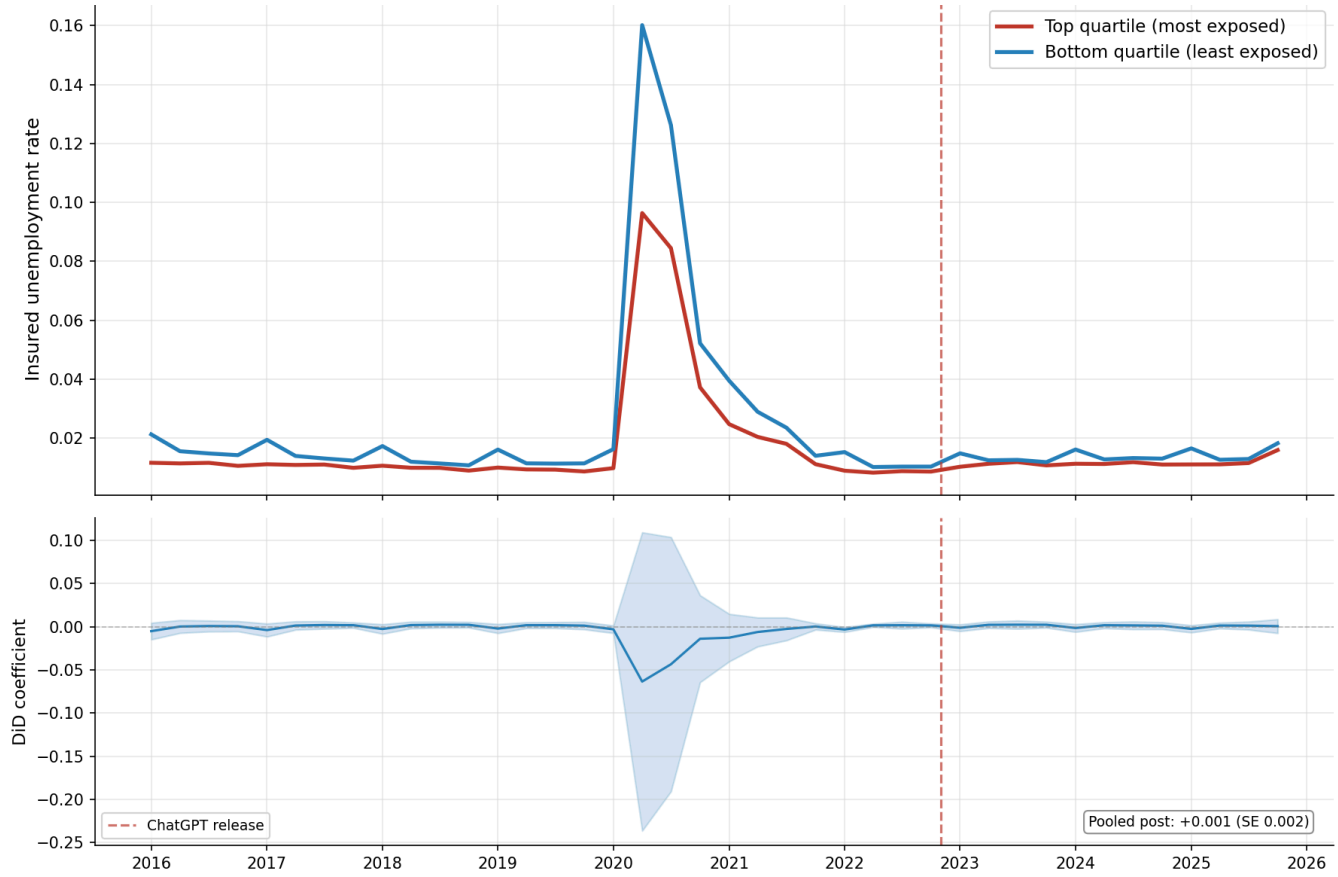


Appendix Figure 2: Sensitivity of the results to the percentile cutoff

Still, when we make quartiles using these categories, there's a stark difference in average observed coverage: the bottom quartile has 1% coverage, on average, compared to 31% in the top quartile. The four highly exposed categories are: Computer & Mathematical, Office & Administrative Support, Business & Financial, and Sales.

Using these numbers in the numerator—and using the number of CPS respondents in that SOC group and state in the denominator—we calculate the insured unemployment rate, which is an estimate of the share of people in that job category currently receiving unemployment benefits. In Appendix Figure 3, we show impacts on the insured unemployment rate. These estimates are qualitatively similar to the main results using only the CPS data. Apart from during the COVID-19 pandemic, the gap between the most and least exposed groups has remained consistent. (This gap is smaller than in the main results, likely because these broad categories group together several different occupations.)

The pooled estimate in the post-ChatGPT period is 0.1 percentage point and insignificant, echoing the main results that use only CPS data.



Appendix Figure 3: Main results using UI claims to measure the unemployment rate

Task granularity

One potential pitfall in the O*NET task database is that the level of specificity in a task statement is somewhat arbitrary. For example, the job Allergists and Immunologists has the task "Document patients' medical histories," and the job Naturopathic Physicians has the task "Document patients' histories, including identifying data, chief complaints, illnesses, previous medical or family histories, or psychosocial characteristics." These are most likely describing the same activities, but Allergists may look more exposed based on AI usage just because their task happens to be described in a generic way, earning a preference from our classifier.

There are multiple ways to address this. First, O*NET provides two higher levels of categorization: 332 IWAs (Intermediate Work Activities), for example "Assign work to others" and "Maintain health or medical records;" and 2,087

DWAs (Detailed Work Activities) including "Assign duties or work schedules to employees" and "Record patient medical histories." Researchers could aggregate the 18,000 task statements to the DWA or IWA level.

Combining tasks by the 2,087 DWAs would group very different work activities, damaging the signal. For example, the DWA "Translate information for others" includes the tasks "Recover data or decrypt seized data" and "Travel with or guide tourists who speak another language." There are even fewer IWAs, so this would make the problem worse.

One could also group by DWA and then condition on some degree of semantic similarity between tasks. This does not work either, because the DWA groupings can cleave otherwise identical tasks. For example, the task "Plan and prepare employee work schedules" is in the "Plan employee work schedules" DWA, while the task "Prepare employee work schedules" is in the "Prepare staff schedules or work assignments" DWA. The tasks differ by one word, but they are in different DWAs.

A potential fix we explored is grouping O*NET tasks with (i) a shared IWA and (ii) semantic similarity of at least 0.7. This leverages the grouping done by O*NET researchers to validate a seeming conceptual match across two tasks. In practice, this mostly agrees with our slightly simpler task grouping described in the main text, with a Spearman correlation coefficient of 0.9. There is still ample room for improvement in the O*NET task framework. Calibrating the specificity of tasks appropriately, and representing their inter-dependency, could be a useful next step.

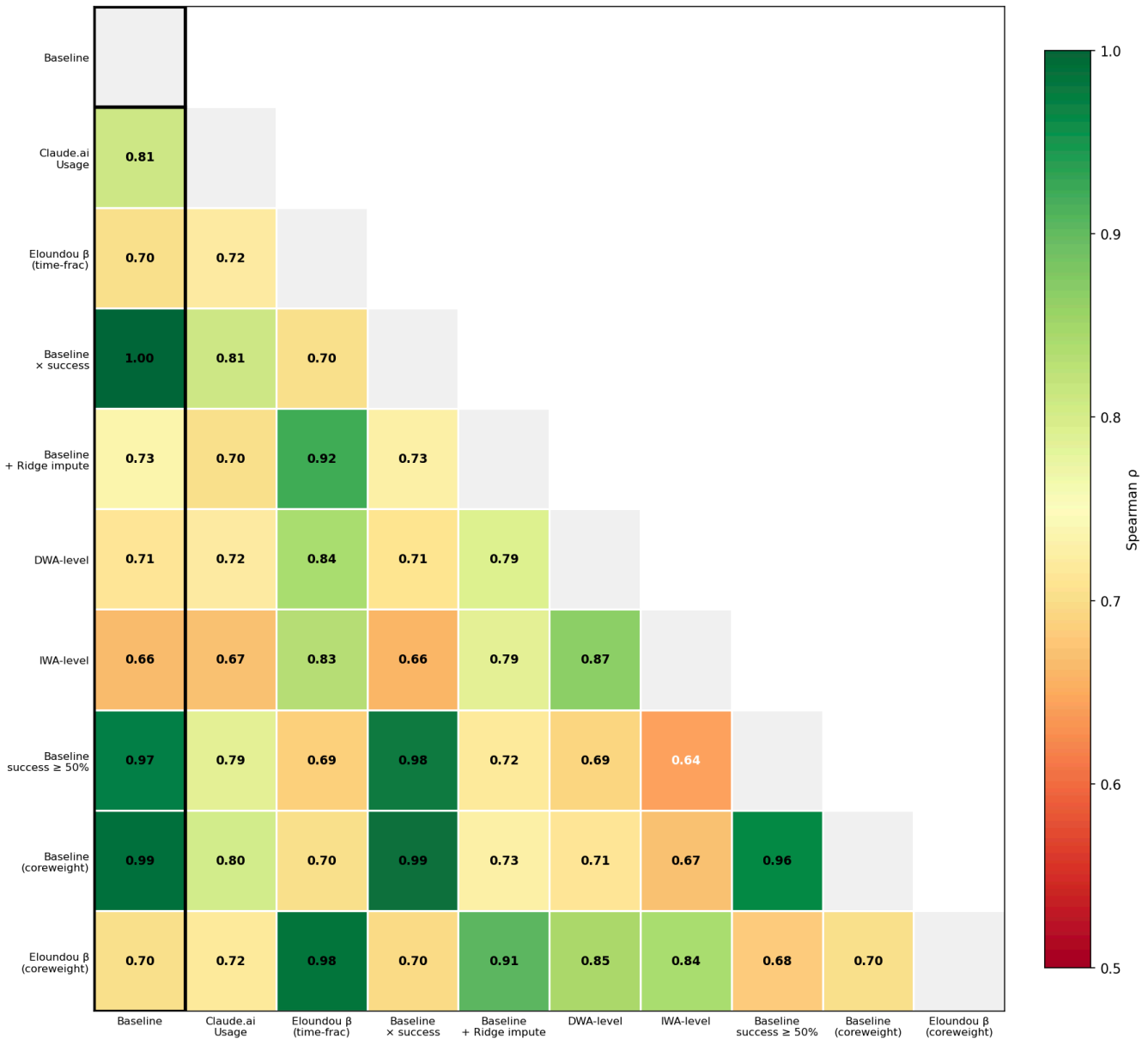
Comparing measures of occupational exposure exposure

There are many potential ways to measure job exposure. Here we explore the correlation between several different formulations. In the heatmap below, each cell gives the Spearman (rank-rank) correlation between the two job-level coverage measures.

- Baseline: this is our primary measure, observed coverage, described above.
- Claude.ai usage: raw usage counts from the Anthropic Economic Index,

summed to the occupation level and weighted by time fraction. This has a high Spearman correlation with our baseline measure (0.81) and the Eloundou et al. (2023) measure (0.72).Eloundou β (time-frac): Mean Eloundou et al. (2023) AI capability score across an occupation's tasks, weighted by time fraction rather than core weight. (The ONET data designates some "core" tasks which Eloundou et al. (2023) give twice the weight in their occupation-level averages.)

- Baseline x success: Baseline, except the task penetration measure is multiplied by the percent of conversations where the user's goal was achieved. In practice, this almost perfectly correlates with our baseline measure, suggesting that the success rate is not strongly correlated with occupation.
- Baseline + Ridge impute: Here we used a Ridge regression and task embeddings to impute usage for tasks that were not observed in the data. This was to address the potential issue that some tasks may be implemented on our platforms but do not appear often enough in our sample to meet the privacy threshold for inclusion in the data, which would underestimate the coverage of low-employment occupations. In practice, the imputed measure looks very similar to β.
- DWA-level: Baseline, except tasks are aggregated to the DWA level before taking the occupation level averages. We describe above why we did not use DWA or IWA to aggregate tasks.
- IWA-level: Baseline, except tasks are aggregated to the IWA level (~332 Intermediate Work Activities, coarser than DWAs) before taking the occupation-level averages, as in Tomlinson et al. (2025).
- Baseline x success over 50%: Baseline with hard gate requiring success equal to or greater than 50%. This is also highly correlated with the baseline.
- Baseline (coreweight): Same task-level values as Baseline, but aggregated to the occupation level using O*NET coreweight instead of time fractions.
- Eloundou β (coreweight): Same as Eloundou β (time-frac) but aggregated using O*NET coreweight instead of time fractions.



Appendix Figure 4: Spearman correlation across different measures of occupational exposure

---

[1] The use-case primitive (which is needed to calculate the share of work-related queries) was introduced in the September data, so in the August data we impute the percent of work-related Claude.ai traffic using a model trained on embeddings of the September tasks.

[2] Of course, some non-work conversations could impact the relevant jobs. People getting nutrition plans and detailed biology instruction from AI may rely less on professional biology tutors and nutritionists. In practice, counting

these transcripts tends to increase noise overall given the range of inquiries around neurology and animal care, for instance. There may be personal use cases of API workflows. We still weight these the same as work-related Claude.ai tasks because personal questions (say, nutrition advice) that are routed through an API call seem more likely to represent incipient automation.

[3] Usage of 100 represents 0.0025% of traffic using the previous two Anthropic Economic Index reports (2M from Claude.ai and 2M from 1P API), which is similar to the median share of time spent on a particular O*NET task (calculated by combining our time fraction estimates and employment estimates from BLS), 0.0014%. See Tomlinson et al. (2025) for a related gating approach.

[4] We incorporated success in an earlier measure of exposure, but its ranking correlates with this simpler version with r=0.999. This suggests that the success rate is not strongly correlated with occupation. We leave the success rate out for this reason, and to make it easier for integration with other data streams.

[5] See the exact elicitation in the Appendix of Eloundou et al. (2023).

[6] The specific data product is ETA 203 "Characteristics of the insured unemployed," available at this link: https://oui.doleta.gov/unemploy/DataDownloads.asp